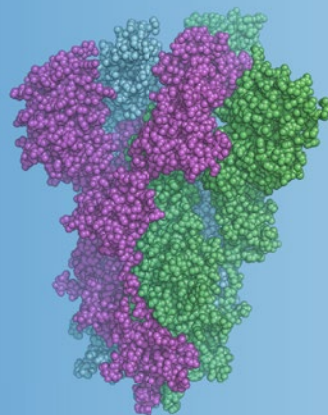


Methods in
Molecular Biology 2305

Springer Protocols



Raymond J. Owens *Editor*

Structural Proteomics

High-Throughput Methods

Third Edition

 Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor

John M. Walker

School of Life and Medical Sciences

University of Hertfordshire

Hatfield, Hertfordshire, UK

For further volumes:

<http://www.springer.com/series/7651>

For over 35 years, biological scientists have come to rely on the research protocols and methodologies in the critically acclaimed *Methods in Molecular Biology* series. The series was the first to introduce the step-by-step protocols approach that has become the standard in all biomedical protocol publishing. Each protocol is provided in readily-reproducible step-by-step fashion, opening with an introductory overview, a list of the materials and reagents needed to complete the experiment, and followed by a detailed procedure that is supported with a helpful notes section offering tips and tricks of the trade as well as troubleshooting advice. These hallmark features were introduced by series editor Dr. John Walker and constitute the key ingredient in each and every volume of the *Methods in Molecular Biology* series. Tested and trusted, comprehensive and reliable, all protocols from the series are indexed in PubMed.

Structural Proteomics

High-Throughput Methods

Third Edition

Edited by

Raymond J. Owens

The Rosalind Franklin Institute, Harwell Science Campus, Didcot, UK

 **Humana Press**

Editor

Raymond J. Owens
The Rosalind Franklin Institute
Harwell Science Campus
Didcot, UK

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-0716-1405-1 ISBN 978-1-0716-1406-8 (eBook)
<https://doi.org/10.1007/978-1-0716-1406-8>

© Springer Science+Business Media, LLC, part of Springer Nature 2008, 2015, 2021

Chapter 9 is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). For further details see license information in the chapters.

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cover Illustration Caption: EM of SARS-CoV-2 spike protein.

This Humana imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.

The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Preface

2021 marks the 50th Anniversary of the Protein Data Bank (PDB), the global initiative for the curation and dissemination of macromolecular structures to the international research community. It is therefore appropriate that the opening chapter in this third edition of *Structural Proteomics* in the *Methods in Molecular Biology* series, by Velankar and colleagues from the PDB, summarizes the current status of the PDB and its role in the future of structural biology. The PDB contains over 160,000 protein structures, but this is still a fraction of the total number of protein sequences available from genomic data. Therefore, in silico methods for predicting protein structure from sequence play an important part in structural proteomics. In the following two chapters, Edmunds and McGuffin and Madhusudhan et al. provide comprehensive and complementary user guides to the bioinformatics tools and resources for ab initio modeling of proteins and complexes, including ligand-docking algorithms.

Structural proteomics has been one of the key drivers for the development of streamlined workflows for sample preparation. The production of high-quality samples for structural studies, particularly mammalian membrane proteins and protein complexes, remains challenging. New protocols for tackling these difficult-to-express targets in higher eukaryote cells (insect and mammalian cells) are described in the chapters by Krasnoselska and van den Heuvel for the transient expression of membrane proteins in mammalian and insect cells, respectively. Novel approaches to protein production include the genome engineering of either the expression host or vector. Poterszman et al. describe a gene editing protocol to introduce purification tags into endogenous proteins for purification of macromolecular complexes. In the next two chapters, the crossover into synthetic biology is covered by Berger and Kubick and colleagues. Protocols for modifying the genome of the baculovirus, a widely used expression vector, to produce a novel synthetic virus are reported by the Berger group. In the next chapter, Kubick et al. describe incorporating non-natural amino acids using mammalian cell-free expression to produce fluorescent labeling of antibodies.

Isotopic labeling of endogenous proteins for NMR has now been extended to mammalian cells, and Baldus et al. describe the production of isotopically labelled microtubules and analysis of their interaction with MT-associated protein by solid-state NMR. X-ray crystallography remains a key technique for structural analysis; Orville and Aller present the state-of-the-art in the use of electron-free lasers for time-resolved crystallography. Since the publication of the last edition of *Methods in Molecular Biology* focused on Structural Proteomics, advances in detector technology and software algorithms have brought microscopy (cryo-EM) to the forefront of structural biology. Therefore, five chapters in this third edition are devoted to aspects of the use of electrons in structural biology. The so-called resolution revolution now means that the structures of large proteins and complexes can now be routinely determined at near-atomic resolution. Experimental and data analysis workflows are described in the chapters by Renault and Sorzano, respectively. The introduction of phase contrast methods has contributed to increasing resolution, and a guide to setting up and troubleshooting the Volta phase plate in cryo-EM data collection is detailed in the chapter by von Loeffelholz and Klaholz. In addition to single particle techniques, modern methods in cryo-EM include cryoelectron tomography and Microcrystal Electron Diffraction (MicroED). The combination of using focused ion beam milling to prepare

lamella thin enough for electrons to penetrate with cryo-EM imaging is providing protein structural information within cells. The cryoelectron tomography workflow for sample preparation and analysis is described in the chapter by Nováček et al. MicroED described by Danelius and Gonen is the newest cryo-EM technique enabling the rapid determination of peptide and organic molecule structures from microcrystalline powders. The technique has important applications in the structural analysis of pharmaceutical compounds and natural products.

I am grateful to all the contributors to this book for sharing their experience and expertise. I would also like to thank the *Methods in Molecular Biology* series editor, John Walker, for his guidance in preparing this volume and Springer for the opportunity to edit the third edition.

Didcot, UK

Raymond J. Owens

Contents

<i>Preface</i>	<i>v</i>
<i>Contributors</i>	<i>ix</i>

PART I STRUCTURAL BIOINFORMATICS

1 The Protein Data Bank Archive	3
<i>Sameer Velankar, Stephen K. Burley, Genji Kurisu, Jeffrey C. Hoch, and John L. Markley</i>	
2 Computational Methods for the Elucidation of Protein Structure and Interactions	23
<i>Nicholas S. Edmunds and Liam J. McGuffin</i>	
3 Methods for Molecular Modelling of Protein Complexes	53
<i>Tejashree Rajaram Kanitkar, Neeladri Sen, Sanjana Nair, Neelesh Soni, Kaustubh Amritkar, Yogendra Ramtirtha, and M. S. Madhusudhan</i>	

PART II PROTEIN PRODUCTION

4 High-Level Production of Recombinant Eukaryotic Proteins from Mammalian Cells Using Lentivirus	83
<i>Ester Behiels and Jonathan Elegheert</i>	
5 Transient Transfection and Expression of Eukaryotic Membrane Proteins in Expi293F Cells and Their Screening on a Small Scale: Application for Structural Studies	105
<i>Ganna O. Krasnoselska, Maud Dumoux, Nadisha Gamage, Harish Cheruvara, James Birch, Andrew Quigley, and Raymond J. Owens</i>	
6 Reproducible and Easy Production of Mammalian Proteins by Transient Gene Expression in High Five Insect Cells	129
<i>Maren Schubert, Manfred Nimtz, Federico Bertoglio, Stefan Schmelz, Peer Lukat, and Joop van den Heuvel</i>	
7 SynBac: Enhanced Baculovirus Genomes by Iterative Recombineering	141
<i>Hannah Crocker, Barbara Gorda, Martin Pelosse, Deepak Balaji Thimiri Govinda Raj, and Imre Berger</i>	
8 Gene Tagging with the CRISPR-Cas9 System to Facilitate Macromolecular Complex Purification	153
<i>Sylvain Geny, Simon Pichard, Arnaud Poterszman, and Jean-Paul Concordet</i>	

- 9 Synthesis of Fluorescently Labeled Antibodies Using Non-Canonical Amino Acids in Eukaryotic Cell-Free Systems 175
Marlitt Stech, Nathanaël Rakotoarinoro, Tamara Teichmann, Anne Zemella, Lena Thoring, and Stefan Kubick

PART III STRUCTURE DETERMINATION

- 10 Solid-State NMR Spectroscopy for Studying Microtubules and Microtubule-Associated Proteins 193
Yanzhang Luo, Shengqi Xiang, Alessandra Lucini Paioni, Agnes Adler, Peter Jan Hooikaas, A. S. Jijumon, Carsten Janke, Anna Akhmanova, and Marc Baldus
- 11 Dynamic Structural Biology Experiments at XFEL or Synchrotron Sources 203
Pierre Aller and Allen M. Orville
- 12 From Tube to Structure: SPA Cryo-EM Workflow Using Apoferritin as an Example 229
Christoph A. Diebolder, Rebecca S. Dillard, and Ludovic Renault
- 13 Image Processing in Cryo-Electron Microscopy of Single Particles: The Power of Combining Methods 257
Carlos Oscar S. Sorzano, Amaya Jiménez-Moreno, David Maluenda, Erney Ramírez-Aportela, Marta Martínez, Ana Cuervo, Robert Melero, Jose Javier Conesa, Ruben Sánchez-García, David Strelak, Jiri Filipovic, Estrella Fernández-Giménez, Federico de Isidro-Gómez, David Herreros, Pablo Conesa, Laura del Caño, Yuniór Fonseca, Jorge Jiménez de la Morena, Jose Ramon Macías, Patricia Losana, Roberto Marabini, and Jose-Maria Carazo
- 14 Setup and Troubleshooting of Volta Phase Plate Cryo-EM Data Collection 291
Ottilie von Loeffelholz and Bruno P. Klaholz
- 15 Cryo-Focused Ion Beam Lamella Preparation Protocol for in Situ Structural Biology 301
Jana Moravcová, Radka Dopitová, Matyáš Pinkas, and Jiří Nováček
- 16 Protein and Small Molecule Structure Determination by the Cryo-EM Method MicroED 323
Emma Danelius and Tamir Gonen
- Index* 343

Contributors

- AGNES ADLER • *NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands*
- ANNA AKHMANOVA • *Cell Biology, Neurobiology and Biophysics, Department of Biology, Faculty of Science, Utrecht University, Utrecht, The Netherlands*
- PIERRE ALLER • *Diamond Light Source Ltd, Didcot, UK*
- KAUSTUBH AMRITKAR • *Indian Institute of Science Education and Research Pune, Pashan, Pune, India*
- MARC BALDUS • *NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands*
- ESTER BEHIELS • *Univ. Bordeaux, CNRS, Interdisciplinary Institute for NeuroScience (IINS), UMR5297 CNRS/UB, Bordeaux, France*
- IMRE BERGER • *Bristol Synthetic Biology Centre BrisSynBio, Biomedical Sciences, School of Biochemistry, University of Bristol, Bristol, UK; School of Chemistry, Max Planck Bristol Centre for Minimal Biology, Bristol, UK*
- FREDERICO BERTOGLIO • *Department of Biotechnology, Institut fuer Biochemie, Biotechnologie und Bioinformatik, Technische Universitaet Braunschweig, Braunschweig, Germany*
- JAMES BIRCH • *Diamond Light Source Ltd, Didcot, UK*
- STEPHEN K. BURLEY • *Research Collaboratory for Structural Bioinformatics Protein Data Bank, Center for Integrative Proteomics Research, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA; Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ, USA; Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ, USA; Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA, USA*
- LAURA DEL CAÑO • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- JOSE-MARIA CARAZO • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- HARISH CHERUVARA • *Diamond Light Source Ltd, Didcot, UK*
- JEAN-PAUL CONCORDET • *Laboratoire Structure et Instabilité des Génomes, Inserm U1154, CNRS UMR 7196, Museum National d'Histoire Naturelle, Paris, France*
- JOSE JAVIER CONESA • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- PABLO CONESA • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- HANNAH CROCKER • *Bristol Synthetic Biology Centre BrisSynBio, Biomedical Sciences, School of Biochemistry, University of Bristol, Bristol, UK*
- ANA CUERVO • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- EMMA DANELIUS • *Department of Biological Chemistry, University of California Los Angeles, Los Angeles, CA, USA; Department of Physiology, University of California Los Angeles, Los Angeles, CA, USA; Howard Hughes Medical Institute, University of California Los Angeles, Los Angeles, CA, USA*
- CHRISTOPH A. DIEBOLDER • *The Netherlands Centre for Electron Nanoscopy (NeCEN), Leiden University, Leiden, The Netherlands*
- REBECCA S. DILLARD • *The Netherlands Centre for Electron Nanoscopy (NeCEN), Leiden University, Leiden, The Netherlands*

- RADKA DOPITOVÁ • *CEITEC, Masaryk University, Brno, Czech Republic*
- MAUD DUMOUX • *The Rosalind Franklin Institute, Didcot, UK*
- NICHOLAS S. EDMUNDS • *School of Biological Sciences, University of Reading, Reading, UK*
- JONATHAN ELEGHEERT • *Univ. Bordeaux, CNRS, Interdisciplinary Institute for NeuroScience, IINS, Bordeaux, France*
- ESTRELLA FERNÁNDEZ-GIMÉNEZ • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- JIRI FILIPOVIC • *Masaryk University, Brno, Czech Republic*
- YUNIOR FONSECA • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- NADISHA GAMAGE • *Diamond Light Source Ltd, Didcot, UK*
- SYLVAIN GENY • *Laboratoire Structure et Instabilité des Génomes, Inserm U1154, CNRS UMR 7196, Museum National d'Histoire Naturelle, Paris, France*
- TAMIR GONEN • *Department of Biological Chemistry, University of California Los Angeles, Los Angeles, CA, USA; Department of Physiology, University of California Los Angeles, Los Angeles, CA, USA; Howard Hughes Medical Institute, University of California Los Angeles, Los Angeles, CA, USA*
- BARBARA GORDA • *School of Biochemistry, Bristol Synthetic Biology Centre BrisSynBio, Biomedical Sciences, University of Bristol, Bristol, UK*
- DAVID HERREROS • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- JOOP VAN DEN HEUVEL • *Department of Structure and Function of Proteins, Helmholtz Zentrum für Infektionsforschung GmbH, Braunschweig, Germany*
- JEFFREY C. HOCH • *BioMagResBank, Department of Molecular Biology and Biophysics, UConn Health, Farmington, CT, USA*
- PETER JAN HOOIKAAS • *Cell Biology, Neurobiology and Biophysics, Department of Biology, Faculty of Science, Utrecht University, Utrecht, The Netherlands*
- FEDERICO DE ISIDRO-GÓMEZ • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- CARSTEN JANKE • *Institut Curie, PSL Research University, CNRS UMR3348, Orsay, France; Université Paris Sud, Université Paris-Saclay, CNRS UMR3348, Orsay, France*
- A. S. JIJUMON • *Institut Curie, PSL Research University, CNRS UMR3348, Orsay, France; Université Paris Sud, Université Paris-Saclay, CNRS UMR3348, Orsay, France*
- AMAYA JIMÉNEZ-MORENO • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- TEJASHREE RAJARAM KANITKAR • *Indian Institute of Science Education and Research Pune, Pashan, Pune, India*
- BRUNO P. KLAHOLZ • *Centre for Integrative Biology (CBI), Department of Integrated Structural Biology, IGBMC (Institute of Genetics and of Molecular and Cellular Biology), Illkirch, France; Centre National de la Recherche Scientifique (CNRS) UMR 7104, Illkirch, France; Institut National de la Santé et de la Recherche Médicale (Inserm) U964, Illkirch, France; Université de Strasbourg, Strasbourg, France*
- GANNA O. KRASNOSELSKA • *Division of Structural Biology, University of Oxford, The Wellcome Centre for Human Genetics, Headington, Oxford, UK*
- STEFAN KUBICK • *Branch Bioanalytics and Bioprocesses (IZI-BB), Fraunhofer Institute for Cell Therapy and Immunology (IZI), Potsdam, Germany; Faculty of Health Sciences, Joint Faculty of the Brandenburg University of Technology Cottbus-Senftenberg, The Brandenburg Medical School Theodor Fontane and the University of Potsdam, Senftenberg, Germany*
- GENJI KURISU • *Protein Data Bank Japan, Institute for Protein Research, Osaka University, Suita, Osaka, Japan*

- OTTILIE VON LOEFFELHOLZ • *Department of Integrated Structural Biology, Centre for Integrative Biology (CBI), IGBMC (Institute of Genetics and of Molecular and Cellular Biology), Illkirch, France; Centre National de la Recherche Scientifique (CNRS) UMR 7104, Illkirch, France; Institut National de la Santé et de la Recherche Médicale (Inserm) U964, Illkirch, France; Université de Strasbourg, Strasbourg, France*
- PATRICIA LOSANA • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- PEER LUKAT • *Department of Structure and Function of Proteins, Helmholtz Zentrum für Infektionsforschung GmbH, Braunschweig, Germany*
- YANZHANG LUO • *NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands*
- JOSE RAMON MACÍAS • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- M. S. MADHUSUDHAN • *Indian Institute of Science Education and Research Pune, Pashan, Pune, India*
- DAVID MALUENDA • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- ROBERTO MARABINI • *Universidad Autónoma de Madrid, Madrid, Spain*
- JOHN L. MARKLEY • *BioMagResBank, Biochemistry Department, University of Wisconsin-Madison, Madison, WI, USA*
- MARTA MARTÍNEZ • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- LIAM J. MCGUFFIN • *School of Biological Sciences, University of Reading, Reading, UK*
- ROBERT MELERO • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- JANA MORAVCOVÁ • *CEITEC, Masaryk University, Brno, Czech Republic*
- JORGE JIMÉNEZ DE LA MORENA • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- SANJANA NAIR • *Indian Institute of Science Education and Research Pune, Pashan, Pune, India*
- MANFRED NIMTZ • *RG Cellular Proteomics, Department of Structure and Function of Proteins, Helmholtz Zentrum für Infektionsforschung GmbH, Braunschweig, Germany*
- JIRÍ NOVÁČEK • *CEITEC, Masaryk University, Brno, Czech Republic*
- ALLEN M. ORVILLE • *Diamond Light Source Ltd, Didcot, UK*
- RAYMOND J. OWENS • *Division of Structural Biology, University of Oxford, The Wellcome Centre for Human Genetics, Headington, Oxford, UK; The Rosalind Franklin Institute, Didcot, UK*
- ALESSANDRA LUCINI PAIONI • *NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands*
- MARTIN PELOSSE • *Bristol Synthetic Biology Centre BrisSynBio, Biomedical Sciences, School of Biochemistry, University of Bristol, Bristol, UK*
- SIMON PICHARD • *Institut de Génétique et de Biologie Moléculaire et Cellulaire, Integrated Structural Biology, Equipe labellisée Ligue Contre le Cancer, Illkirch, France; Centre National de la Recherche Scientifique, UMR7104, Illkirch, France; Institut National de la Santé et de la Recherche Médicale, U1258, Illkirch, France; Université de Strasbourg, Illkirch, France*
- MATYÁŠ PINKAS • *CEITEC, Masaryk University, Brno, Czech Republic*
- ARNAUD POTERSZMAN • *Institut de Génétique et de Biologie Moléculaire et Cellulaire, Integrated Structural Biology, Equipe labellisée Ligue Contre le Cancer, Illkirch, France; Centre National de la Recherche Scientifique, UMR7104, Illkirch, France; Institut National de la Santé et de la Recherche Médicale, U1258, Illkirch, France; Université de Strasbourg, Illkirch, France*
- ANDREW QUIGLEY • *Diamond Light Source Ltd, Didcot, UK*

- NATHANAËL RAKOTOARINORO • *Branch Bioanalytics and Bioprocesses (IZI-BB), Fraunhofer Institute for Cell Therapy and Immunology (IZI), Potsdam, Germany*
- ERNEY RAMÍREZ-APORTELA • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- YOGENDRA RAMTIRTHA • *Indian Institute of Science Education and Research Pune, Pashan, Pune, India*
- LUDOVIC RENAULT • *The Netherlands Centre for Electron Nanoscopy (NeCEN), Leiden University, Leiden, The Netherlands*
- RUBEN SÁNCHEZ-GARCÍA • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- STEFAN SCHMELZ • *Department of Structure and Function of Proteins, Helmholtz Zentrum für Infektionsforschung GmbH, Braunschweig, Germany*
- MAREN SCHUBERT • *Department of Biotechnology, Institut fuer Biochemie, Biotechnologie und Bioinformatik, Technische Universitaet Braunschweig, Braunschweig, Germany*
- NEELADRI SEN • *Indian Institute of Science Education and Research Pune, Pashan, Pune, India*
- NEELESH SONI • *Indian Institute of Science Education and Research Pune, Pashan, Pune, India*
- CARLOS OSCAR S. SORZANO • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- MARLITT STECH • *Branch Bioanalytics and Bioprocesses (IZI-BB), Fraunhofer Institute for Cell Therapy and Immunology (IZI), Potsdam, Germany*
- DAVID STRELAK • *National Centre for Biotechnology (CSIC), Madrid, Spain*
- TAMARA TEICHMANN • *Branch Bioanalytics and Bioprocesses (IZI-BB), Fraunhofer Institute for Cell Therapy and Immunology (IZI), Potsdam, Germany*
- DEEPAK BALAJI THIMIRI GOVINDA RAJ • *Council for Scientific and Industrial Research CSIR, Pretoria, South Africa*
- LENA THORING • *Branch Bioanalytics and Bioprocesses (IZI-BB), Fraunhofer Institute for Cell Therapy and Immunology (IZI), Potsdam, Germany*
- SAMEER VELANKAR • *Protein Data Bank in Europe, European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK*
- SHENGQI XIANG • *NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands; MOE Key Lab for Membrane-less Organelles & Cellular Dynamics, School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, China*
- ANNE ZEMELLA • *Branch Bioanalytics and Bioprocesses (IZI-BB), Fraunhofer Institute for Cell Therapy and Immunology (IZI), Potsdam, Germany*

Part I

Structural Bioinformatics



Chapter 1

The Protein Data Bank Archive

Sameer Velankar, Stephen K. Burley, Genji Kurisu, Jeffrey C. Hoch,
and John L. Markley

Abstract

Protein Data Bank is the single worldwide archive of experimentally determined macromolecular structure data. Established in 1971 as the first open access data resource in biology, the PDB archive is managed by the worldwide Protein Data Bank (wwPDB) consortium which has four partners—the RCSB Protein Data Bank (RCSB PDB; rcsb.org), the Protein Data Bank Japan (PDBj; pdbj.org), the Protein Data Bank in Europe (PDBe; pdbe.org), and BioMagResBank (BMRB; www.bmrb.wisc.edu). The PDB archive currently includes ~175,000 entries. The wwPDB has established a number of task forces and working groups that bring together experts from the community who provide recommendations on improving data standards and data validation for improving data quality and integrity. The wwPDB members continue to develop the joint deposition, biocuration, and validation system (OneDep) to improve data quality and accommodate new data from emerging techniques such as 3DEM. Each PDB entry contains coordinate model and associated metadata for all experimentally determined atomic structures, experimental data for the traditional structure determination techniques (X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy), validation reports, and additional information on quaternary structures. The wwPDB partners are committed to following the FAIR (Findability, Accessibility, Interoperability, and Reproducibility) principles and have implemented a DOI resolution mechanism that provides access to all the relevant files for a given PDB entry. On average, >250 new entries are added to the archive every week and made available by each wwPDB partner via FTP area. The wwPDB partner sites also develop data access and analysis tools and make these available via their websites. wwPDB continues to work with experts in the community to establish a federation of archives for archiving structures determined using integrative/hybrid method where multiple experimental techniques are used.

Key words Protein Data Bank, Worldwide Protein Data Bank (wwPDB), Macromolecular structure archive, Deposition, Biocuration, and Validation system OneDep, Validation task forces, PDBx/mmCIF, X-ray crystallography, NMR spectroscopy, 3DEM, Integrative hybrid methods

1 Introduction

Established in 1971, the Protein Data Bank (PDB; pdb.org) is the single global archive of experimentally determined macromolecular structures [1]. It was the first open access digital data archive in the life sciences. Today, the PDB contains >175,000 structures of proteins, nucleic acids, carbohydrates, and their complexes. PDB

structure data contribute to mechanistic understanding of the function of biological macromolecules, and PDB is recognized as a core biological data resource [2]. Analyses of citations of PDB data demonstrate that structures in the PDB are extensively reused over many years after they are first made public [3, 4]. Recent analysis has also shown that open access to all structural data supports not only fundamental research and education in biology and medicine but also translational research, such as discovery and development of new drugs [5].

In recognition of the global nature of structural biology, the PDB is managed by an international consortium, the Worldwide Protein Data Bank (wwPDB; wwpdb.org; [6]) with four partners—the RCSB Protein Data Bank (RCSB PDB; rcsb.org; [7]), the Protein Data Bank Japan (PDBj; pdj.org; [8]), the Protein Data Bank in Europe (PDBe; pdbe.org; [9]), and BioMagResBank (BMRB; bmrb.wisc.edu; [10]). The wwPDB partners share a common vision to “Sustain freely accessible, interoperating Core Archives of structure data and metadata for biological macromolecules as an enduring public good to promote basic and applied research and education across the sciences.”

The PDB archival resource benefits the global structural biology (“depositor”) community by ensuring uniform and robust data preservation, attributing credit for data deposition even if it is not accompanied by a peer-reviewed publication, and housing data for development of better methods for structure determination and validation. Beyond structural biologists, the open access nature of the PDB benefits other consumers of macromolecular structures, such as the entire structural bioinformatics community and over 400 biomedical data resources that utilize macromolecular structures to derive biological insights. Basic and applied researchers worldwide, in fields as diverse as medicine, enzymology, polymer physics, mathematics, art, and education, access the PDB structures directly or via other resources. Engagement with user communities is thus central to the continued development of the PDB archive. The wwPDB consortium works with the structural biology and the broader user communities to develop policies and processes for data deposition, biocuration, validation, and data distribution.

2 History of Macromolecular Structure Archiving

Early efforts to elucidate three-dimensional structures of biological macromolecules led to the first structures of the DNA double helix [11], myoglobin [12], and hemoglobin [13]. From these very early days of structural biology, atomic coordinates were exchanged on an ad hoc basis with other researchers to facilitate better understanding of function at the molecular level. In 1971, discussions at the Cold Spring Harbour (CSH) Symposium on “Structure and

Function of Proteins at the Three Dimensional Level,” resulted in a formal proposal to establish a repository in the USA in collaboration with a team in the UK for archiving results of structure determination experiments. This led to the establishment of the PDB at the Brookhaven National Laboratory (BNL; [14]) and a collaboration with the Department of Chemistry at Cambridge University, UK, which housed the Cambridge Crystallographic Data Centre beginning in 1965 and had extensive experience in archiving crystallographic data for small molecules [15, 16]. Duplicate copies of the master PDB files were maintained at Brookhaven (USA), Cambridge (UK), and Tokyo (Japan), and in 1979, the data became available on magnetic tape from the Institute for Protein Research, Osaka University (Japan) [17]. It was appreciated from the outset that well-designed and well-documented data standards and formats were crucial to the success of the PDB, bringing about the development of the PDB format in 1972 [18]. Despite the limitations imposed by the punched card format, the original PDB format successfully served the community for four decades until its formal replacement by the PDBx/mmCIF format of the wwPDB.

From its inception, the community has played active roles in development of the PDB, and committees were established under the auspices of the International Union of Crystallography (IUCr) to define the minimum data content and policies for data deposition [19]. The IUCr guidelines were published in 1989 and mandated deposition of coordinates and experimental data to the PDB prior to publication [20]. Community action also prompted most major journals and funding agencies to adopt these guidelines, thus making structural data open access for the entire scientific community. Multiple mirror sites were established across the world to provide easy access to the structure data available in the PDB [21]. In 1996, the European Bioinformatics Institute (EMBL-EBI), an outstation of the European Molecular Biology Laboratory (EMBL) established the Macromolecular Structure Database (MSD), later rebranded as Protein Data Bank in Europe (PDBe), which set up a collaboration with the PDB at BNL for accepting depositions [22]. In 1999, management of the PDB was transferred from BNL to the Research Collaboratory for Structural Bioinformatics Protein Data Bank at Rutgers University (RCSB PDB; [23]). In 2000, the Protein Data Bank Japan (PDBj) was established by the Institute for Protein Research, Osaka University, to archive the data from the Structural Genomics (Protein 3000) project sponsored by the Japanese government. In 2002, EMBL-EBI established the Electron Microscopy Data Bank (EMDB; [24]) as a resource that archives the electric potential maps and associated metadata from electron microscopy experiments. Atomic coordinates derived from these maps continued to be deposited in the PDB. In 2003, RCSB PDB, PDBe, and PDBj established the wwPDB organization to manage the PDB archive as a single global

archive of macromolecular structure data [6]. In 2006, BioMagResBank (BMRB) joined the consortium [10, 25].

3 Role of the Worldwide Protein Data Bank in 3D Structure Data Archiving

The vision of the wwPDB consortium (*see* Subheading 1) is translated into practical steps through the Consortium's mission statement, which commits the partners to (a) managing the PDB archive as a public good according to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability [26]; (b) providing expert deposition, validation, biocuration, and remediation services at no charge to data depositors worldwide; (c) ensuring universal open access to public domain structural biology data with no limitations on usage; and (d) developing and promoting community-endorsed data standards for archiving and exchange of global structural biology data.

Community engagement is a major consortium activity. Over the intervening years, updates to the deposition guidelines were formulated by various task forces and subsequently endorsed by the wwPDB Advisory Committee: for example, deposition of structure factor data and NMR restraints became mandatory in 2008, NMR chemical shifts in 2010, and from 2016 deposition of atomic coordinates from 3DEM must be accompanied or preceded by the deposition of electric potential maps to EMDB. wwPDB was also instrumental in bringing together community experts in Validation Task Forces (VTFs) for X-ray [27], NMR [28], and 3DEM [29], who advised wwPDB on suitable metrics and software tools for validation of experimental data, atomic coordinates, and assessment of the fit between them. By 2016, these recommendations were largely implemented within the wwPDB validation pipeline [30], which is a component of the OneDep system for deposition, validation, and biocuration [31]. wwPDB has also hosted a number of workshops and meetings on specific issues: for example, the 2015 wwPDB/CCDC/D3R Ligand validation workshop provided valuable feedback on improving validation of bound ligands in the PDB [32]. These recommendations have recently been implemented in collaboration with Global Phasing, Ltd. [33] allowing wwPDB to distribute validation reports with richer and improved ligand information to depositors and other users of the PDB (Fig. 1).

To ensure continuous feedback from the community on developing data standards and archiving, the wwPDB has established a working group that brings together all the major software developers and experts who meet regularly to review the existing data standards and provide input on any changes/updates or additions required to meet the evolving requirements of structural biology community (<https://www.wwpdb.org/task/mmcif>). wwPDB also supported and collaborated with the NMR software community

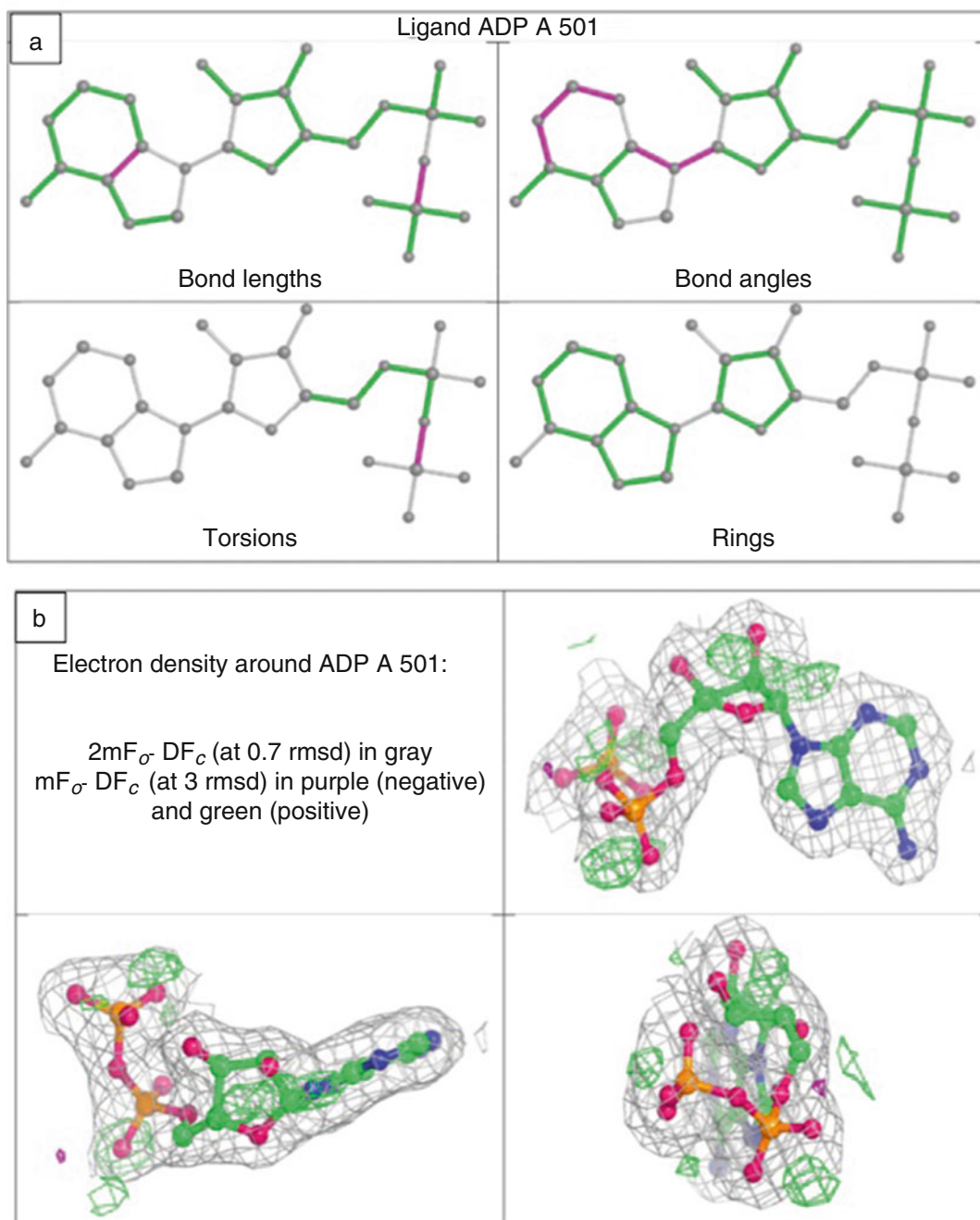


Fig. 1 Improved ligand validation identifies issues with ligands in PDB entries. **(a)**: an updated depiction of geometric quality from the Mogul software, highlighting bond length, bond angle, torsion angle, and ring outliers (purple). **(b)** electron density associated with the ligand shown along the three principle axes, and visualizing the experimental evidence supporting the placement and identity of the ligand. These images are generated for ligands identified as by depositors as “of interest,” i.e., usually having a biological role rather than facilitating crystallization

efforts to establish an NMR Exchange Format (NEF), a subset of the NMR-STAR format [34], as a robust mechanism to exchange NMR experimental data between different software packages and simplify deposition of these data to PDB and BMRB archives.

Anticipating the development of integrative/hybrid methods (I/HM), wwPDB organized a workshop that brought together a large number of community experts representing a variety of experimental techniques to discuss the archiving and validation requirements for coordinate models and associated experimental data when more than one experimental method contributes to structure determination. The outcome of the workshop addressed two main issues related to archiving of I/HM structure models and reviewed the state-of-the-art for validation of structure models, experimental data and fit between I/HM models and associated experimental data [35]. wwPDB partners are working closely with the community experts to implement these recommendations and have established a prototype system for deposition of I/HM models and associated experimental data [36].

4 Data Standards for 3D Biostructure Data Archiving

The PDB format was developed in the early 1970s to ensure that all the data in the archive are represented in a consistent and accurate fashion enabling the entire scientific community to exploit macromolecular structure data [18]. The original format was based on the then state-of-the-art standard used to store computer programs—the 80-column Hollerith format was used for punched cards. Although the PDB format served the community well, it imposed limitations on the maximum size of the structure model (e.g., 99,999 atoms and 62 polymeric chains) that can potentially be represented using this format. By the mid-1990s, structural biologists were depositing large structures that could not be represented by a single PDB format file [37]. These limitations were recognized by the early 1990s, and in order to circumvent them a new extensible format was proposed, the Macromolecular Crystallographic Information File (mmCIF), which was an extension of the CIF format and dictionary adopted by the small molecule crystallography community [38]. CIF and mmCIF are themselves based on the STAR framework [39]. The mmCIF framework is able to faithfully represent macromolecular structures and associated rich metadata [40–42] by describing the information through data items grouped into categories. Meaningful relationships are defined between the different categories, providing the necessary mechanism to test and impose data integrity. mmCIF dictionary definitions for each data item also include validation criteria, such as allowed ranges for numerical values or controlled vocabularies (enumerations) that can enforce a well-defined set of rules to test the values at the

time of deposition, ensuring that the archived data is as high quality as possible. In the early 1990s, the mmCIF dictionary contained data categories and items necessary for describing structures determined using X-ray crystallography, but over time data categories and items for metadata related to NMR and EM were added to the dictionary making the framework applicable to all entries in the PDB. To recognize this flexibility and applicability to other experimental techniques, the updated framework is called the PDB Exchange (PDBx) or PDBx/mmCIF dictionary. More recent additions to the dictionary include data categories for describing metadata related to small angle scattering experiments [43], I/H methods [44], and predicted structural models. At the time of writing, the public PDBx/mmCIF dictionary (Version 5) contains a total of 643 categories containing 6836 items spanning the description of the experimental sample (1117 items), model coordinates (2133 items), experimental data, and experimental setup for various supported techniques (1753 items for MX, 257 for solution and solid-state NMR, 1024 for EM), relationships to entries in other data resources, authorship and citation information, and the audit trail of changes made to the entry. The RCSB PDB currently acts as the archive keeper, ensuring disaster recovery of PDB data and coordinating weekly updates.

The PDBx/mmCIF data dictionary serves as a stable and extensible framework for the definition and representation of structural biology data. The wwPDB formally adopted PDBx/mmCIF framework in 2007 for internal use and in 2014 it became the official master format of the PDB archive, at which point the legacy PDB format was deprecated. The PDBx/mmCIF dictionary and format underpins the wwPDB global deposition, validation, and biocuration system, OneDep. The dictionary is publicly available (<http://mmcif.wwpdb.org/>). In addition to serving as a comprehensive framework for data archiving by wwPDB, in 2011 the PDBx/mmCIF format was adopted by developers of major macromolecular crystallography (MX) software packages as a vehicle for data exchange. Consequently, all major MX software packages for structure determination and refinement output richer and more complete deposition-ready PDBx/mmCIF formatted files. The PDBx/mmCIF dictionary is updated frequently to reflect the changes wrought by advances in structure determination techniques and refinement methods. To ensure the continued dialog with the community, such changes are presented to and endorsed by the PDBx/mmCIF working group, which oversees the development of the dictionary and whose membership includes developers of structural biology software packages and representatives of the wwPDB consortium. The PDBx/mmCIF data model is also translated into an XML schema allowing an XML-based representation of PDB archive data (PDBML) [45]. To facilitate semantic integration of macromolecular structure data with other biomedical data

resources, the PDB data is also represented in RDF (Resource Description Framework) [46].

In addition to 3D biostructure data, the PDB archive contains a number of reference data dictionaries, also represented with the PDBx/mmCIF framework. The Chemical Component Dictionary (CCD) [47] describes all unique chemical components observed in the PDB, including the essential chemical definition such as atom names, atom connectivity/bond order, and stereochemistry of each compound, as well as the systematic name, synonyms, chemical formulae, and standard structure descriptors, such as InChI and SMILES. Where available the CCD also includes idealized coordinates for the component. The Biologically Interesting Molecule Reference Dictionary (BIRD) [48] describes larger biologically relevant molecules, e.g., peptide containing antibiotics (e.g., vancomycin), and those formed when multiple CCD components are covalently connected to each other. In addition to describing such covalent connectivity and linking to the CCD, the BIRD dictionary includes common names and synonyms for such larger molecules and, where available, provides cross-references to other data resources where they may also be described.

In addition to the PDBx/mmCIF data dictionary described above, wwPDB also uses the NMR-STAR format [34] for representing NMR experimental data. 3DEM experiment results are archived and distributed by EMDB [49], which uses the EMDB-XML schema as the data model for metadata and CCP4 map format [50] for electric potential maps.

Increasingly, experimental data originating from more than one experimental technique are used to derive information about large macromolecular machines, including structural information across a wide range of spatial resolutions. Such integrative/hybrid structure determination approaches often result in multi-scale structural models, only parts of which may have sufficient experimental data yield atomic coordinates, while other parts of a structure may need to be modeled differently, e.g., as coarse-grained “beads” or low-resolution shapes/volumes. Archiving of such mixed, multi-scale models is a challenging task. Anticipating these developments in integrative/hybrid structure determination methods, following the guidelines from the wwPDB I/HM task force, the PDBx/mmCIF dictionary has been extended to include additional data categories and items to describe multi-scale models [44].

5 3D Biostructure Data Deposition, Biocuration, Validation, and Distribution

The wwPDB global deposition, biocuration, and validation system, OneDep, was launched in 2014 and provides a single global portal for deposition to the PDB and EMDB archives. To better support the user communities in different time zones, the OneDep system is

implemented at all the partner sites allowing for distribution of depositions based on geography with PDBj processing all the depositions originating from Asia and Middle East, RCSB PDB handling all the depositions from Americas and Oceania, and PDBe handling all European and African depositions.

The OneDep system is based on the extensible PDBx/mmCIF framework described above and supports deposition of structure data from all the experimental methods accepted by the PDB and EMDB archives. The extensibility of PDBx/mmCIF framework also provides a mechanism for extending support for new experimental methods in the future. As discussed above, the PDBx/mmCIF dictionary includes validation criteria for individual data items providing a mechanism to improve data quality and integrity for deposited data. The wwPDB biocuration team undertakes continuous review of the validation criteria in the PDBx/mmCIF dictionary and of the deposition and biocuration procedures with the view of improving both their efficiency and the quality of the PDB archive.

The wwPDB validation pipeline, which is integrated in the OneDep system and available as a standalone server (validate.wwpdb.org), is also accessible programmatically (<https://www.wwpdb.org/validation/onedep-validation-web-service-interface>) [30]. During deposition, depositors must review and accept a preliminary validation report prior to data submission and issuance of a PDB code. The validation reports are meant as a checkpoint to discover any major issues with the uploaded data, and if issues are identified, the depositor is encouraged to critically examine their uploaded data and, if needed, to upload revised files. Once the deposition is submitted, an appropriate accession code (PDB and/or EMDB) is assigned, and the entry is transferred for biocuration. The biocuration process [51] includes checking descriptions of all the chemical components to make them consistent with the CCD definitions. The biocurators review the sample description, including polymer sequences and the organism taxonomy and add cross-references to UniProtKB [52] and NCBI taxonomy [53] data resources. Added value annotations, such as secondary and quaternary structure and ligand binding sites, are derived and added to the entry.

In addition to standardizing the data representation and value-added annotations, the biocuration process also generates the official wwPDB validation report, which differs slightly from the preliminary one, as it takes into account, standardized nomenclature, which may not have been utilized prior to biocuration. Depositors are strongly encouraged to include these official wwPDB validation reports as part of manuscript submission to journals to assist referees in assessing the scientific results described in the publications. wwPDB also strongly encourages journal editors to make submission of these reports mandatory, and we are

grateful that a number of journals have followed this recommendation. Standardization of various validation metrics across the PDB archive enabled comparison and ranking of entries, which may be important when selecting a suitable dataset [54, 55] for a particular study. The recent update to the wwPDB validation software clearly identifies ligands that are not supported by electron density (Fig. 1b). Mandatory deposition of experimental data, the more widespread use of validation tools during structure refinement, and the wwPDB validation pipeline have contributed to a general trend of improved quality of structures in the PDB [30, 54, 55].

wwPDB is actively working with software developers and structural biology community to simplify the deposition process and make it more efficient by harvesting data, a task greatly facilitated by the adoption of the PDBx/mmCIF framework by MX structure determination and refinement software. Multiple software packages, including Phenix [56], CCP4 [57] and Global Phasing (<https://www.globalphasing.com/buster/>), already export PDBx/mmCIF format data and metadata. From July 2019 onwards, all OneDep MX structure depositions require upload of PDBx/mmCIF formatted atomic coordinate files [58]. We anticipate that in the near future enhanced data harvesting within these software packages will lead to a further improvement of PDB data completeness and quality. wwPDB are also working with the NMR and EM communities with the aim of adapting the corresponding software to allow export of deposition-ready PDBx/mmCIF formatted files for deposition via OneDep.

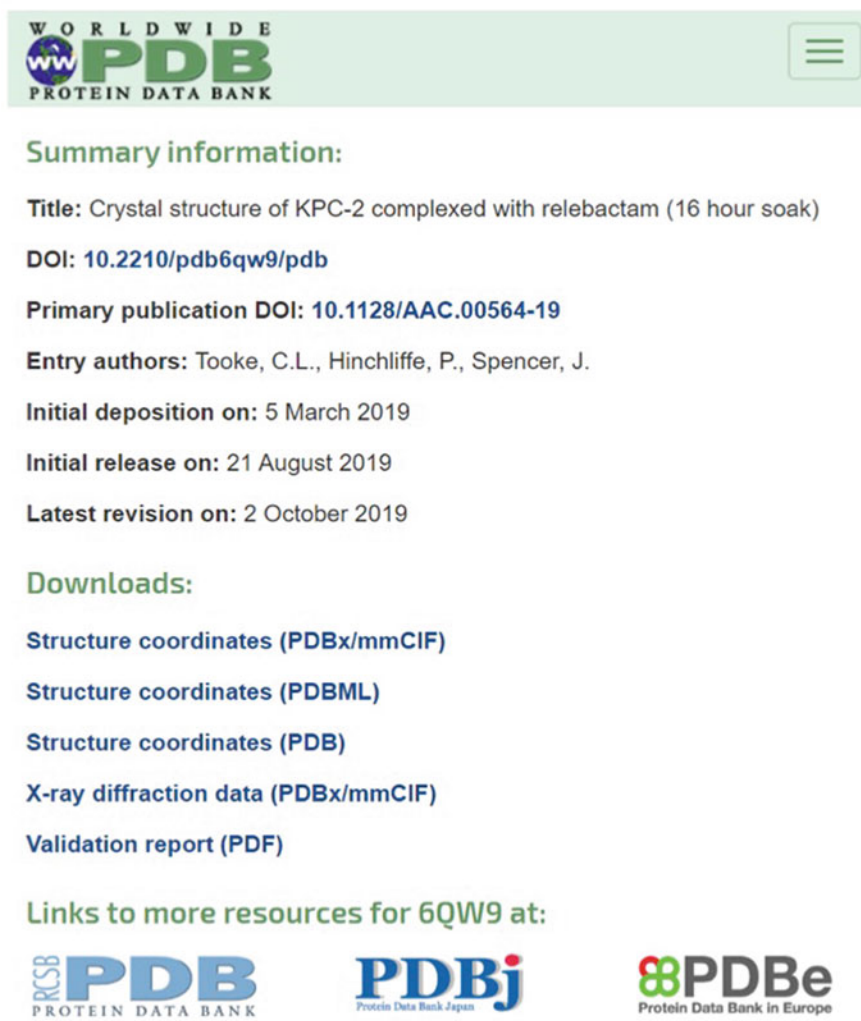
While small angle scattering data alone are insufficient to derive an atomic resolution model, it is often combined with experimental techniques that offer atomic or near-atomic resolution (i.e., MX, NMR, or 3DEM). In such multi-method approaches, small angle X-ray/neutron scattering (SAS) provides additional restraints to either help determine a de novo structure or to evaluate members of an ensemble of possible structures (e.g., [59]). It is also often used to verify if the quaternary assembly of the sample in its crystalline form is compatible with solution state [60]. SAS data are frequently archived by the Small Angle Scattering Biological Data Bank (SASBDB; sasbdb.org) [61], but deposition is not yet mandatory. To ensure accurate cross-referencing of SAS data used in the context of a PDB structure determination, wwPDB partners, and the SASBDB team have implemented a working system allowing deposition of SAS data to SASBDB during the course of a OneDep deposition session to the PDB and/or EMDB archives.

Rapid progress in structural biology and improvements in computational approaches often warrant re-examination of older experimental datasets to obtain a more complete or otherwise improved structure (e.g., better ligand geometry or fit to experimental data). In some cases, improved structures were not deposited to the PDB because of the prior policy requiring issuance of a

new PDB accession code if a new structure is deposited following original release, breaking the link to the peer-reviewed publication. To remedy this issue and to preserve the link to the original publication even if new atomic coordinates are deposited, wwPDB partners implemented versioning of PDB entries. This new OneDep capability allows depositors-of-record to update atomic coordinates of previously released PDB entries, while maintaining the same PDB accession code. To accommodate versioning, wwPDB partners maintain a separate PDB FTP area that serves all the major versions (i.e., affecting atomic coordinates, polymer sequences, and/or chemical representation) of PDB entries (<ftp-versioned.wwpdb.org>), while the more familiar FTP area (<ftp.wwpdb.org>) serves only the latest versions.

New and revised PDB entries are added to the wwPDB FTP site and to individual wwPDB partner FTP sites on a weekly basis (Wednesdays at 00:00 UTC), making the latest versions of all entries available to users. A detailed description of how to access and download the latest archive is available from the wwPDB website (<https://www.wwpdb.org/ftp/pdb-ftp-sites>). As part of commitment to supporting methods development and application efforts within the structural bioinformatics and cheminformatics communities, the weekly release process includes releasing a subset of PDB data each Friday, 4 days before the full release of the PDB archive. This advanced release subset includes amino acid or nucleotide sequence for each unique polymer molecule, the description of all new ligands in the form of InChI strings, and the crystallization pH value for each new entry, assisting various prediction challenges (e.g., CASP [62], CAPRI [63], CAMEO [64], and CELPP [65]) that support computational methods development.

The PDB archive has evolved from its inception in 1971. Today, the wwPDB provides, both atomic coordinates and rich metadata, associated experimental data, validation reports, and value-added data (e.g., quaternary assembly information and chemical reference data). Hence, the complete description of a PDB entry is no longer confined to the atomic coordinate model file and includes various additional data files. To facilitate access to all relevant files for any given entry, wwPDB registers Digital Object Identifiers (DOIs), which resolve to dedicated wwPDB web pages for each PDB entry. These pages show basic information about the entry, link to all relevant data and metadata files on the wwPDB FTP site and also link to web pages at the individual wwPDB partner sites, which offer further value-added information, visualization, and analysis tools (Fig. 2). Scientific journals and data resources are strongly encouraged to link to PDB data via the DOI mechanism outlined above.



WORLDWIDE PDB
PROTEIN DATA BANK

Summary information:

Title: Crystal structure of KPC-2 complexed with relebactam (16 hour soak)

DOI: [10.2210/pdb6qw9/pdb](https://doi.org/10.2210/pdb6qw9/pdb)

Primary publication DOI: [10.1128/AAC.00564-19](https://doi.org/10.1128/AAC.00564-19)

Entry authors: Tooke, C.L., Hinchliffe, P., Spencer, J.

Initial deposition on: 5 March 2019

Initial release on: 21 August 2019

Latest revision on: 2 October 2019

Downloads:

[Structure coordinates \(PDBx/mmCIF\)](#)

[Structure coordinates \(PDBML\)](#)

[Structure coordinates \(PDB\)](#)

[X-ray diffraction data \(PDBx/mmCIF\)](#)

[Validation report \(PDF\)](#)

Links to more resources for 6QW9 at:

[RCB PDB](#) [PDBj](#) [PDBe](#)

PROTEIN DATA BANK Protein Data Bank Japan Protein Data Bank in Europe

Fig. 2 The Digital Object Identifiers (DOIs) for PDB entries resolve to the newly developed wwPDB web pages providing access to all relevant files for the selected entry. The pages are designed to show basic information about the entry and link to all the wwPDB member sites that offer further value-added information, visualization, and analysis tools

6 State of the Protein Data Bank Archive

The PDB archive was established in 1971 with seven protein structures determined by X-ray crystallography. As other structure determination methods developed, 3D structures determined by NMR and 3DEM were also deposited to the PDB. Over the past 48 years, the PDB has grown steadily, reaching 10,000 structures in 1999, 100,000 in 2014 and exceeding 150,000 in 2019. Since 2015, more than 11,000 new structures are added to the PDB annually.

While the majority of structures in the PDB continue to be determined using X-ray crystallography (82% in 2019), other techniques play their role with a steady number of structures determined by NMR (3% in 2019), and a rapidly increasing number by 3DEM (14% in 2019), with a small percentage by a variety of other techniques such as neutron diffraction, electron crystallography, and others (<1%). Recent advances in 3DEM (e.g., direct electron detectors and new image processing and computational methods) stimulated an increase in the number of high-resolution (better than 4 Å) 3DEM structures deposited to the PDB (Fig. 3). Advances in 3DEM have also enabled routine studies of larger macromolecular machines, hitherto inaccessible to MX and NMR.

7 Distributed Data Dissemination and Value-Added Annotations

With more than two million daily structure data file downloads, information stored in the PDB is being used across the entire breadth of scientific research and education communities, literally from agriculture to zoology [4]. Virtually, all PDB data consumers are not experts in structural biology. Each wwPDB partner—RCSB PDB (rcsb.org), PDBj (pdbj.org), PDBe (pdbe.org), and BMRB (bmr.b.wisc.edu)—maintains an independent website and develops advanced visualization and analysis tools to enable these diverse user communities to access macromolecular structure data. In pursuing their goals of serving the users, each partner also undertakes data enrichment activities that add value (e.g., biological context annotations) or provide easy access to the PDB structures. One such effort, a collaboration between RCSB PDB, PDBe, and the Central European Institute of Technology (CEITEC) recently resulted in a launch of a common interactive 3D viewer for macromolecular structures, Mol-star (<http://molstar.org>) that integrates most of the features from commonly used web-based viewers NGL [66] and LiteMol [67]. Mol-star is also capable of displaying multi-scale models making it ready for the PDB to accept 3D structures produced by I/H methods. The Structure Integration with Function, Taxonomy, and Sequence (SIFTS) project [68], maintained by the PDBe and Protein Function teams at EMBL-EBI, provides residue level mappings between PDB and UniProtKB entries. SIFTS data facilitates transfer of annotations between protein sequences and protein structures. It is kept up to date with each release of the PDB and UniProtKB data resources. SIFTS data is shared with all the wwPDB partners and provides a mechanism for each site to integrate additional annotations, such as protein domains, or sites of post-translational modifications.

While the underlying PDB structures are identical at all wwPDB partners, for some use cases, value-added annotations available from one wwPDB partner may need to be combined

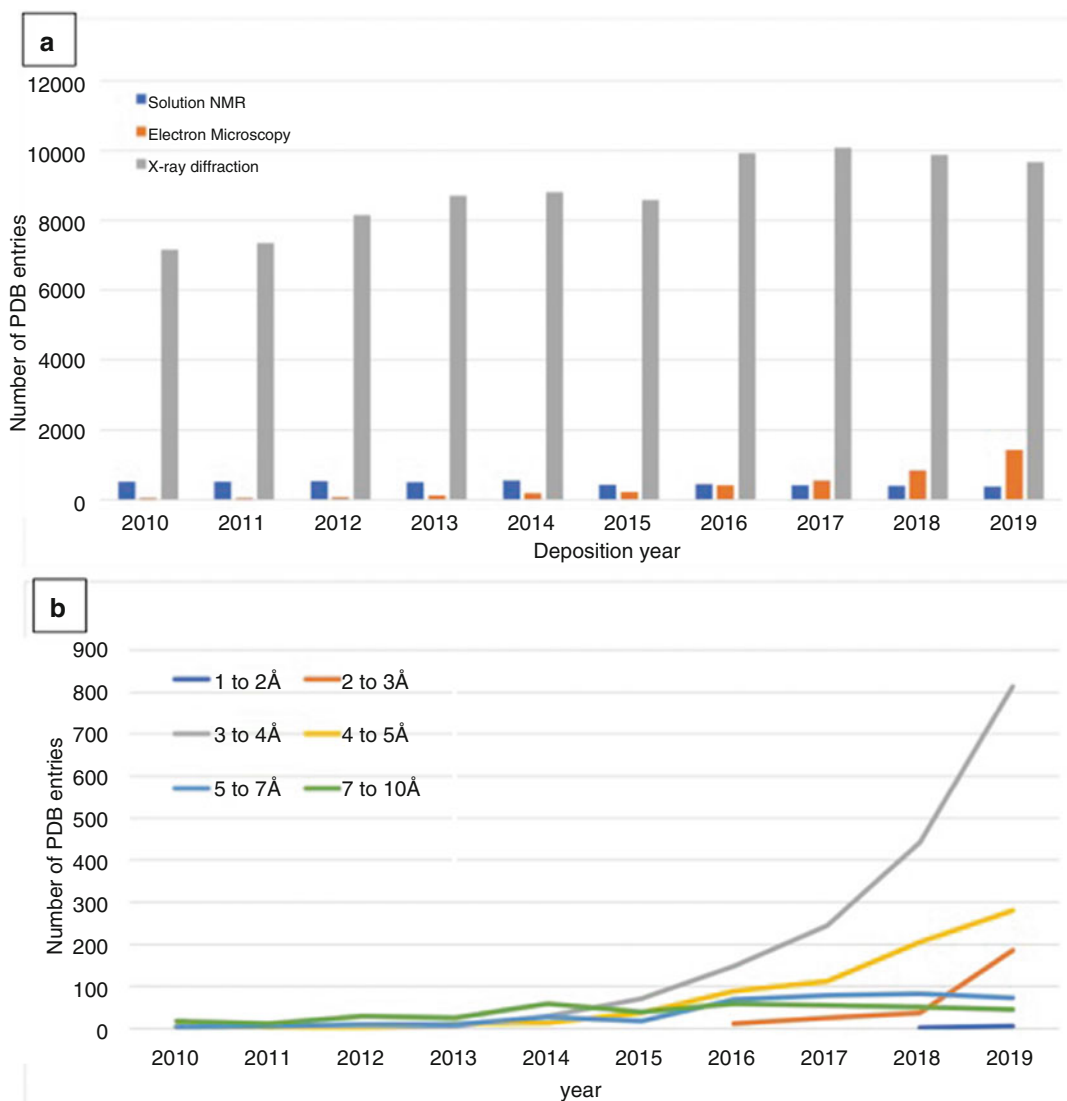


Fig. 3 Growth in cryo-3DEM. **(a)**: number of PDB structures deposited each year colored by the experimental technique used to determine the structure (MX—gray, NMR—blue, 3DEM—orange). **(b)** number of structures determined using 3DEM at different resolution ranges. As is evident from the plot, number of 3DEM structures determined at atomic resolution is increasing due to the introduction of direct electron detectors and rapid advances in the 3DEM structure determination methods

with data from another partner. Users may also be unaware of the various types of data available from each partner site. To help users to more efficiently access enriched, up-to-date annotations, wwPDB is developing a Next Generation PDB archive FTP area (NextGen PDB). NextGen PDB will combine all the information that is currently available in the PDB archive with added value annotations contributed by individual wwPDB partners. The

NextGen archive will thus make it easier to integrate structure data with information from other biomedical data resources.

8 Future of Structural Biology and the Role of the Worldwide PDB

Structural biology is witnessing rapid advances in experimental methods and the field now spans a broad length-scale range from atoms to individual proteins to molecular machines to organelles to cells and tissues. Advances in 3DEM and X-ray Free Electron Laser Serial Femtosecond Crystallography enable studies of multiple states or time-resolved behavior of biological macromolecules, providing insights into dynamics and molecular mechanisms of biological processes. These advances are complemented by novel molecular dynamics approaches, which utilize restraints derived from multiple experimental methods to improve *in silico* studies of 3D biostructure dynamics.

Advances in I/H methods, present unique challenges for archiving and validating the diverse experimental data and potentially multi-scale structural models with both atomic and non-atomic representation. To help address these emerging challenges, wwPDB established a Hybrid Methods Task Force and organized its inaugural meeting in 2014. Initial recommendations on data archiving and validation from this group were published as a white paper [35]. A follow-up meeting was arranged as a satellite to the Biophysical Society Meeting in 2019 [69]. One of the recommendations is to establish a novel, federated approach to I/H method data archiving. Experimental data would be deposited to the PDB (for MX), BMRB (for NMR), EMDB (for 3DEM), and federated specialist data archives, such as SBGRID [70], EMPIAR [71], PRIDE [72], or SASBDB [61]. This federated approach will ensure that the quality of experimental data quality is assessed by subject matter experts. In keeping with current best practices, the multi-scale I/H method structures would be deposited to the PDB. Data contributed to federated resources would be cross-referenced to each other to ensure that links between each I/H method structure and associated experimental data are preserved. This federated approach will support validation of I/H method structures against diverse aggregations of experimental data. Another recommendation was to develop a comprehensive data model to represent multi-scale structures. The initial meeting in 2014 was followed by development of a prototype system for archiving multi-scale models from I/H methods [36, 44]. This prototype (PDB-Dev: pdb-dev.wwpdb.org) allows wwPDB partners to gather a number of use cases, which will inform further extensions of the PDBx/mmCIF framework to support archiving of such datasets. wwPDB Hybrid Methods Task Force will continue

working closely with community experts on defining suitable validation metrics for each experimental modality.

In 2021, the structural biology community will achieve a major milestone of 50 years of continuously and consistently making their data open access through the PDB archive. The wwPDB consortium is organizing a number of meetings to celebrate the 50th anniversary of the PDB, including discussions on the future of the archive. During these celebrations, the wwPDB partners will reaffirm their commitment to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability [26] and to continued productive engagement with data depositors and data consumers worldwide.

Acknowledgments

The Protein Data Bank in Europe is supported by European Molecular Biology Laboratory-European Bioinformatics Institute; Wellcome Trust [104948]; Biotechnology and Biological Sciences Research Council [BB/G022577/1, BB/J007471/1, BB/K016970/1, BB/K020013/1, BB/M013146/1, BB/M011674/1, BB/M020347/1, BB/M020428/1, BB/P024351/1]; European Union [284209], ELIXIR, and Open Targets. The RCSB PDB is jointly funded by the National Science Foundation (DBI-1832184), the National Institutes of Health (R01GM133198), and the United States Department of Energy (DE-SSC0019749). PDBj is funded by the National Bioscience Database Center of Japan Science and Technology Agency (JST-NBDC), the Basis for Supporting Innovative Drug Discovery and Life Science Research of Japan Agency for Medical Research and Development (AMED-BINDS), and the Joint Usage / Research Center project assigned to Institute for Protein Research, Osaka University, by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. BMRB is supported by US National Institutes of Health (NIH) grant R01GM109046. We gratefully acknowledge contributions from John Berrisford, Aleks Gutmanas, Eldon L. Ulrich, Jasmine Young, and John Westbrook, and all wwPDB staff members present and past. We would like to acknowledge wwPDB collaborators and partners at the EMDB, SASBDB, CCP4, CCPEM, CCPN, and the global structural biology and bioinformatics communities.

References

1. wwPDB Consortium (2019) Protein data Bank: the single global archive for 3D macromolecular structure data jointly managed by the worldwide protein data bank. *Nucleic Acids Res* 47(D1):520–528
2. Durinx C, McEntyre J, Appel R et al (2016) Identifying ELIXIR core data resources. *F1000Res* 5. <https://doi.org/10.12688/f1000research.9656.2>
3. Bousfield D, McEntyre J, Velankar S et al (2016) Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources. *F1000Res* 5. <https://doi.org/10.12688/f1000research.7911.1>
4. Burley SK, Berman HM, Christie C et al (2018) RCSB protein data Bank: sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci* 27(1):316–330
5. Westbrook JD, Burley SK (2019) How structural biologists and the protein data bank contributed to recent FDA new drug approvals. *Structure* 27:211–217
6. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide protein data Bank. *Nat Struct Biol* 10:980
7. Burley SK, Berman HM, Bhikadiya C et al (2019) RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* 47:D464–D474
8. Kinjo AR, Bekker GJ, Wako H et al (2018) New tools and functions in data-out activities at protein data Bank Japan (PDBj). *Protein Sci* 27:95–102
9. Armstrong DR, Berrisford JM, Conroy MJ et al (2020) PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res* 48:D335–D343
10. Ulrich EL, Akutsu H, Doreleijers JF et al (2008) BioMagResBank. *Nucleic Acids Res* 36:D402–D408
11. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171:737–738
12. Kendrew JC, Bodo G, Dintzis HM et al (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181:662–666
13. Perutz MF, Rossmann MG, Cullis AF et al (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature* 185:416–422
14. (1971) Crystallography: protein data Bank. *Nat New Biol* 233:223–223
15. Kennard O, Watson DG, Town WG (1972) Cambridge crystallographic data centre. I. Bibliographic file. *J Chem Doc* 12:14–19
16. Groom CR, Bruno IJ, Lightfoot MP, Ward SC (2016) The Cambridge structural database. *Acta Crystallogr B Struct Sci Cryst Eng Mater* 72:171–179
17. The Protein Data Bank Newsletter Nr 10, Oct 1979 (1979) ftp://ftp.wwpdb.org/pub/pdb/doc/newsletters/bnl/news10_oct79.pdf
18. Bernstein FC, Koetzle TF, Williams GJ et al (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112:535–542
19. Berman HM (2008) The protein data bank: a historical perspective. *Acta Crystallogr A* 64:88–95
20. (1989) Commission on biological macromolecules. *Acta Crystallogr A* 45:658
21. Sussman JL, Lin D, Jiang J et al (1998) Protein data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* 54:1078–1084
22. Keller PA, Henrick K, McNeil P et al (1998) Deposition of macromolecular structures. *Acta Crystallogr D Biol Crystallogr* 54:1105–1108
23. Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. *Nucleic Acids Res* 28(1):235–242
24. Henrick K, Newman R, Tagari M, Chagoyen M (2003) EMDep: a web-based system for the deposition and validation of high-resolution electron microscopy macromolecular structural information. *J Struct Biol* 144:228–237
25. Markley JL, Ulrich EL, Berman HM et al (2008) BioMagResBank (BMRB) as a partner in the worldwide protein data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J Biomol NMR* 40:153–155
26. Wilkinson MD, Dumontier M, Aalbersberg IJ (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3:160018
27. Read RJ, Adams PD, Arendall WB et al (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure* 19:1395–1412

28. Montelione GT, Nilges M, Bax A et al (2013) Recommendations of the wwPDB NMR validation task force. *Structure* 21:1563–1570
29. Henderson R, Sali A, Baker ML et al (2012) Outcome of the first electron microscopy validation task force meeting. *Structure* 20:205–214
30. Gore S, Sanz Garcia E, Hendrickx PM et al (2017) Validation of structures in the protein data bank. *Structure* 25:1916–1927
31. Young JY, Westbrook JD, Feng Z et al (2017) OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure* 25:536–545
32. Adams PD, Aertgeerts K, Bauer C et al (2016) Outcome of the first wwPDB/CCDC/D3R ligand validation workshop. *Structure* 24:502–508
33. Smart OS, Bricogne G (2015) Achieving high quality ligand chemistry in protein-ligand crystal structures for drug design. In: Scapin G, Patel D, Arnold E (eds) *Multifaceted roles of crystallography in modern drug discovery*, Dordrecht, 2015. Springer, Netherlands, pp 165–181
34. Ulrich EL, Baskaran K, Dashti H et al (2019) NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. *J Biomol NMR* 73:5–9
35. Sali A, Berman HM, Schwede T et al (2015) Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure* 23:1156–1167
36. Burley SK, Kurisu G, Markley JL et al (2017) PDB-dev: a prototype system for depositing integrative/hybrid structural models. *Structure* 25:1317–1318
37. Jacobson RH, Zhang XJ, DuBose RF, Matthews BW (1994) Three-dimensional structure of beta-galactosidase from *E. coli*. *Nature* 369:761–766
38. Hall SR, Allen FH, Brown ID (1991) The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallogr A* 47:31
39. Hall SR (1991) The STAR file: a new format for electronic data transfer and archiving. *J Chem Inf Comp Sci* 31:326–333
40. Westbrook JD, Bourne PE (2000) STAR/mmCIF: an ontology for macromolecular structure. *Bioinformatics* 16:159–168
41. Fitzgerald PM, Westbrook JD, Bourne PE et al (2005) The macromolecular dictionary (mmCIF). In: Hall SR, McMahon B (eds) *International tables for crystallography*, vol G. International tables for crystallography. Springer, Dordrecht, pp 295–443
42. Westbrook J, Henrick K, Ulrich EL, HM B (2005) The protein data bank exchange dictionary. In: *International tables for crystallography*, vol G. Springer, Dordrecht, pp 195–198
43. Kachala M, Westbrook J, Svergun D (2016) Extension of the sasCIF format and its applications for data processing and deposition. *J Appl Crystallogr* 49:302–310
44. Vallat B, Webb B, Westbrook JD et al (2018) Development of a prototype system for archiving integrative/hybrid structure models of biological macromolecules. *Structure* 26:894–904
45. Westbrook J, Ito N, Nakamura H et al (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics* 21:988–992
46. Kinjo AR, Suzuki H, Yamashita R et al (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res* 40: D453–D460
47. Westbrook JD, Shao C, Feng Z et al (2015) The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the protein data Bank. *Bioinformatics* 31:1274–1278
48. Dutta S, Dimitropoulos D, Feng Z et al (2014) Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers* 101:659–668
49. Abbott S, Iudin A, Korir PK et al (2018) EMDB web resources. *Curr Protoc Bioinformatics* 61:5. 10 11–15 10 12
50. Winn MD, Ballard CC, Cowtan KD et al (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67:235–242
51. Young JY, Westbrook JD, Feng Z et al (2018) Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. *Database (Oxford)* 2018. <https://doi.org/10.1093/database/bay002>
52. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47:D506–D515
53. Sayers EW, Beck J, Brister JR et al (2020) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 48:D9–D16
54. Shao C, Liu Z, Yang H et al (2018) Outlier analyses of the Protein Data Bank archive using a probability-density-ranking approach. *Sci Data* 5:180293

55. Smart OS, Horsky V, Gore S et al (2018) Worldwide Protein Data Bank validation information: usage and trends. *Acta Crystallogr D Struct Biol* 74:237–244
56. Liebschner D, Afonine PV, Baker ML et al (2019) Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* 75:861–877
57. Pottertton L, Agirre J, Ballard C et al (2018) CCP4i2: the new graphical user interface to the CCP4 program suite. *Acta Crystallogr D Struct Biol* 74:68–84
58. Adams PD, Afonine PV, Baskaran K et al (2019) Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). *Acta Crystallogr D Struct Biol* 75:451–454
59. Lemak A, Wu B, Yee A et al (2014) Structural characterisation of a flexible two-domain protein in solution using small angle X-ray scattering and NMR data. *Structure* 22:1862–1874
60. Schlundt A, Tants JN, Sattler M (2017) Integrated structural biology to unravel molecular mechanisms of protein-RNA recognition. *Methods* 118:119–136
61. Kikhney AG, Borges CR, Molodenskiy DS et al (2020) SASBDB: towards an automatically curated and validated repository for biological scattering data. *Protein Sci* 29:66–75
62. Moulton J, Fidelis K, Kryshchukovych A et al (2018) Critical assessment of methods of protein structure prediction (CASP)-round XII. *Proteins* 86(Suppl 1):7–15
63. Lensink MF, Nadzirin N, Velankar S, Wodak SJ (2019) Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins*. <https://doi.org/10.1002/prot.25870>
64. Haas J, Gumienny R, Barbato A et al (2019) Introducing "best single template" models as reference baseline for the continuous automated model evaluation (CAMEO). *Proteins* 87:1378–1387
65. Wagner JR, Churas CP, Liu S et al (2019) Continuous evaluation of ligand protein predictions: a weekly community challenge for drug docking. *Structure* 27:1326–1335
66. Rose AS, Bradley AR, Valasatava Y et al (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics* 34:3755–3758
67. Sehnal D, Deshpande M, Varkova RS et al (2017) LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. *Nat Methods* 14:1121–1122
68. Dana JM, Gutmanas A, Tyagi N, Qi G et al (2019) SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res* 47:D482–D489
69. Berman HM, Adams PD, Bonvin AA et al (2019) Federating structural models and data: outcomes from a workshop on archiving integrative structures. *Structure* 27:1745
70. Morin A, Eisenbraun B, Key J et al (2013) Collaboration gets the most out of software. *elife* 2:e01456
71. Iudin A, Korir PK, Salavert-Torres J et al (2016) EMPIAR: a public archive for raw electron microscopy image data. *Nat Methods* 13:387–388
72. Perez-Riverol Y, Csordas A, Bai J et al (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 47:D442–D450



Chapter 2

Computational Methods for the Elucidation of Protein Structure and Interactions

Nicholas S. Edmunds and Liam J. McGuffin

Abstract

Biologists are increasingly aware of the importance of protein structure in revealing function. The computational tools now exist which allow researchers to model unknown proteins simply on the basis of their primary sequence. However, for the non-specialist bioinformatician, there is a dazzling array of terminology, acronyms, and competing computer software available for this process. This review is intended to highlight the key stages of computational protein structure prediction, as well as explain the reasons behind some of the procedures and list some established workarounds for common pitfalls. Thereafter follows a review of five one-stop servers for start-to-finish structure prediction.

Key words Tertiary structure, Homology modeling, Template-based, Template-free, Sequence, Alignment, Refinement, Quality assessment, Docking, Quaternary structure

1 Introduction

Understanding macromolecular 3-D structure remains a major ambition for molecular biologists. This is due, not only to the therapeutic potential offered by nucleic acid–protein and protein–ligand interactions as new medicinal drug targets, but also to many wider applications of protein structure knowledge including agricultural crop improvement or even biofuel development [1].

Computational or *in silico* methods for the determination of protein structure are becoming ever more widespread and important in fulfilling this ambition. This is fundamentally the consequence of two phenomena: firstly, that the ability to elucidate protein sequences from genomic information continues to outpace the capability of experimental methods to determine the structure of these newly sequenced proteins [1], despite advances in X-ray crystallography technique and improvements of NMR and cryo-EM accuracy and resolution; and secondly, the continuing assertion that structure implies function in protein biology and that, in turn, sequence determines structure. Therefore, the sequence to

structure gap continues to grow and manual experimental techniques are unlikely to close this in the near future [2].

Since the creation of the Protein Data Bank (PDB) [3] in 1971, there has been an increasing reliance on curated sequence and structural repositories by the molecular biology community. Furthermore, along with community-wide experiments such as CASP (Critical Assessment of techniques for protein Structure Prediction) and CAPRI (Critical Assessment of the PRediction of Interactions)—see Subheading 6 for more details, growth in the area of *in silico* methods has led to an explosion in predicted protein structures [4]. This has mainly occurred through the rise of homology (or template-based) modeling and has in turn driven the associated proliferation of prediction software and data repositories, which are now available to research communities via the internet.

In this chapter, we will attempt to explain some of the main techniques used in 3-D protein structure prediction along with decoding a number of acronyms commonly encountered within the field; and secondly, to clarify the wide array of software packages and databases that now exist and, in the process, reference and analyze some key representative examples.

2 A Brief Summary of Protein Classification and Data Repositories

Proteins can be classified in a number of ways; in terms of primary structure or sequence similarity; secondary structure and associated motifs; tertiary structure and associated folds and domains and an emerging categorization based on protein–protein interactions (PPI) [1]. In addition, and perhaps related more closely to secondary structure classification than any of the others, is the grouping of proteins into classes and families on the basis of evolutionary relationship. The following describes a little about resources that fall into these classification categories.

In the case of primary structure, there are a number of databases containing information on amino acid sequences of which probably the most important from a structural prediction point of view is the Protein Knowledge Base—UniProtKB/TrEMBL [5]. This vast protein sequence database consists of the Universal Protein Resource (UniProt from PIR) which evolved from the early manually annotated SWISS-Prot sequence database (1986) allied to the automatically annotated TrEMBL sequence database administered by the European Bioinformatics Institute (EBI). The resource also contains UniRef a clustering service which lists groups of related sequences together and UniParc, an additional development intended to represent a complete and comprehensive non-redundant database of all known protein sequences with each sequence listed only once with a unique identifier (see Table 1).

Table 1
Protein sequence databases

Name	Description	Website
UniProtKB [5]	Repository for sequence, taxonomy, annotation, ontology, and classification information including TrEMBL (automatically annotated sequences)	www.uniprot.org/help/uniprotkb
UniParc [5]	Non-redundant database of all known protein sequences	www.uniprot.org/help/uniparc
UniRef [5]	Clustering service of related sequences	www.uniprot.org/help/uniref

Tools for assessing sequence similarity and alignment based on sequence database searches are discussed in Subheading 4.1 below.

Information on classifying proteins according to secondary structure is most easily obtained from the structural classification repositories [1]; Pfam [6] (from the EBI, classifies proteins into families based on domain similarity), SCOP [7] (Structural Classification Of Proteins—classifies into family, superfamily, and fold similarity), and CATH [8] (from UCL, classifies proteins into class, architecture, topology, and homologous families on the basis of domain similarity) and each of these has a website with full information on their classification system and how best to interpret it. These databases contain a great deal of evolutionary and relationship information as well as links to other software and are widely referenced by many 3-D prediction algorithms.

For novel protein sequences whose structures are not recorded in any existing database, the most widely accepted methods of secondary structure prediction (also referenced below) are those based on the Dictionary of Protein Secondary Structure algorithm (DSSP) [9] and these include PSIPred [10] and JPred4 [11] although it is possible to find many others via links within the ExPASy Bioinformatics Resource Portal.

The major resource for known tertiary structure information is, of course, the PDB (Protein Data Bank) [12] although a number of alternative databases can be found including those at the NCBI and EBI webpages (see Table 2). These have links to many classification and prediction resources. Again, the SIB (Swiss Institute of Bioinformatics) resource portal ExPASy may be useful with links to nextProt [14] (a human protein knowledge base), STRING [15] as well as Swiss-Model [16] (see Subheading 5.5).

Probably the most comprehensive quaternary and protein–protein interaction database is PDBe-PISA [13] (Proteins Interfaces, Structures, and Assemblies) that is hosted by the EBI although SIB’s SMTL (Swiss-Model Template Library) [16] and STRING are also useful for studying interactions and networks.

Table 2
Protein structure and classification databases

CATH [8]	Structural classification into class, architecture, topology, and homology	www.cathdb.info/
Pfam [6]	Protein family classification (EBI)	https://pfam.xfam.org/
SCOP [7]	Structural classification of proteins	http://scop.mrc-lmb.cam.ac.uk/scop/
PDB [3]	The protein data bank, from wwPDB, a collaboration of PDBe (UK), PDBj (Jpn), and BMRB (US)	www.rcsb.org/pdb
PDBe-PISA [13]	Proteins, interfaces, structures, and assemblies database for protein–protein interactions and quaternary structures	www.ebi.ac.uk/msd-srv/prot_int/pistart.html
nextProt [14]	Human protein knowledge base	https://www.nextprot.org/
STRING [15]	Alternative protein–protein interaction knowledge base from the SIB	https://string-db.org/

3 Types of Structure Prediction; Comparative Versus Ab Initio Modeling

The most successful form of structure prediction to emerge over the last 25 years is comparative modeling [12]. At its most basic, this is the process of modeling a protein with an unknown tertiary structure on the basis of sequence similarity to those with known structures.

Proteins that have a matching sequence (sequence identity above 30% as a rule of thumb) [17] are deemed homologs and can be used as templates on the presumption that sequence similarity suggests a common functional evolutionary ancestor. A similar structure can therefore be inferred from a similar sequence.

This approach is known variously and almost interchangeably as Comparative Modeling (CM), Homology Modeling (HM), and Template-Based Modeling (TBM) (although true homology modeling relies on an established evolutionary relationship between proteins rather than just a distant sequence similarity or shared domain). For the rest of the chapter, we will refer to this process as Template-Based Modeling or TBM.

Ab initio modeling, on the other hand attempts to use the so-called physics-based rules and routine, e.g., torsion angles in the protein carbon backbone, hydrophobicity ratings, bond length calculations, and van der Waals interactions, to predict the folding and hence tertiary structure of a protein from sequence alone, i.e., without comparison with a template [18]. This is often alternatively

termed *de novo* modeling, although, strictly speaking *de novo* modeling may include some type of sequence fragment check against a database whereas true *ab initio* techniques should model from sequence alone. A complication that might be encountered is that a number of programs now include a certain level of *ab initio* modeling embedded within their TBM calculations (e.g., the Rosetta algorithm [2, 19]) or to help resolve unstructured parts of the suggested model (e.g., Phyre2 [20]). However, there are other programs that offer a complete *ab initio* modelling service (e.g., QUARK, FALCON as well as ROSETTA).

The following sections will concentrate on describing TBM only, as this is likely to be the most useful route for the general molecular biologist who is not part of a specialist protein modeling group, and the technique is applicable to the majority of new protein targets.

4 Stages in Template-based Modeling (TBM)

TBM is a multi-step process [1], often made to appear seamless by publicly accessible webserver programs (see Table 3 below for a list). However, the identification of suitable homologs to use as templates is often not an insignificant task, and there are a number of technical solutions employed across various platforms to ensure that the templates used in model building are as relevant as possible. Another problematic stage in the modeling process is the sorting, scoring, and ranking of the (often) many alternative models (termed decoys) that are built [2]. These two stages remain the greatest challenge in TBM with the latter potentially more challenging than the former due to the nature of selecting the closest model to the native protein whose structure is unknown.

Rangwala and Kapris, 2010 [1] split the process of TBM (comparative modeling in their review) into five distinct stages: Selection of templates, Alignment of sequences, Model building, Quality evaluation, and Refinement, and in the following sections we have highlighted a similar but updated sequence of events routinely used by the protein modeling community.

The flowchart below gives an overall guide to the way the sequence fits together and the decision points that drive the process. It must be noted, however, that these stages are in-built and often invisibly merged in most public webserver making it unclear which distinct stage is being carried out at any one time. For those wishing to perform TBM in a more hands-on manner, there are specialist programs which can be downloaded and run separately from many of the website listed in Table 3, but for most non-specialist bioinformaticians these sections represent background information as the majority of your modeling needs will be catered for by using the full structure prediction webserver described in Subheading 5.

Table 3
Tertiary structure prediction tools

IntFOLD [21]	A high-performance server developed by the McGuffin group, offering a suite of programs for tertiary and quaternary predictionSpecializing in model quality assessment	https://www.reading.ac.uk/bioinf/IntFOLD/
I-TASSER [22]	A powerful threading-based online server offering a number of services in addition to modeling	https://zhanglab.ccmb.med.umich.edu/I-TASSER/
MODELLER [23]	Downloadable program for 3-D structure prediction. Users must provide their own alignment data	https://salilab.org/modeller/
MULTICOM [24]	Part of an online toolbox for structure prediction hosted by the university of Missouri	http://sysbio.mnet.missouri.edu/multicom_cluster/
Pcons [25]	Online server specializing in quality assessment (Stockholm university)	http://pcons.net/
Phyre2 [20]	Online full-service server from the structural bioinformatics Group at Imperial College part of Genome3D	http://www.sbg.bio.ic.ac.uk/phyre2
Predict protein (PP) [26]	Developed by RostLab (university of Munich) offering full prediction service	https://www.predictprotein.org/
RaptorX [27]	Online server (Xu group, University of Chicago), specializing in predicting sequences with no close homologs	http://raptorx.uchicago.edu/
ROBETTA [28] (Rosetta [19, 29])	Online server (Baker lab, University of Washington) full structure prediction using the powerful Rosetta algorithm	http://robeta.bakerlab.org/
SWISS-MODEL [16]	Comprehensive online server; both tertiary and protein interaction prediction by the SIB (Swiss Institute of Bioinformatics)	https://swissmodel.expasy.org/

4.1 Sequence Alignment and Template Identification

The initial task is that of identifying one or more suitable homologs to use as templates on which to base the model (see Table 4 for a list of programs). The amino acid sequence of the protein of interest, the target protein, will be run against a database of sequences, often the UniprotKB or a non-redundant derivative thereof. Here, the first problem is encountered; evolutionarily related proteins often have a greater level of structure conservation than sequence conservation [20]. Therefore, it is possible that simply aligning the whole of your target sequence against a sequence from another protein will produce a poor match. Most sequence alignment programs (e.g., Uniprot-align [5] and PSI-BLAST [30]) will therefore attempt local sequence alignment where sequences are cut into sections that are then cross-aligned [35]. The rationale is that protein domains may swap places over time and therefore one needs to search the whole sequence for matches rather than a simple pairwise comparison. Even with successful alignments there is a high probability of missing sequence sections (deletions), additional sections (insertions), and substitutions where amino acids have been replaced with others. For this reason, sequence alignments are scored from a BLOSUM matrix [18] that attempts to give good scores for amino acid conservation or replacement in non-structured parts of the protein (loop regions) and penalties for missing sections or replacement of amino acids in ordered secondary structure regions. A number of programs will also employ a secondary structure consensus check between target and templates at this stage [20] to increase confidence in final template selection, a popular choice of program being PSIPRED (UCL).

Table 4
Protein sequence search and alignment tools

BLAST [30]	Basic local alignment tool (also see PSI-BLAST a more sensitive version)	https://blast.ncbi.nlm.nih.gov/
ClustalW [31]	Multiple sequence alignment using traditional sequence profiling	https://embnet.vital-it.ch/software/ClustalW.html
Clustal Omega [32]	Multiple sequence alignment tool using HMM profiling	https://www.ebi.ac.uk/Tools/msa/clustalo/
EMBOSS [33]	Global alignment (needle option) and local alignment (water option)	https://www.ebi.ac.uk/Tools/psa/emboss_needle/emboss_water/
FASTA [33]	A simple local alignment tool	https://www.ebi.ac.uk/Tools/sss/fasta/
HH-blits [34]	Popular hidden Markov model (HMM) alignment site	https://toolkit.tuebingen.mpg.de/tools/hhblits
HMMER [34]	Sequence search tool using hidden Markov models (HMM) prediction	http://hmmer.org/ (to download) https://www.ebi.ac.uk/tools/hmmer/search/phmmer (online)

4.2 Loop Identification and Side-Chain Packing

Many homologous proteins will share not only a certain agreement in sequence identity but also in secondary structure, folds, and overall configuration. However, it is quite frequent for related proteins to differ in the length of the unstructured loop regions that connect secondary structure as well as the order of the individual folds or domains. For this reason, researchers have often been obliged to take the extra step of loop building in order to account for longer or shorter unstructured regions between folds. Many contemporary programs now include loop building as an automatic function [20], but optimization of loops and unstructured regions still occurs in refinement programs (see below). Side-chain packing is another element of model building which has become absorbed into the regular functioning of modern modeling programs [36], but which is still an important part of refinement procedures. Often the last part of refining a model will be to assess clashes or unlikely contacts between amino acid side chains and attempt to modify angles and residue positions slightly in order to resolve these.

4.3 QA and Ranking Models

Once models are constructed by the modeling software the importance of assessing their quality is necessary for two reasons. The first, which is discussed further in the following section, is to rate the models on general agreement with known protein structures, in other words, have you built a native-like potentially functional model or is it so far beyond acceptable structural limits as to be unlikely to exist? The second is the task of assessing which of your models matches your protein's native structure the best and therefore should be at the top of your ranking list.

In general, single-model quality assessment methods (those assessing each model individually) employ a number of physical checks to assess the models' structural integrity. These range from residue environment compatibility, e.g., hydrophobicity and solvent accessibility to structural features, such as secondary structure compatibility and assessment of backbone torsion angles [12]. Users are then presented with scores showing how well the model conforms to hypothetical 3D norms. One problem that must be borne in mind when interpreting these plausibility checks is that a model may score well because it conforms to pre-programmed ideals and so be ranked above a model which displays some structural defects but nevertheless is much closer to the native structure.

The second issue of ranking models may be relatively simple if all that was required was to select the best model on the basis of its resemblance to the template. However, with lower sequence identities the key question becomes, how closely does resemblance to the template suggest closeness to the native structure? Ranking models' resemblance to a native structure that is unknown will always be a subjective process and so consensus assessment has been developed in an attempt to overcome this.

Consensus methods use scores from a number of different programs, and many include a clustering stage in which models are clustered together on the basis of structural similarity, selecting those that lie close to the largest clusters. Consensus assessment can often out-perform single methods, with clustering working well when templates and models show a close structural relationship [37]. However, if there is a large variability in templates leading to a significant number of low-quality models or very few models in the first place, clustering and consensus methods that include them can prove less reliable.

As can be imagined, the distinction between the disciplines of assessment for ranking and final model quality assessment has become blurred and the processes now overlap somewhat.

Model quality is, to a large extent, dependent on the evolutionary distance between the target protein and the template(s) used to model it [1]. When working with low sequence identity, target-template 3-D similarity naturally decreases meaning that models may contain significant errors. As stated, model quality assessment assigns a predictive score to a model [12] in an attempt to rate its accuracy or similarity to the native protein prior to any confirmatory experimental structure being available and over the years a number of approaches have been developed.

Early versions of quality checks focused on stereochemical calculations measuring, amongst others, bond angles, steric clashes, and Ramachandran outliers. Others were based on calculating an energy score based on the model's perceived distance from a hypothetical free energy minimum. The so-called energy function checks fell broadly into two groups: those calculating a statistical score by analyzing the model against known protein structures and those calculating an empirically derived energy score from force field and molecular dynamic data. The shortcomings of these quality checks were, as mentioned before, that models could have perfectly reasonable stereochemical profiles and a low energy conformation but neither guaranteed similarity to the unknown native structure.

Current MQAPs (a selection listed in Table 5) attempt to overcome these shortcomings by combining a number of approaches. Firstly, as well as giving a global score for the overall model many programs will also give a local, or per residue score which assesses each amino acid residue and the favorability of the surroundings in which it finds itself in the proposed chain (factors like solvent accessibility, secondary structure compatibility, and side-chain contacts may be assessed). Secondly, in addition to basic stereochemical checks and energy considerations most MQAPs will perform a clustering routine [37] where potential models (decoys) are clustered on the basis of their conformation similarities. Models representative of large clusters are assumed to have a higher likelihood of resembling the native structure than remote models. Lastly, to increase the statistical confidence of the

Table 5
A selection of Model Quality Assessment Program servers (MQAPs)

ModFOLD6 [21]	A resource for estimates of model accuracy (EMA), using a hybrid quasi-single model approach	https://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD6_form.html
PCons [25]	Analyses models for recurring 3-D structural patterns and assigns a commonality score	http://pcons.net/index.php?about=pcons
ProQ3 [38]	Based on Rosetta, including all-atom (ProQRosFA) and centroid (ProQRosCen) energy functions	http://proq3.bioinfo.se/
QMEAN [39]	The sum of four measures; backbone torsion angles, C β interactions, all atom interactions, solvation score	https://swissmodel.expasy.org/qmean/

final score, neural networks can be used to perform an all-against-all comparison of conformations and then calculate a probability score [12]. The advantage of using neural networks is not only their ability to handle vast amounts of data but also the ability to train the networks to recognize native conformations from decoys using a training set of experimentally solved structures.

See Section 8 Notes (Table 11) for a table of scores commonly encountered with model quality assessment, refinement, and ranking output.

4.4 Refinement

Refinement is the process of taking a raw model and attempting to improve its quality score by making small changes to the 3D structure in the hope and expectation that the newly produced model will be closer to the native protein than the original. Refinement programs essentially perform two separate functions; the first is one of sampling, that is, to create improved 3D models from those already built by the modeling software (often by MD employing the AMBER or CHARMM force fields) and the second is one of scoring these models, mostly via energy functions (such as DFIRE, RWPlus, and Rosetta), so that improvements can easily be identified [36]. It is in the second function that refinement programs overlap significantly with model quality assessment programs and the process of MQA and refinement can often be iterative as shown in Fig. 1 below.

As well as performing two functions, refinement programs can be broadly split into two types. First are those, sometimes referred to as manual programs, which perform very computationally intensive functions such as molecular dynamics (MD) and Monte Carlo statistical simulations and may also be augmented by applying knowledge-based constraints. These tend to be programs available to download and run locally in Linux or available to run from specialist research groups who complete in the CASP experiments. Second are the automated server-style programs that are available

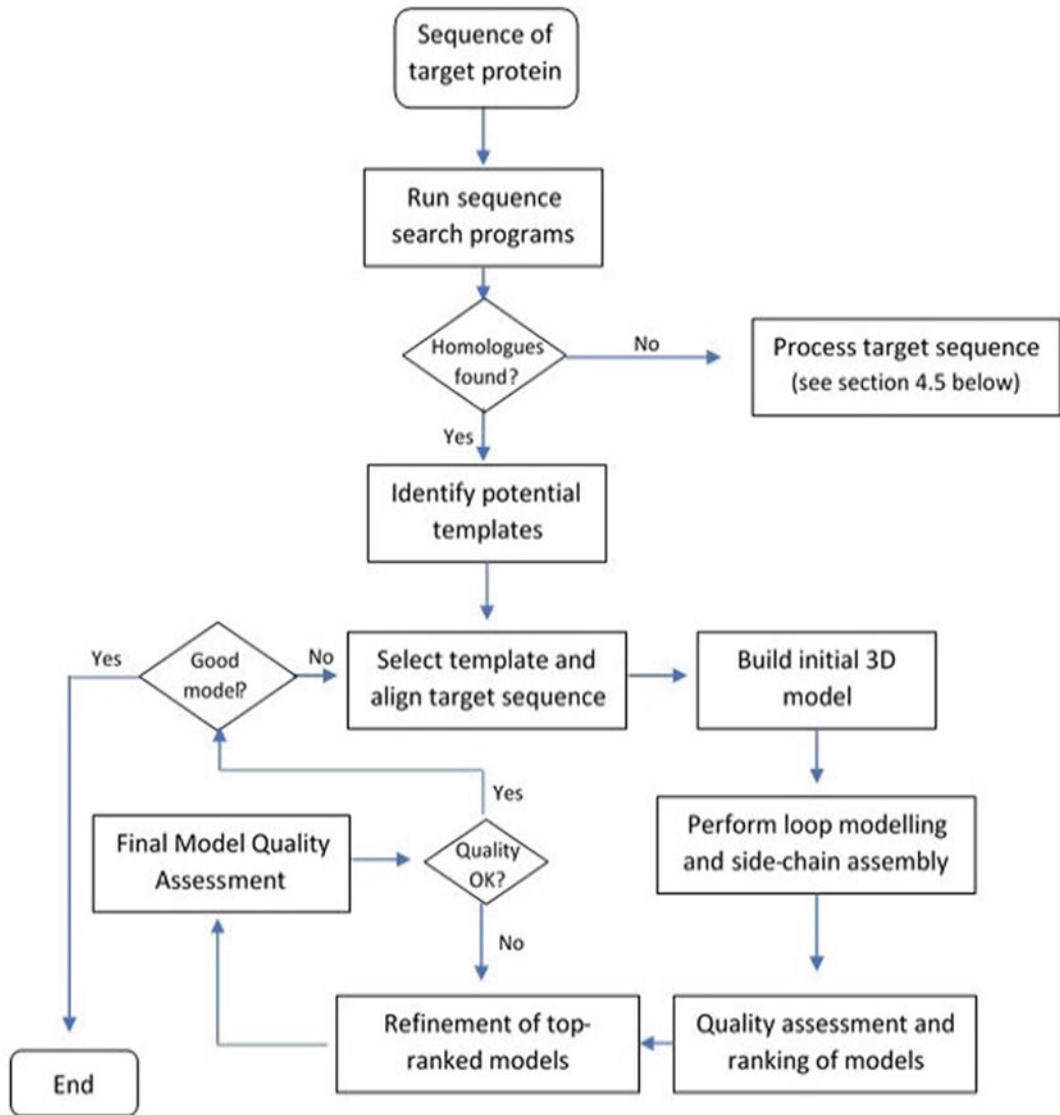


Fig. 1 A flow chart of the key stages in template-based modeling

via public webpages. These tend to be quicker and focus more on computationally less-intensive methods such as side-chain optimization and less stringent energy minimization functions [40]. The second group tend to make more conservative changes to the models, which is often desirable if the models are of reasonably good quality in the first place. Table 6 lists a number of publicly available refinement servers.

Table 6

Publicly available refinement webserver. (Reproduced from *Methods for the Refinement of Protein Structure 3D Models*, 2019 (Adiyaman R and McGuffin LJ) with permission from International Journal of Molecular Science [36])

PREFMD [41]	Developed by the Feig group, based on molecular dynamics (MD)	http://feiglab.org/prefmd
locPREFMD [42]	As above but focussed on local (per residue) quality	http://feig.bch.msu.edu/web/services/locprefmd/
GalaxyRefine [43]	From the Seok group, focused on side-chain repacking	http://galaxy.seoklab.org/refine
KoBaMIN [44]	Energy minimization strategies using a knowledge-based force field	http://csb.stanford.edu/kobamin
Princeton TIGRESS 2.0 [45]	Combines many strategies from other servers, scored well in CASP experiments	http://atlas.engr.tamu.edu/refinement/
ModRefiner [46]	Multi-step algorithm for side-chain optimization with physics and knowledge-based force fields	http://zhanglab.ccmb.med.umich.edu/ModRefiner
3DRefine [47]	Optimization of H-bonds and energy minimization with physics and knowledge-based force fields	http://sysbio.rnet.missouri.edu/3Drefine/
ReFOLD [48]	A quasi single-model approach with H-bond optimization and MD, using ModFOLD, from the IntFOLD server	http://www.reading.ac.uk/bioinf/ReFOLD/
FG-MD [49]	MD-based algorithm using TM-align to identify analogous fragments from the PDB	http://zhanglab.ccmb.med.umich.edu/FG-MD/

4.5 What if your Model Is Not a Good One?

If your model does not score well when subjected to quality assessment programs and attempted refinement, then it is likely that the template, on which it is based, is not a good match. Checking back to the flow chart in Fig. 1, we can see that problems may become obvious much earlier than this if there are few or no homologs identified for your target sequence. In either case, there are a number of avenues that may lead to an improvement in the model quality. These are summarized in Fig. 2 below and the following sections where one or more options may be necessary.

4.6 Disorder and Secondary Structure Prediction

One possible reason that your chosen modeling software fails to produce a good model of your target protein may be that it contains some intrinsically disordered regions (IDRs). Many proteins contain flexible regions in place of well-defined secondary structure [50], and these regions have been linked with a number of functions including recognition and binding of ligands and DNA,

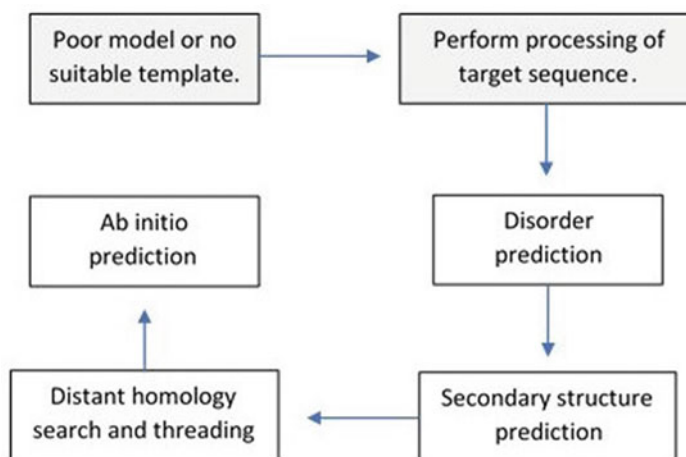


Fig. 2 A flow chart showing some alternative modeling strategies

signaling and cell cycle control or even potential phosphorylation sites. In many of these cases, the phenomenon of disorder-to-order is only observed upon binding and so the protein, in its native-unbound state, will be unlikely to comply with programmed expectations of 3D structure. Disorder prediction may therefore give some clues as to why models are poorer than expected.

In a similar way, it may be worth checking the predicted secondary structure of your target. Although modern modeling software is very good at recognizing folds and domains that occur at different positions in polypeptide chains, there is the possibility that multi-domain proteins containing long loops and areas of disorder will be poorly scored and ranked with the available software. It is therefore worth checking secondary structure agreement between target protein and the templates and/or the models generated, to inform your interpretation of the models you are presented with. Indeed, McGuffin writing in 2010 [12] asserted that simple scores based on secondary structure compatibility can be very effective model quality assessment and be used to filter out models with incorrectly or poorly formed secondary structures. See Table 7 for a list of disorder prediction tools.

4.7 Distant Homology Searches, Fold Recognition, and Threading Programs

In order to negate the limitations of sequence alignment, particularly where sequence identity is below that 35% threshold, the process of protein fold recognition was developed [1, 20]. This technique employs the rationale that evolutionary homologs often display less structural divergence than sequence divergence [35] and therefore less reliance on matching sequence and more on matching fold structure can result in less clutter of sequence-related but structurally distant template suggestions. Fold recognition commonly involves statistical methods (e.g., Hidden Markov

Table 7
Protein disorder and secondary structure prediction tools

JPred4 [11]	Secondary structure prediction online server	www.compbio.dundee.ac.uk/jpred/
PSIPred [10]	Hosted by UCL, London. Secondary structure prediction with links to associated applications	http://bioinf.cs.ucl.ac.uk/psipred/
Disopred [51]	Recognition of disordered regions	http://bioinf.cs.ucl.ac.uk/psipred/
IUPred [50]	Predictions of intrinsically unstructured proteins	https://iupred2a.elte.hu/
PrDOS [52]	Protein DisOrder prediction system	http://prdos.hgc.jp/cgi-bin/top.cgi

Table 8
Tools when no close matches are found

THREADER [53]	Fold recognition methods for predicting protein structure	http://bioinf.cs.ucl.ac.uk/software_downloads/threader/
GenTHREADER [54]	Rapid fold recognition, matching sequences against PDB chains assuming an evolutionary link	http://bioinf.cs.ucl.ac.uk/web_servers/
pGenTHREADER [54]	Highly sensitive fold recognition using profile–profile comparison	http://bioinf.cs.ucl.ac.uk/web_servers/
pDomTHREADER [54]	Highly sensitive homologous domain recognition using profile–profile comparison	http://bioinf.cs.ucl.ac.uk/web_servers/
HHPred [34]	Tertiary structure prediction and threading, part of the HH-suite of programs	https://toolkit.tuebingen.mpg.de/tools/hhpred
MUSTER [55]	MUlti-sources ThreadER, a threading algorithm combining sequence profile–profile alignment with structural information	https://zhanglab.cmb.med.umich.edu/MUSTER/
SPARKS-X [56]	Fold recognition software	http://sparks-lab.org/yueyang/server/SPARKS-X/

Models—HMM) [20] to compare sequence profiles of targets with potential templates and identify the most suitable ones from which to construct 3D models. Traditionally, threading methods were also developed which would fit or “thread” target sequences into the backbones of existing structures and then evaluate suitable templates using statistical energy potentials. Stand-alone individual fold recognition and threading techniques have enjoyed success in previous CASP experiments and include those listed below in Table 8. However, there is now a question as to whether their

predictive powers have reached a plateau [57], as most successful servers now deploy a combination, or consensus, of alternative techniques.

4.8 *Ab Initio or (Template) Free Modeling Methods*

Ab initio modeling which is essentially synonymous with template-free modeling (TFM) is a technique that applies physics-based rules in order to estimate the structure of a target sequence using the sequence as the only input [18]. These programs do not query the PDB or any other database, instead relying on the application of physical algorithms to build the model from scratch.

The algorithms used will be very similar to those discussed so far, focusing on torsion angles, hydrophobicity, secondary structure agreement as well as energy minimization and molecular dynamic technics. The computational power necessary to cope with the many degrees of freedom that present themselves in these cases is significant and many ab initio predictive servers run on either integrated CPU networks, powerful GPUs (graphical processing units), or neural networks and support vector machines (SVMs)—FALCON (a remote template alignment program employing a significant number of ab initio routines within its algorithms) harnesses the power of 20,000 volunteer CPUs for example [58]. QUARK represents a pure ab initio prediction methodology (there are others) whereas FALCON and Robetta (in the form of the upgraded ROSIE site—see Notes, Subheading 8) include a certain amount of ab initio routines behind the scenes while performing model building (see Table 9 for weblinks).

QUARK is typical of many of the modern ab initio prediction sites which now tend to use small fragments (1–20 residues long) and reference their own fragment database [59].

Here, it might be prudent to briefly mention the recent development of TFM programs specializing in amino acid contact prediction. The two leading proponents of this technology are Google DeepMind, using the Alphafold algorithm, and DMPfold. Alphafold uses a system of contact distance and angle predictions that are then solved by gradient descent mathematics [60]. DMPfold works

Table 9
A sample of available *Ab initio* or de novo modeling software

FALCON [58]	Software specializing in aligning query proteins with conserved regions	http://protein.ict.ac.cn/TreeThreader/
QUARK [59]	Structure prediction and protein folding to construct 3D models from amino acid sequence only	https://zhanglab.ccmbl.med.umich.edu/QUARK/
ROSETTA [19, 28, 29]	ROBETTA server (robot-Rosetta) provides ab initio folding and structure prediction, as well as fragment selection	http://rosetta.bakerlab.org/

slightly differently by predicting inter-atomic distances, torsion angles, and main chain hydrogen bonding to drive the folding prediction. Both use powerful neural networks and have reported success with CASP tertiary structure targets; DMPfold predicted 56% of folds correctly in CASP13 targets and AlphaFold led the field with 72% correct [61].

5 Comprehensive or Integrated Structure Prediction Webservers

The Swiss Institute of Bioinformatics (SIB) website (https://www.expasy.org/proteomics/protein_structure) has links to many publicly available programs designed to perform specific stages of the prediction process as well as those which perform the full service from start to finish. OmicX (<https://omictools.com>) is another useful website with an abundance of well-categorized resource links. It must also be mentioned that some of the above-mentioned server programs also offer complete sequence to 3-D model functionality or are part of a webserver suite or collection of programs designed to complement each other, for example, the UCL PSIPRED workbench (<http://bioinf.cs.ucl.ac.uk/psipred/>) allows one or many stages of the protein prediction pipeline to be undertaken at any one time with a simple tick-box system.

Below we will limit our focus to five leading one-stop webserver and describe briefly their mode of action and any advantages or special features they provide. They are listed in alphabetical order. IntFOLD is an integrated protein structure and function server consisting of a suite of interlinked programs developed by the McGuffin group and hosted by the University of Reading. As with many stand-alone servers, IntFOLD uses its own algorithms along with those from numerous other servers in order to multiply the power of template selection and accuracy of predicted models [21, 62].

5.1 IntFOLD

INPUT: IntFOLD simply accepts the sequence of the target protein of interest. There is the option to provide a name for the job and an email address to which the results page link can be sent. Click on the *IntFOLD submission* link to be taken to the latest version of the program. If an email address is not submitted, users should be sure to bookmark or save the link to the results page as it will be lost upon navigation away from the page.

MODE: IntFOLD works broadly on a two-step process; first, is a single template modeling step with Accuracy Self Estimate (ASE) scoring followed by a second multiple template modeling step, again with ASE scoring.

The first step of template identification harnesses the power of 14 separate algorithms, six stand-alone fold recognition programs—SP3, SPARKS-2, HHsearch, COMA, SPARKS-X, and CNFSearch, and the eight threading programs comprising the

LOMETS package. Each individual algorithm may submit up to 10 templates (140 in total), which are then run through the IntFOLD server's clustering and scoring algorithm ModFOLDClust2.

The second step involves an iterative multi-template modeling (MTM) regime using the cluster scores to rank the templates found in **step 1**. Firstly, the top two alignments are used to construct an initial model, this is then compared to models made using the top ranked plus any other template, the best model is selected based on amino acid coverage of the models. This is performed twice more for the evolving model before selected models are re-scored with ModFOLDclust. The 4-stage iterative model building and comparison process is then repeated. Additionally, I-TASSER and HHPred [34] are used to build three models each and these are added to the model group from the iterative process which are then fed into a ranking and refinement loop. Using ModFOLD6_Rank [21] and reFOLD algorithms, models are continuously ranked and refined via molecular dynamics procedures and the final top five-ranked models from this cyclic process constitute the IntFOLD output.

OUTPUT: The output file lists the top five models ranked by global model quality score and accompanied by a color-coded p-value. The following sections are also included; Disorder prediction, Domain Boundary prediction, Binding site prediction, and full quality assessment results. These are comprehensively described and explained on the IntFOLD Webserver help page (https://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD_help.html#examples) and so will not be repeated here. Users may download the data files for the predictions via the hyperlinks on the results page.

CREDENTIALS : In CAMEO server benchmarking IntFOLD4 was rated second on the common subset comparison (1-year performance 2016–17) and IntFOLD5 was rated first in 3-D data results for 3 months (Oct 2018–Jan 2019). The McGuffin group has also been competitively ranked in numerous recent CASP experiments [4].

5.2 I-TASSER

Developed and administered by Zhang Lab of the University of Michigan, the acronym stands for Iterative Threading ASSEMBly Refinement [22].

INPUT: In addition to the basic sequence in FASTA format, I-TASSER allows users to specify additional restraint data if known, for example, distance restraints in the form of atom contacts. If users would like to specify particular proteins to be used as homologs, their PDB codes can be entered and there is also the facility to upload a complete 3D homolog structure in PDB file format should that be required. Users can also take advantage of TASSER's threading credentials by excluding close sequence homologs and going below the usual cut-off of 25% sequence identity.

MODE: I-TASSER is a suite of programs. The initial fold recognition is carried out by the LOMETS meta-server with subsequent fragment threading by the MUSTER [55] algorithm. The fragments are then assembled into potential models with loop sections built by ab initio methods as necessary. SPICKER then selects the best models by clustering on a lowest energy basis and the process is verified by parallel model-build using TM-Align. The models are then re-clustered, and the final model is constructed using REMO software.

OUTPUT: Submissions can take 1–2 days to run by the end of which users will be emailed a results webpage link. The results are extensive and include a secondary structure visual display, solvent accessibility display, and a B-factor graph showing variation along the mode (*see Note 1*). Following this is an interactive list of the templates used as well as the top five models viewable in a Jmol-style graphical user interface. Each of the model files is downloadable and accompanied by a C-score, TM score, and RMSD. Included at the bottom of the results page are some potentially useful sections on predicted co-factors and binding sites, enzyme potential data, and gene ontology information.

CREDENTIALS: I-TASSER was ranked as the top server in CASP 7, 8, 9, and 10.

5.3 Phyre2

This is an updated version of the Phyre server that has been completely rewritten with the emphasis on both enhanced technical attributes and usability. The acronym stands for Protein Homology/analogY Recognition Engine V 2.0 and is run by the Structural Bioinformatics Group at Imperial College, London, making up part of the Genome3D collaboration between UCL, Imperial, Cambridge, and Bristol universities [20].

INPUT: Phyre2 can be accessed from the Phyre2 homepage, which will accept a sequence in FASTA format as well as an email address for results. It can also be accessed via the Genome3D page (<http://genome3d.eu>) where a FASTA, keyword or UniProt id submission returns a list of matches that, upon selection, lead to a predicted domains page. Here there are links to CATH and SCOP for protein classification information and Phyre2 for 3-D modeling (as well as links to some other Genome3D annotation software).

MODE: As with many servers, Phyre2 makes use of a number of other programs. Alignment and template detection is now upgraded from a PSI-BLAST search to a HMM-based fold library scan using HHsearch/HHpred software. Secondary structure is also predicted using PSIPRED. Phyre2 has a sophisticated mechanism for the management of insertions, deletions, and disordered or missing loop regions; employing a fragment-matching library and testing dihedral angle and energy scores to ensure the lowest possible perturbation in the structure as potential fragments are inserted. There is also an acknowledgment of the persistent

problem of few templates or templates that only match one domain for a multi-domain target. Here the *ab initio* modeling software Piong, which is designed to work as a virtual ribosome, is employed to build as much of the model as necessary. Lastly, DISOPred software predicts areas of disorder and the R3 protocol uses a rotamer library to orientate amino acid side chains.

OUTPUT: Results are emailed to users with a link to the results page. The page is split into four sections; firstly, a model based on the top-ranked template which can be viewed interactively in Jmol; secondly, a detailed graphic of predicted secondary structure and potential disorder scores; third is nice graphic of all templates and the percentage alignment for each, these are interactive and link to the fourth section below which lists all templates' structures and PDB information. These are downloadable individually, and there is a Download as zip option for the whole results page (*see Note 2*).

CREDENTIALS : Phyre2 is an older server that was been ranked sixth in CASP9 and tenth in CASP10. However, the authors are keen to point out that there is only 2–3% difference other servers' performance (measured by GDT_TS) [20], (with the exception of I-TASSER which scored slightly better in cases where only remote homologs exist).

5.4 Robetta

Robetta is the public-facing webpage of the Rosetta server prediction program developed by the Baker lab at the University of Washington, USA, and now administered by the Rosetta Commons group. Rosetta has a long history as a competitor in CASP and Robetta is a free-to-use front end-running the powerful Rosetta algorithms that have been so successful [19, 28, 59].

INPUT: Users must register in order to run jobs on Robetta. There are essentially three options upon registration; Rosetta comparative modeling (CM), Rosetta *ab initio* modeling (AB), or a fully automated pipeline. Users can paste (FASTA) or upload an amino acid sequence and also upload templates or alignments of their own if required. It is also possible to add custom distance constraints, if known. Users are only allowed one job at a time and jobs are run on a two-stage process; firstly, the identification of templates and secondly domain 3-D modeling. Users will be required to pick a domain to model after stage one and may submit only one domain at a time to conserve computing power (*see Note 3*).

MODE: Robetta essentially runs four separate algorithms for template selection and alignment; these are RaptorX, HHPred, SPARKS-X, and Map align. As above, users are able to upload their own templates and alignment data if they wish to bypass this stage. Rosetta algorithms then perform 3-D modeling on a domain by domain basis and also check potential interface areas by Alanine scanning (each amino acid is in-turn replaced by Alanine and the effect on the calculated binding energy computed) for binding and interaction prediction.

OUTPUT: Jobs typically take 1–2 days to run and users receive access to the results page via email. The results are comprehensive and include a multi-server secondary structure annotation with disorder predictions plus interactive RasMol annotations of the top five models, which can be colored by error estimation. Graphical error plots of distances (in Å) between C α atoms of the model compared to the native structure also accompany each model. The results page is interactive and a click on each domain will reveal the templates and alignments used to build it as well as a cluster graph showing its position relative to the average. For comparative modeling, a predicted confidence value equivalent to GDT_TS is provided. For ab initio modeling, a predicted confidence value equivalent to TM-score of the top 10 Rosetta scoring models is provided instead.

CREDENTIALS : Robetta has competed in CAMEO since 2014 and cites its success in terms of LDDT score (Local Distance Difference Test—which evaluates inter-atomic distances). Robetta averages around 69 (0–100 where higher scores are better). The error estimates included in results are also evaluated through CAMEO and Robetta achieves an average model confidence score of 0.85.

5.5 *Swiss-Model*

This was the first fully automated server developed over 20 years ago and is now a comprehensive website with enhanced functionality administered by the Swiss Institute of Bioinformatics (SIB) [16].

INPUT: As well as a FASTA sequence users can input the UniProt accession code for the target. There also exists the facility to upload potential template files, but familiarity with the SIB Swiss-PDBViewer, also known as DeepView, will likely be necessary for this.

MODE : There are a number of key features to SWISS-MODEL. It is designed specifically to run HMM modeling, via HHblis [34] software, on the SWISS-MODEL Template Library (STML); an amalgamated version of the SWISS PROT and PDB databases augmented with derived data allowing the differentiation between bound ligands and solvent molecules. SWISS-MODEL will also run a BLAST search and check secondary structure via PSIPRED before allowing the user a choice between automated or manual selection of the templates found. If manual mode is selected, the templates are listed along with their Global Mean Quality Estimation score (GMQE—essentially an average of QMEAN [39] scores applied to each individual amino acid) and information on predicted ligands, oligomeric state, and sequence alignment. Users are able to select any number of templates and these are then displayed in a 3-D structural super-position as well as a 2-D cluster graph of evolutionary distance. Users can then choose their potential templates based on clustering, domain matches, and sequence identity scores.

SWISS-MODEL will then build an all-atom model using ProMod II software with a back-up comparison built using MODELLER [23].

OUTPUT: Users get a comprehensive listing of model coordinates, target-template alignment, step-by-step modeling log, information on potential oligomeric state, potential ligands, and co-factors as well as a QMEAN score, all of which can be downloaded. The models within the graphical interface are also colored by QMEAN to show areas of higher and lower confidence.

6 CASP and CAMEO

To give some context to the programs and rating credentials presented in Subheading 5, it is worth expanding here on the CASP and CAMEO community-wide experiments (first referenced in the introduction) which form the arena in which modeling expertise is tested and advanced.

The CASP experiment has been running as a biannual blind tertiary structure prediction competition since its inception by John Moult and associates in 1994 [63]. The purpose has been to provide a vehicle for the objective assessment of the prediction capability of *in silico* groups globally with the added benefit of shared practice and identification of technical advancement. Organizers source soon-to-be-solved crystal or NMR 3-D structures from researchers and invite *in silico* prediction groups to solve the structure before revealing the answers and scoring groups' efforts around 9 months later [35]. These experiments have seen the discipline of *in silico* protein structure prediction rise in integrity over the past 25 years with CASP1 attracting 35 invited predictor groups [63] compared to CASP6, (run 10 years later in 2004) which received over 30,000 predictions from 200 predictor teams [35] and CASP8 (2008) representing peak predictor participation with 253 groups across 24 countries worldwide [64].

Since the time of its inception to the latest version the focus of the CASP experiment has changed and expanded from mostly *ab initio* modeling to comparative methods (TBM) which are able to exploit the wealth of structural information now available (by CASP10 (2012) there were 1393 distinct folds available in the PDB and a total of 87,000 solved protein structures [65].

CAMEO (Continuous Automated Model EvaluatiOn—see Fig. 3) is a server-based experiment run along similar lines to CASP but differing in that participating servers must be fully automated with no human intervention in the prediction process. Servers receive their targets on a weekly basis and have 3 days in which to complete the prediction and return results to CAMEO. The ratings and metrics on the relative successes of the servers is a good indication of their competitiveness and likelihood of providing a good quality model.

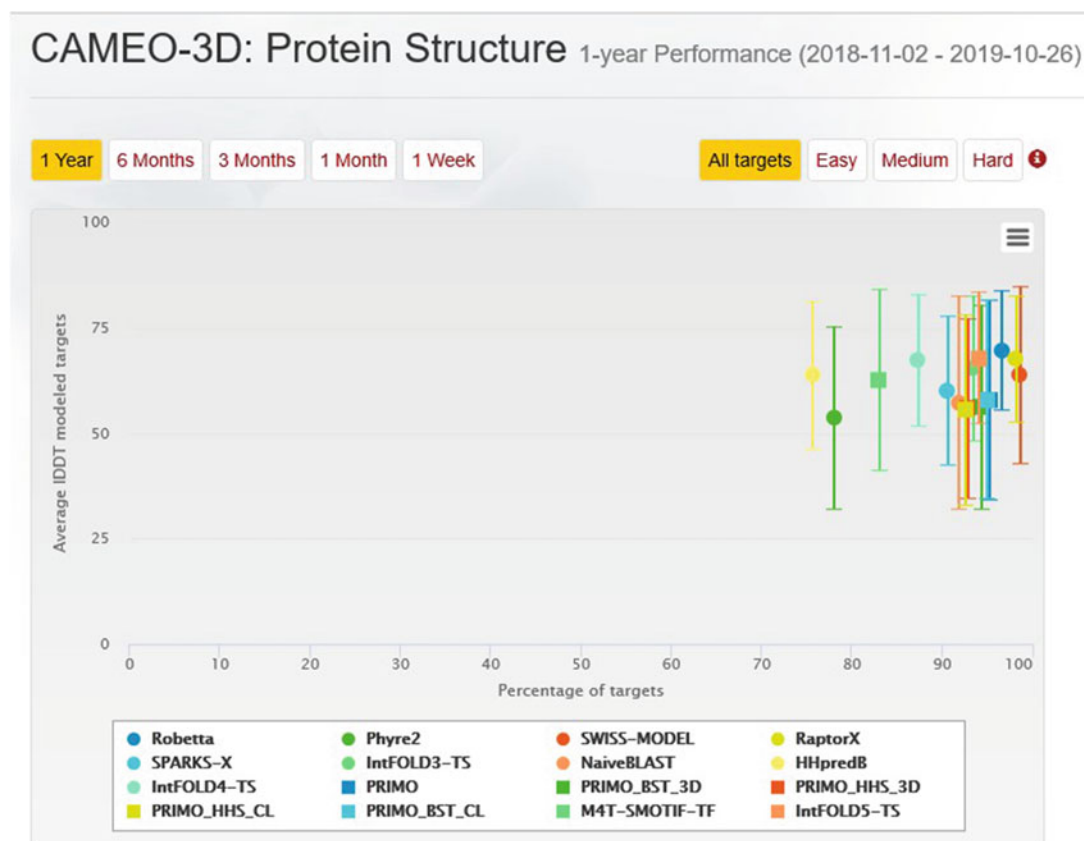


Fig. 3 A screenshot from the CAMEO website showing participating servers. (Taken from https://www.cameo3d.org/sp/1-year/difficulty/all/?to_date=2019-10-26)

7 Protein-Protein Interactions (PPI) and Quaternary Structure Prediction

While both CASP and CAMEO experiments include predictions of the interaction of proteins to form dimers and some higher level oligomers, the third community-wide prediction competition CAPRI (Critical Assessment of Prediction of Interactions) forms the area of expertise in PPI and quaternary structure prediction. However, communities are now merging somewhat with CASP 11 (2014) and CASP 12 (2016) seeing joint CASP-CAPRI collaborations on many prediction targets, representing a crossover of docking and homology modeling expertise.

7.1 Docking Programs

Many program routines currently used in the CAPRI experiment were originally developed to predict protein docking interfaces with either ligands or with themselves to form homodimers which explains the inclusion of the word “dock” in many program names. Although these programs can often perform a protein-ligand docking function, the ones listed here have been developed

Table 10
Docking-based PPI modeling software

GRAMM-X [66], ZDOCK [67], and MEGADOCK [68]	Fast Fourier transform (FFT)-based programs
FRDOCK [69]	Modified FFT technique (Chacon et al., 2009), using a reduced 3-D search space to save time and computer power yet reportedly achieving a comparable level of accuracy
PatchDock [70]	Uses image segmentation techniques to map the contours of the surface of a protein followed by shape complementarity and symmetry to fit the protein surfaces together
Hex [71]	Uses spherical harmonics (D. Ritchie)
RosettaDock [28]	Uses a combination of side-chain orientations and free-energy calculations linked to its probability-based Monte Carlo algorithm
LZerD [72]	A unique approach identifying Zernike 3-D shape descriptors followed by complementarity calculations
ClusPro [73]	Models are clustered together depending on the location of the interface residues, the logic being that the size of clusters is proportional their probability of representing the native model
HADDOCK [74]	A physics-based scoring function based on a combination of van der Waal's interactions, electrostatics, and desolvation measures

to focus primarily on protein–protein interactions. If a program specifically for docking is required, a popular choice is Autodock Vina.

A number of different docking approaches have been developed to predict protein–protein interactions. A favorite technique is the use of a Fast Fourier Transform (FFT) to search all possible binding modes in a 6-D search space (3 rotational and 3 translational) [66] but there are others based on shape complementarity, spherical harmonics, and identification of Zernike shape descriptors as well as those employing more traditional physics-based measurements such as energy minimization, side-chain orientation, and solvent accessibility.

See Table 10 for a list and brief description of some of the main players in the prediction of interactions and quaternary structure via docking algorithms.

All approaches have had success over the rounds of CAPRI experiments with ClusPro scoring a success rate of 5 high and 3 medium quality models, followed by HADDOCK with 4 high and 1 medium (from 12 targets) in 2009 and LZerD scoring 4 high and 3 medium models from 20 targets in 2016 (data from the server modeling section of CAPRI [73]). RosettaDock has also enjoyed success, predicting all 5 small targets with medium to high accuracy in rounds 3–5 [28] as well as being ranked second

in the 2014 predictor server rankings [75]. All servers are listed with their varying levels of success in the 2014 CAPRI round 30 [75] at <http://www.capri-docking.org/resources/#performance-of-docking-servers-in-capri>. It must be added that most success in protein interaction prediction has come in the form of predicting dimers and certain higher order oligomers exhibiting spherical symmetry with hetero complexes continuing to present problems [76]. Analysis of the joint CASP/CAPRI experiments by Lensink et al. (2016) [75] suggests that, in general, docking approaches to predicting quaternary structures performed better than template-based modeling due, in part, to the increased difficulty of finding reliable oligomeric crystal templates in the PIR database. Therefore, although an increasing number of 3-D modeling programs will offer a likely quaternary structure for a target sequence it may be worth bearing in mind the additional difficulties that this process involves when considering the accuracy of the final model.

7.2 The Evolution of Docking Methods

Although docking programs can produce very good models of homodimers, they are less well adapted to identifying quaternary structure straight from sequence especially for hetero or larger complexes. While some of the programs listed above have been adapted to predict higher level homomers, e.g., MZDock and MultiLZerD (as demonstrated by Nakamura et al. (2017)) [77], their use often still requires a catalog of specialist software and results can be variable. One server to both beef-up its computing power and allow easy user input directly from a webpage interface is MEGADOCK 4.0 (accessible as MEGADOCK-Web <http://www.bi.cs.titech.ac.jp/megadock-web/>).

Other specialist quaternary prediction sites that are publicly available via a webpage and require only sequence data in FASTA format as input include SWISS-MODEL, QuaBingo, and Galaxy.

Bertoni et al. (2017) [78] reported their attempt to go from sequence straight to quaternary structure using SWISS-MODEL that samples multiple template databases as well as adding a co-evolution distance measure score—termed PPI fingerprint. If it is considered possible to build a quaternary model using SWISS-MODEL, the quaternary structure quality estimate (QSQE) score will be included in the output.

Another study, Tung et al. (2016) [79] reported their description of the program QuaBingo that identifies conserved domains using the BLOCKS database of motifs based on SWISSPROT. QuaBingo also adds a pseudo amino acid descriptor (PseACC) that takes into account the hydrophobic-hydrophilic character of individual residues. QuaBingo can be accessed from <http://predictor.nchu.edu.tw/QuaBingo>.

Galaxy also has a homomer prediction facility based on a simple FASTA sequence submission (<http://galaxy.seoklab.org/>) as Galaxy-Homomer.

8 Notes

1. When using I-TASSER:

Models are selected by clustering and although there is good evidence that clustering improves model identification [37], care should be taken when a target sequence has few homologs as clustering may be less powerful. Also, the ranking of the models by cluster size presents the potential for a good model (higher C-score) being omitted from the top of the models list as it appears in a smaller cluster. Results should be checked for these issues.

2. When using Phyre 2:

Phyre2 has a number of ad-on functions that may be useful.

BackPhyre is a genome search tool allowing users to search for homologs to their solved structure in specific genomes.

One to one threading can be used if users have biological information indicating that a specific protein should be used as the template. A file can be uploaded.

Phyre Alarm is a scanning service which checks fold libraries on a weekly basis and updates users who have not found a good template match in their initial modeling attempt.

Phyre Investigator give access to extra information on model quality analysis, alignment confidence, and Ramachandran analysis as well as catalytic site, mutation analysis, and potential interface detection.

Lastly, users can opt for Batch Analysis, where up to 100 jobs can be scheduled to run automatically and Job Manager that gives access to a page with all previously run jobs.

3. When using Robetta:

Rosetta software is available to download if users would prefer to run the algorithm locally from the command line. There is also an option to download pyRosetta for those interested in running the software via Python. From the Robetta homepage are links to the latest Rosetta incarnation called ROSIE. This has links to a whole host of functional characterization programs (one could say a whole lotta Rosie!) and would be worth visiting.

A list of scoring functions often encountered in protein structure prediction is given in Table 11.

Table 11
A list of scoring functions often encountered in protein structure prediction

Predictive scores (for model quality assessment)	
C-score	(I-TASSER). This is a confidence score calculated for threading template alignments. Scores range from -5 to 2 with higher scores indicating a better alignment
E-value	(BLAST and RAPTOR). Related to p-value, for two sequences with n alignments, E-value represents the expected number of false alignments having greater than n correctly aligned positions. The closer to 0 the better
LG score	(PCons). Essentially a p -value for the significance of a structural similarity match. A significant threshold would be $1 \times 10^{-1.5}$ (0.031), so anything below this figure would represent a potentially good match between a model and the target
MaxSub score	Identifies the largest set of $C\alpha$ atoms that superimpose well over two structures so focusing on well-predicted regions. Produces a score between 0 and 1 with 1 being the best, normalised for the size of the overlap so that larger sequences do not automatically score better than shorter ones
ProQ score	(PCons). This is the $-\log$ of LG score, e.g., for a significant LG score of $1 \times 10^{-1.5}$ The ProQ score would be 1.5 . Therefore, 1.5 and upwards are good scores
p -Value	The proportion of models with a particular score that do not share any similarity with the native structure, i.e., will have the same alignment purely by chance. $<0.001 = 1/1000$ chance (or less) that the model is incorrect; <0.01 less than a $1/100$ chance; <0.05 , less than a $1/20$; <0.1 less than a $1/10$; >0.1 likely to be a poor model with little or no similarity to the native structure
Qmean score (qualitative model energy analysis)	The simplest form of this, Qmean4, is the sum of four measures; geometric analysis of the torsion angles of the carbon backbone, CB interactions, all atom interactions, and a solvation score (QMean6 additionally includes a secondary structure agreement score and a solvent accessibility agreement as percentages. A Qmean4 of 1 is good with 0 considered acceptable but, as with Z-score, a negative figure indicates a poorer fit. Qmean scores are often transformed into Z-scores for ease of comparison with experimentally determined structures
S-score	(PCons). A global super-position score calculated as a transformation of RMSD on a per amino acid residue basis. 1 would represent a perfect score and 0 a useless model
TM-score	This is a measure of the similarity of two protein structures based on a weighted RMSD score, i.e., small RMSD values are weighted more strongly than large scores in an attempt to overcome the distortion of RMSD for good models with local errors. Scores can range from 0 to 1 with >0.5 representing a strong match and < 0.17 a match no better than random

(continued)

Table 11
(continued)

Z-score	A Z-score is an expression of the number of standard deviations from the mean structure of the templates. A Z-score of zero would indicate that a template represents the mean structure, a negative score would indicate a worse fit than the mean whereas a positive score would indicate a better fit. However, it must be remembered when dealing with normal distributions and standard deviations, the further one travels from the mean, in any direction, the more likely one is to be looking at an outlier and the true value is likely to be close to the mean
Observed scores (obtained when a model is compared to the true structure)	
Global model quality score	The global model quality scores range between 0 and 1. In general, scores less than 0.2 indicate there may be incorrectly modeled domains and scores greater than 0.4 generally indicate more complete and confident models, which are highly similar to the native structure
GDT_TS (Global distance test total score)	A CASP observed score. Explanations may be found at http://predictioncenter.org/casp13/doc/help.html#GDT_TS
B-factor	Often known as a temperature factor, this measurement is traditionally supplied with crystallographic structures as a measure of the displacement of individual atoms from their true position. Measured in angstroms squared, 0 would be a perfect score with anything below 30 Å ² considered as acceptable and anything greater than 60 Å ² , questionable (for reference a 15 Å ² score would equate to a mean displacement of an atom by 0.44 Å and 60 Å ² , a mean displacement of 0.87 Å)
RMSD (root mean square deviation)	This usually refers to the average distance of all amino acid pairs in two compared structures. Some programs will give a global score for the whole structure whereas others may give local scores per amino acid residue. Measured in Å, a good score would be <2.0 although this will depend on the resolution of the templates used to calculate the model. This measure, although widely quoted, is particularly sensitive to the problem of local alignment error discussed below

References

1. Rangwala H, Karypis G (2010) Introduction to protein structure prediction. In: Rangwala, Karypis (eds) Introduction to protein structure prediction: methods and algorithms. John Wiley & Sons
2. Cao R, Bhattacharya D, Adhikari B, Li J, Cheng J (2015) Large-scale model quality assessment for improving protein tertiary structure prediction. *Bioinformatics* 31(12): i116–i123
3. Berman HM, Westbrook J, Feng Z et al (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242
4. McGuffin LJ, Adiyaman R, Maghrabi A et al (2019) IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Res* 47: W408–W413
5. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47:D506–D515
6. El-Gebali S et al (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* 47:D427–D432
7. Andreeva A, Howorth D, Chothia C et al (2014) SCOP2 prototype: a new approach to

- protein structure mining. *Nucleic Acids Res* 42:D310–D314
8. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res* 45(D1):D289–D295
 9. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637
 10. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292:195–202
 11. Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* 43:W389–W394
 12. McGuffin LJ (2010) Model quality prediction. In: Rangwala, Karypis (eds) *Introduction to protein structure prediction: methods and algorithms*. John Wiley & Sons
 13. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372:774–797
 14. Zahn-Zabal M, Michel PA, Gateau A, Nikitin F et al (2020) The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Res* 48(D1):D328–D334
 15. Szklarczyk D, Gable A, Lyon D et al (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47:D607–D613
 16. Biasini M, Bienert S, Waterhouse A et al (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* 42:W252–W258
 17. Buenavista M, Roche D, McGuffin LJ (2012) Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics* 28:1851–1857
 18. Guo J, Ellrott K, Xu Y (2008) A historical perspective of template-based protein structure prediction. In: Zaki, Bystroff (eds) *Protein structure prediction*, 2nd edition, methods in molecular biology, vol 413. Springer
 19. de Oliveira HP, Shi J, Deane C et al (2015) Building a better fragment library for de novo protein structure prediction. *PLoS One* 10:e0123998
 20. Kelley LA, Mezulis S, Yates CM et al (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 10:845–858
 21. McGuffin LJ, Shuid AN, Kempster R et al (2018) Accurate template-based modeling in CASP12 using the IntFOLD4-TS, ModFOLD6, and ReFOLD methods. *Proteins* 86:335–344
 22. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER suite: protein structure and function prediction. *Nat Methods* 12:7–8
 23. Webb B, Sali A (2016) Comparative protein structure modeling using modeller. *Curr Protoc Bioinformatics* 54, John Wiley & Sons, Inc.:5.6.1–5.6.37
 24. Wang Z, Eickholt J, Cheng J (2010) MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics* 26:882–888
 25. Wallner B, Elofsson A (2005) Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics* 21:4248–4254
 26. Yachdav G, Kloppmann E, Kajan L et al (2014) PredictProtein—an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res* 42:W337–W343
 27. Ma J, Wang S, Zhao F, Xu J (2013) Protein threading using context-specific alignment potential. *Bioinformatics (Proceedings of ISMB 2013)* 29(13):i257–i265
 28. Park H, Kim D, Ovchinnikov S, Baker D (2018) Automatic structure prediction of oligomeric assemblies using Robetta in CASP 12. *Proteins* 86:283–291
 29. Simons K, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268:209–225
 30. Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
 31. Larkin MA, Blackshields G, Brown NP et al (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948
 32. Sievers F, Wilm A, Dineen DG et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. *Mol Syst Biol* 7:539
 33. Madeira F, Park YM, Lee J et al (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 47:W636–W641

34. Zimmermann L, Stephens A, Nam SZ et al (2018) A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J Mol Biol* 430(15):2237–2243
35. Moulton J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 15:285–289
36. Adiyaman R, McGuffin LJ (2019) Methods for the refinement of protein structure 3D models. *Int J Mol Sci* 20:2301
37. McGuffin LJ, Roche DB (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics* 26:182–188
38. Uziela K, Shu N, Wallner B, Elofsson A (2016) ProQ3: improved model quality assessments using Rosetta energy terms. *Sci Rep* 6:33509
39. Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27:343–350
40. Feig M (2017) Computational protein structure refinement: almost there, yet still so far to go. *Wiley Interdiscip Rev Comput Mol Sci* 7: e1307
41. Heo L, Feig M (2018) PREFMD: a web server for protein structure refinement via molecular dynamics simulations. *Bioinformatics* 34:1063–1065
42. Feig M (2016) Local protein structure refinement via molecular dynamics simulations with locPREFMD. *J Chem Inf Model* 56:1304–1312
43. Heo L, Park H, Seok C (2013) GalaxyRefine: protein structure refinement driven by side-chain repacking. *Nucleic Acids Res* 41:384–388
44. Rodrigues JPGLM, Levitt M, Chopra G (2012) KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Res* 40:323–328
45. Khoury GA, Smadbeck J, Kieslich CA et al (2017) Princeton_TIGRESS 2.0: high refinement consistency and net gains through support vector machines and molecular dynamics in double-blind predictions during the CASP11 experiment. *Proteins Struct Funct Bioinform* 85:1078–1098
46. Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys J* 101:2525–2534
47. Bhattacharya D, Cheng J (2013) 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. *Proteins* 81:119–131
48. Shuid AN, Kempster R, McGuffin LJ (2017) ReFOLD: a server for the refinement of 3D protein models guided by accurate quality estimates. *Nucleic Acids Res* 45:W422–W428
49. Zhang J, Liang Y, Zhang Y (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* 19:1784–1795
50. Dosztányi Z (2018) Prediction of protein disorder based on IUPred. *Protein Sci* 27:331–340
51. Jones DT, Cozzetto D (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 31:857–863
52. Ishida T, Kinoshita K (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res* 35: W460–W464
53. Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. *Nature* 358:86–89
54. Lobley A, Sadowski MI, Jones DT (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* 25:1761–1767
55. Wu S, Zhang Y (2008) MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* 72:547–556
56. Yang Y, Faraggi E, Zhao H, Zhou Y (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. *Bioinformatics* 27:2076–2082
57. Skolnick J, Zhou H (2017) Why is there a glass ceiling for threading based protein structure prediction methods? *J Phys Chem B* 121:3546–3554
58. Wang C, Zhang H, Zheng W-M et al (2015) FALCON@home: a high-throughput protein structure prediction server based on remote homologue recognition. *Bioinformatics* 32:462–464
59. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* 80:1715–1735
60. Hutson M (2019). AI protein-folding algorithms solve structures faster than ever. *Deep*

- learning makes its mark on protein-structure prediction. In: *Nature NEWS*, ISSN 1476-4687. <https://www.nature.com/articles/d41586-019-01357-6>. Accessed 31 Oct 2019
61. Greener J, Kandathil S, Jones D (2019) Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat Commun* 10:3977
 62. Nealon J, Philomina L, McGuffin L (2017) Predictive and experimental approaches for elucidating protein-protein interactions and quaternary structures. *Int J Mol Sci* 18:2623
 63. Moult J et al (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins* 23:ii-iv
 64. Moult J, Fidelis K, Kryshchuk A, Tramontano A (2011) Critical assessment of methods of protein structure prediction (CASP)—round IX. *Proteins* 79(Suppl 10):1-5
 65. Moult J, Fidelis K, Kryshchuk A et al (2014) Critical assessment of methods of protein structure prediction (CASP) — round x. *Proteins* 82:1-6
 66. Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res* 34:W310-W314
 67. Pierce BG, Wiehe K, Hwang H et al (2014) ZDOCK server: interactive docking prediction of protein-ProteinComplexes and symmetric multimers. *Bioinformatics* 30:1771-1773
 68. Hayashi T, Matsuzaki Y, Yanagisawa K et al (2018) MEGADOCK-Web: an integrated database of high-throughput structure-based protein-protein interaction predictions. *BMC Bioinformatics* 19:62
 69. Garzon JJ, López-Blanco JR, Pons C et al (2009) FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics* 25:2544-2551
 70. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 33:W363-W367
 71. Macindoe G, Mavridis L, Venkatraman V et al (2010) HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Res* 38:W445-W449
 72. Peterson LX, Kim H, Esquivel-Rodriguez J et al (2017) Human and server docking prediction for CAPRI round 30-35 using LZerD with combined scoring functions. *Proteins* 85:513-527
 73. Vajda S, Yueh C, Beglov D et al (2017) New additions to the ClusPro server motivated by CAPRI. *Proteins* 85:435-444
 74. Vangone A, Rodrigues JP, Xue LC et al (2017) Sense and simplicity in HADDOCK scoring: lessons from CASP-CAPRI round 1. *Proteins* 85:417-423
 75. Lensink M, Velankar S, Kryshchuk A et al (2016) Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. *Proteins* 84(Suppl 1):323-348
 76. Lafita A, Bliven S, Kryshchuk A et al (2018) Assessment of protein assembly prediction in CASP12. *Proteins* 86:1-399
 77. Nakamura T, Oda T, Fukasawa Y, Tomii K (2018) Template-based quaternary structure prediction of proteins using enhanced profile-profile alignments. *Proteins* 86(Suppl 1):274-282
 78. Bertoni M, Kiefer F, Biasini M et al (2017) Modelling protein quaternary structure of homo- and heterooligomers beyond binary interactions by homology. *Sci Rep* 7:10480
 79. Tung C-H, Chen C-W, Guo R-C et al (2016) QuaBingo: a prediction system for protein quaternary structure attributes using block composition. *Biomed Res Int* 2016:9480276



Chapter 3

Methods for Molecular Modelling of Protein Complexes

Tejashree Rajaram Kanitkar, Neeladri Sen, Sanjana Nair, Neelesh Soni, Kaustubh Amritkar, Yogendra Ramtirtha, and M. S. Madhusudhan

Abstract

Biological processes are often mediated by complexes formed between proteins and various biomolecules. The 3D structures of such protein–biomolecule complexes provide insights into the molecular mechanism of their action. The structure of these complexes can be predicted by various computational methods. Choosing an appropriate method for modelling depends on the category of biomolecule that a protein interacts with and the availability of structural information about the protein and its interacting partner. We intend for the contents of this chapter to serve as a guide as to what software would be the most appropriate for the type of data at hand and the kind of 3D complex structure required. Particularly, we have dealt with protein–small molecule ligand, protein–peptide, protein–protein, and protein–nucleic acid interactions.


Most, if not all, model building protocols perform some sampling and scoring. Typically, several alternate conformations and configurations of the interactors are sampled. Each such sample is then scored for optimization. To boost the confidence in these predicted models, their assessment using other independent scoring schemes besides the inbuilt/default ones would prove to be helpful. This chapter also lists such software and serves as a guide to gauge the fidelity of modelled structures of biomolecular complexes.

Key words Molecular docking, Protein-biomolecular complexes, 3D structure modelling, Scoring and sampling

1 Introduction

All biological processes are mediated by various molecular interactions. These include interactions between protein and protein, protein and small molecule ligands, protein and DNA, etc. Characterizing these interactions is essential for gaining biological insights. Experimental characterization is often cumbersome, expensive, and/or difficult to perform. Computational methods [1–4] are hence routinely used to model the 3D structures of the complexes resulting from such interactions.

The computational methods fall into two broad categories—(a) Those that exploit information from a related or homologous template structure (henceforth referred to as template-based methods) or (b) methods that attempt to model the 3D structures of



Modelling Approach	Protein - small molecule	Protein - peptide	Protein - protein	Protein - nucleic acid	Large macromolecular assembly
Homology based		MODELLER			
		SPOT-Peptide	Template based methods HOMCOS InterPreTS Threading based methods IwRAP SPRING Structu2Net Coev2Net COTH	Template based methods PRIME 2.0 TFModeller Threading based methods ModelX	
Template based docking		GalaxyPepDock	PRISM	MPRDOCK	
Binding site prediction/ Restraint prediction	DEPTH Prankweb COACH 3DLigandSite ProBis-CHARMMing PockDrug CavityPlus FPocket LigSite Castp PASS Concavity SURFNET	ACCLUSTER PepSite SPRINT-Str PeptiMap	CPDRT BIPSPi EVCoupling Complex	DBSI DISPLAR DR_Bind DNABindProt PRNA aaRNA	IMP Rosetta
Docking	Autodock Vina GOLD AutoDock DOCK FlexX Surflex GLIDE LigandFit SwissDock SLIDE	Rosetta FlexPepDock DynaDock PepCrawler PEP-Fold DINC 2.0 Surflex-Dock PepATTRACT MdockPep CABS-Dock AnchorDock ClusPro PeptiDock PIPER-FlexPepDock	HADDOCK ClusPro ZDOCK Hex	HDock NPDOCK 3dRPC PatchDock FTDock	
Model assessment methods	Scoring functions in DOCK DOCK and Autodock intersection of best scoring poses	FOLDX	PIZSA		

Fig. 1 Spreadsheet of select protein complex structure modelling methods that can be used depending on the information available. The boxed methods that span various sections indicate applicability of the method in multiple categories

complexes without any predetermined structural bias. Such methods are often referred to as *ab initio* or as template-free algorithms, include the various docking programs. Many contemporary algorithms make use of a hybrid of methods (a) and (b) to predict the structure of the interacting 3D complex. Figure 1 shows a spreadsheet of many such methods.

Most computational methods employ similar protocols for predicting the structures of the complexes—viz, sampling different conformations and then evaluating/scoring them to find the most optimal mode of association. Each of the algorithms differs in strategies they use for these sampling and scoring steps [5].

This chapter is written to serve as a practical guide to model complexes of (1) protein–small molecule ligands, (2) protein–peptide, (3) protein–protein, (4) protein–nucleic acid (DNA/RNA), and (5) macromolecular assemblies. In each subsection, one or a few representative methods are highlighted while some information is provided about alternate techniques. The choice of representative method has been based on our familiarity, the ease of access (with a preference for freeware) and overall popularity. We believe that

once predictions have been made their assessment is crucial in deciding their benefit or applicability, and we list a few such software that can be used for assessment.

We believe that the relevance of this chapter is enhanced given the current circumstances, when there is an all-out effort to discover or design therapeutic agents and vaccines against SARS CoV2.

2 Modelling Protein–Small Molecule Complexes

Modelling protein–small molecule complexes is important for a wide range of applications from gaining insights into processes such as metabolism to designing therapeutics. While naturally occurring small molecules (~50–1500 daltons [6]) are integral components of metabolic and sensory pathways [7], synthetic small molecules (>500 daltons) find applications in designing therapeutic agents.

We envisage two different situations that would warrant the need for modelling protein–small molecule ligand complexes—(a) to find a suitable small molecule ligand for a given target protein, and (b) to find protein targets of a given small molecule. In both cases, we would also want to find the exact binding pose of a small molecule onto a particular target protein. The sections below cover the situation (a) in some length along with an illustrative example of finding suitable small molecule inhibitors to the Nipah virus glycoprotein [8]. The issues discussed in Subheadings 2.2–2.4 below are also applicable to the situation (b).

2.1 *Selecting the Small Molecule Library*

When searching for putative binding small molecule ligands of given target proteins, it is essential to utilize a screening library. Two such popular libraries are PubChem [9] and ZINC [10]. PubChem hosts ~103 million chemical compounds annotated by physical and chemical properties, biological activities, toxicity, etc. One can create appropriate subsets based on the desired properties of the small molecules. The ZINC database hosts ~230 million commercially available compounds categorized into pre-created subsets such as FDA approved drugs, derivatives of natural products, and so on. The compounds in the ZINC database are also available in docking friendly file formats. User defined subsets based on physical and/or biological properties can also be easily created.

For our example of finding an appropriate inhibitor to the Nipah glycoprotein, we selected the ZINC12 clean drug-like subset. More on this in Subheadings 2.2–2.4.

Small molecules can also be selected from various other online libraries such as DrugBank, ChEMBL, ChemSpider, KEGG, ChEBI, and Ligand Depot [11–16].

2.2 Predicting Small Molecule Binding Pockets on the Target Protein

Many docking software that attempt to predict/build the complexes of proteins with their small molecule ligands often scan the entire protein surface for suitable binding pockets for the ligands. This exercise makes screening a large number of compounds computationally expensive and time consuming. This problem can be circumvented by localizing potential small molecule-binding sites and then having the software scan these sites to conformationally optimize the protein–ligand complex.

A small molecule-binding pocket is a cavity on or inside the protein that can potentially harbor a ligand [17]. Several methods such as ProBiS-CHARMMing, 3DLigandSite, PrankWeb, and PockDrug-Server are among others that predict the binding pockets given a 3D structure of a protein predict the binding pockets given a 3D structure of a protein [17–28]. DEPTH (<http://cospi.iiserpune.ac.in/depth/htdocs/index.html>) is one such method that uses the depth of amino acid residues along with the evolutionary information to predict putative binding pockets. The DEPTH server takes 3D structure of a protein as input (*see* **Notes 1 and 2**) and assigns probability scores to each of the amino acids to be a part of a binding pocket. A user tuneable cut-off score can be used to select binding pockets. These predicted binding pockets can then be used as an input to docking programs.

For instance, DEPTH predicts two binding pockets on the surface of Nipah glycoprotein (PDB ID: 3D11). Interestingly, one of the predicted pockets overlaps with the region where the glycoprotein interacts with host cells proteins. Each of these pockets can be used for docking.

2.3 Docking Small Molecules on a Target Protein: Sampling the Ligand Conformation and Scoring

Molecular docking, similar to other computational procedures, involves a sampling and scoring protocol. There are various sampling schemes such as the systematic incremental approach [29], shape-based sampling [30], genetic algorithms [31], fragment-based approaches [32], and Monte Carlo simulations [33]. The sampling generates various conformations of the small molecules called poses that are evaluated by a scoring scheme. The scoring includes physics-based scoring schemes, empirical scoring functions or knowledge-based potentials [34].

Autodock [35] is one of several popular docking programs (Refer Table 1 for other docking methods) that uses a Lamarckian genetic algorithm for sampling conformations. A semi-empirical free energy force field is used to predict the binding free energy. Binding poses of a small molecule can be sampled on the entire protein surface, or it can be restricted to binding pockets (such as the two pockets predicted by DEPTH for Nipah glycoprotein). Along with exploring the poses of the small molecule, protein side chain conformations can also be sampled to account for their flexibility (flexible docking). The tutorial http://autodock.scripps.edu/faqs-help/tutorial/using-autodock-4-with-autodocktools/2012_ADTtut.pdf describes the docking procedure in detail.

Table 1
Non-exhaustive list of protein–small molecule docking methods

Tool	URL
AutoDock [35]	http://autodock.scripps.edu/downloads/autodock-registration/autodock-4-2-download-page/ (standalone)
AutoDock Vina [36]	http://vina.scripps.edu/download.html (standalone)
DOCK [37]	http://dock.compbio.ucsf.edu/Online_Licensing/index.htm (standalone)
FlexX [32]	https://www.biosolveit.de/FlexX/ (standalone)
GOLD [31]	https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/ (standalone)
GLIDE [29]	https://www.schrodinger.com/glide (standalone)
LigandFit [33]	Not available
SwissDock [38]	http://www.swissdock.ch/ (standalone)

2.4 Shortlisting the Compounds

One method of selecting potential small molecule ligands is based on the energy values of their docked poses. If a small molecule is already known to bind a given pocket (control molecule), the score of its complex with the protein can be used as a cut-off or guide to shortlist other ligands. All the complexes where small molecules are docked at this pocket that have energies better than (or similar to) the cut-off can be considered as potential binders. In cases where control molecules are unknown, shortlisting the ligands is challenging and a consensus of more than one docking method can be employed. The intersection of the top “N” best scoring ligands from various docking software can be further subjected to the structural superimposition of the protein to calculate the ligand RMSD between poses predicted by different docking tools. All compounds that have ligand RMSD better than a preset threshold can be shortlisted for further validation. Such a jury approach ensures predictions with increased confidence [8, 39].

For the Nipah glycoprotein, a subset of small molecules from the ZINC12 database was scanned on the DEPTH predicted binding sites using two docking software, Dock and AutoDock. 9 putative ligands were identified from the top scoring 150 molecules that overlap between Dock and Autodock runs. Such small molecules can then be experimentally tested to confirm their inhibitory activity.

An alternative to docking for finding the exact binding pose of a particular small molecule onto a given target protein is searching a structural database for regions of geometric and physico-chemical similarity of the binding pocket [40].

3 Modelling Protein–Peptide Complexes

Several proteins such as MHCs and membrane proteins interact with peptides [41, 42]. Such interactions are estimated to account for 15–40% of known protein–protein interactions [43]. Because peptides are usually associated with low levels of toxicity and are easy to synthesize [44], they make for attractive therapeutic agents [45]. In this section, we explore the different approaches for modelling protein–peptide complexes.

3.1 Predicting Binding Sites for Peptide Ligands

Similar to the modelling approaches described in protein–small molecule ligand modelling (Refer Subheading 2), some protein–peptide complex modelling methods require the binding site information. ACCLUSTER (<http://zougrouptoolkit.missouri.edu/accluster>) [46] is one of the several software [47–49] that can be used to predict peptide-binding sites on the surface of a given protein (*see Note 3*). ACCLUSTER uses the standard 20 amino acids as probes to detect the poses that form stable chemical interactions with the protein surface. These poses are spatially clustered, and the largest clusters are predicted as potential binding sites. We tested the ability of ACCLUSTER to predict the peptide-binding sites on HLA-B27 major histocompatibility complex that is known to bind to antigenic peptides. Starting with a crystal structure of HLA-B27 (PDB ID 6PYL), without its peptide ligand, the true antigenic peptide-binding site was one of the predictions.

3.2 Modelling Protein–Peptide Complexes

As with most of the methods that deal with modelling complexes, the input here is the known 3D structure of the target protein. The method of choice would depend on the information available about the peptide. If the structure and sequence of the peptide is not known, the structure of protein–peptide complex can be predicted using tools such as SPOT peptide [50]. If the sequence of the binding peptide is known, the 3D structure of the complex can be modelled using tools such as GalaxyPepDock [50], Rosetta FlexPepDock [52], and HADDOCK [53] (Refer to Table 2 for various methods of protein–peptide complex modelling).

3.2.1 Predicting the Sequence of the Peptides and the Structure of the Protein–Peptide Complex

The methods in this category fall into two classes, (a) knowledge-based [67, 68] and (b) de novo [54, 69, 70]. Knowledge-based methods make use of known structural information to predict the structure. The de novo methods, however, are independent of the known structural information and generally make use of physics-based principles to predict the structure of the complex. In this section, we describe a prediction of the peptide that is most likely to bind histone transferase (Histone-lysineN-methyl transferase 2A) using the knowledge-based method, SPOT-Peptide (<http://sparks-lab.org/tom/SPOT-peptide>) [50]. The 3D structure of

Table 2
Non-exhaustive list of protein–peptide complex modelling methods

Tool	Algorithm	URL
Pro_Ligand [54]	De novo	Not available
SPOT-Peptide [50]	Knowledge-based	http://sparks-lab.org/tom/SPOT-peptide/
GalaxyPepDock [51]	Template-based docking	http://galaxy.seoklab.org/pepdock
Rosetta FlexPepDock [52]	Local docking	http://flexpepdock.furmanlab.cs.huji.ac.il/
DynaDock [55]		Not available
PepCrawler [56]		http://bioinfo3d.cs.tau.ac.il/PepCrawler/
HADDOCK peptide docking [57]		http://milou.science.uu.nl/services/HADDOCK2.2/haddock.php
PEP-FOLD 3 [58]		http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD3
AutoDock Vina [36]		http://vina.scripps.edu/download.html (standalone)
DINC 2.0 [58]		http://dinc.kavrakilab.org/
Surflex-Dock [60]		https://omictools.com/surflex-dock-tool (standalone)
pepATTRACT [61]		http://bioserv.rpbs.univ-paris-diderot.fr/services/pepATTRACT/
MDockPeP [62]		http://zougrouptoolkit.missouri.edu/mdockpep/
CABS-dock [63]	Blind docking	http://biocomp.chem.uw.edu.pl/CABSdock
AnchorDock [64]		Not available
ClusPro PeptiDock [65]		https://peptidock.cluspro.org/
PIPER-FlexPepDock [66]		http://piperfpd.furmanlab.cs.huji.ac.il/

the histone transferase is used as the target protein. SPOT-Peptide superimposes this target protein on a library of known peptide-binding proteins to identify suitable templates. Models are built using these templates and are assessed using DFIRE [71] and evolutionary alignment score. The models that are favorably scored from either of the scoring schemes are then filtered by a score based on template similarity, SP-score to get final predicted models.

All the predictions are associated with the three assessment scores and a list of residues of the protein that interact with the peptide. SPOT-peptide was able to successfully reproduce the transferase and peptide complex as one of the top predictions. The predicted complex is comparable to the crystal structure, with a peptide backbone RMSD of ~ 2.5 Å.

3.2.2 Docking Peptides onto Target Proteins

Docking a given peptide onto a protein can be guided by a template. Template-based methods rely on structures of homologous complexes to model the 3D structure. If homologous templates are unavailable, template-independent docking algorithms are employed [71].

Template-Based Docking of Protein–Peptide Complexes

GalaxyPepDock (<http://galaxy.seoklab.org/pepdock>) [51] is a template-based docking program that uses structural similarity of the protein and sequence similarity of the peptide to identify the templates. To predict the complex of ubiquitin Nedd4 with the peptide PPXY (a motif of arrestin-related domain-containing protein-3), GalaxyPepDock takes a 3D structure of the ubiquitin and the sequence PPXY as inputs. Multiple models are generated by GalaxyTBM [72, 73] for each homologous template identified by structural and interaction similarity. Top 10 best energy models for each template are refined by energy-based optimization and are presented as final predicted models. The predicted complexes are associated with details such as templates used for protein and peptide, sequence alignments, structure similarity score, interaction similarity score, accuracy, and the residues on the protein predicted to interact with the peptide. The predicted model for ubiquitin and motif peptide complex (excluding the known crystal structure template) was built using a template with high structural similarity assessed by a metric called TM-score [74].

Local Docking of Protein– Peptide Complexes

Given a peptide sequence and a protein structure on which a binding pocket has been identified (Refer to Subheading 3.1), local docking can be used to predict the 3D structure of the complex. One such method is Rosetta FlexPepDock (<http://flexpepdock.furmanlab.cs.huji.ac.il/>) [52, 75]. The input is an approximate protein–peptide complex (*see Note 4*) where the peptide is placed near the binding pocket. The initial complex can be built using standard homology modelling tools. If the structure of homologs is not available, an initial peptide conformation can be manually constructed and placed in the vicinity of the binding site using tools such as Chimera [64]. Rosetta FlexPepDock refines the initial complex structure in 200 independent FlexPepDock simulations. 100 of these are performed in a high-resolution mode, whereas, the other 100 are performed with a low-resolution pre-optimization followed by a high-resolution refinement step. These are then ranked according to the Rosetta full-atom energy score. Ten best scoring complexes are presented as final predictions.

Along with the initial approximate model, atomic constraints, if known, can also be provided. To better assess the predicted structure, a reference structure can be used as a comparison standard. The reference structure is often a structure of a similar interaction and is used to calculate RMSDs of the predicted complex. If the reference structure is not given as an input, RMSDs are calculated with respect to the starting conformation (input protein–peptide complex). Users can select the representative atoms for RMSD calculation, the default selection is peptide backbone heavy atoms.

Blind Docking of Protein–Peptide Complexes

When little or nothing is known about the peptide-binding site or the peptide conformation, we can take recourse to blind docking. The software AnchorDock performs blind docking by employing a variation of molecular dynamics (MD) simulations [64, 76]. The inputs are the structures of the target protein and a peptide with an extended initial conformation. A free peptide folding simulation is performed with explicit solvent to get a peptide conformation for docking. It localizes the conformational space by identifying the most probable peptide-binding regions on the surface of the protein called anchoring spots using ANCHORS MAP [77]. Once the anchoring spots are identified, an anchor-driven simulated annealing simulation is applied to the free peptide conformation around the anchoring spots. The simulation trajectories are clustered based on backbone RMSD and ranked based on the average potential energy of the system to get the final protein–peptide complexes (the one with the least energy). Refer to https://link.springer.com/protocol/10.1007/978-1-4939-6798-8_7 for a detailed protocol [77].

3.3 Assessing Predicted Models with Various Scoring Schemes

The modelled complexes can then be assessed by various protein–peptide complex scoring schemes such as the FoldX suite that computes the interaction energy. In principle, all the protein–protein assessment scores can also be used here. For more details on this, refer to Subheading 4.4.

4 Modelling Protein–Protein Complexes

Only ~6% of all estimated protein interactions have experimentally solved structures in the PDB leaving a substantial number of them structurally uncharacterized [78]. Computational methods can aid in modelling these uncharacterized structures. Similar to protein–peptide complex modelling methods, the protein–protein complex modelling methods have two broad categories, i.e., template-based prediction and docking. The section below describes some of these methods for dimeric complexes. Modelling of multimeric interactions is covered in Subheading 5.

4.1 Template-Based Prediction of Structure of a Protein–Protein Complex Given Structures of the Target Proteins

PRISM [78] (<http://cosbi.ku.edu.tr/prism/>) is one of the several (Refer Table 3) template-based docking programs that predict the structure of the complex when the structures of both the target proteins are known. PRISM was used to model the falcipain–cystatin complex [79]. Falcipain is a cysteine protease that is inhibited by cystatin. PRISM takes the structures of the targets falcipain

and cystatin as inputs. The surface of the targets is then scanned through a library of known protein–protein interfaces to identify a template interface (based on the structural match). Models are built using the identified template interface and are assessed using an energy function, FiberDock [92]. The lower the energy, the better is the model. PRISM built the best scoring model for the falcipain–cystatin complex using a template of Cathepsin B (a cysteine protease) and stefin A (inhibitor of cysteine protease) complex. The predicted model had the binding regions and relative orientation of the two proteins similar to that of the native falcipain–cystatin complex (PDB ID:1YVB) [93].

In addition to the structure of targets, the template interface, if known, can also be provided as an input. PRISM will then only

Table 3
Non-exhaustive list of protein–protein complex modelling methods

Tool	Algorithm	URL
Interactome 3D [78]	Template-based	http://interactome3d.irbbarcelona.org/
HOMCOS [80]	Template-based	http://strcomp.protein.osaka-u.ac.jp/homcos/
PRISM [79]	Template-based	http://prism.cccb.ku.edu.tr/
iWRAP [81]	Template-based	http://groups.csail.mit.edu/cb/iwrap/
InterPreTS [82]	Template-based	http://www.russelllab.org/cgi-bin/tools/interprets.pl
SPRING [83]	Template-based	http://zhanglab.ccmb.med.umich.edu/spring/
Struct2Net [84]	Template-based	http://groups.csail.mit.edu/cb/struct2net/webserver/
Coev2Net [85]	Template-based	http://groups.csail.mit.edu/cb/coev2net/
COTH [86]	Template-based	http://zhanglab.ccmb.med.umich.edu/COTH/
ZDOCK [87]	Docking	http://zdock.umassmed.edu/
Hex [88]	Docking	http://hexserver.loria.fr/
ClusPro [89]	Docking	https://cluspro.bu.edu/login.php
HADDOCK [90]	Docking	http://www.bonvinlab.org/software/haddock2.2/
InterEVDock2 [91]	Docking	http://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock2/

sample over the specified interface instead of the entire library of protein–protein interfaces.

4.2 Template-Based Prediction of Structure of a Protein Complex When the Structures of the Constituent Target Proteins Are Not Known

To predict the structure of the protein–protein complex, where the structure of the two (or more) target proteins and that of their complex is unknown, template-based prediction methods (Refer to Table 3) can be used. HOMCOS (<http://homcos.pdbj.org/>) [79, 93] is one such method that is based on dimeric threading. It uses homologous dimeric templates to predict structures of complexes. Here, we use an example of constructing a complex of kinase CDK5 and Cyclin B1, with which it is known to interact specifically, as described comprehensively elsewhere [94]. To predict the structure of their complex, their structures (if structures are known) or sequences are inputs to the HOMCOS server. The HOMCOS server identifies the homologous dimeric templates for the target proteins CDK5 and Cyclin B1, by performing two rounds of BLAST over the PDB database, one for each of the given target. Of the detected dimeric templates, only those that involve homologs of both, CDK5 and Cyclin B1 are used to build models. The models are associated with statistics such as the percentage identity of aligned residues and contact residues, the number of aligned contact residues and number of contact residues in the template homolog. Model selection can be assisted by these statistics and is at the discretion of the user.

The HOMCOS server depends on dimeric homologous templates to predict structures. In the absence of such templates, monomer threading followed by oligomer mapping approaches can be used. SPRING(<https://zhanglab.ccmb.med.umich.edu/spring/>) [83] is one such method (Refer to Table 3) that models dimeric complexes. SPRING was used to model a homodimeric complex of a peroxidase, 1-Cys peroxiredoxin [83]. To model the homodimeric complex, SPRING takes sequences of the two targets as inputs. In this case, the sequence of 1-Cys peroxiredoxin is used as both the targets. For each of the query proteins, the SPRING algorithm searches templates for threading. The target sequences are threaded onto each of the interacting monomers of the template. The models are then evaluated based on the SPRING score that is a composite of a threading Z-score, a structural alignment score (TM-align score), and a contact-based potential. Models are ranked based on the SPRING score. The best scoring model of the dimeric complex of 1-Cys peroxiredoxin had a TM-score of 0.75 and interface RMSD of 3 Å (*see Note 5*).

4.3 Protein–Protein Docking

Protein–protein docking can be used when no suitable templates for modelling the protein–protein complexes are available. Docking samples various conformations/configurations in which the two proteins can associate with each other and scores them to identify the most probable mode/pose of association.

Similar to protein–small molecule docking, protein–protein docking can be either local, here the search is localized with the help of user provided restraints, or blind, where the entire surface of the protein is sampled. The following section deals with local and blind docking of protein–protein complexes.

4.3.1 *Restraint-Based Local Docking of Protein– Protein Complexes*

As mentioned earlier (Subheading 2.2), local docking methods try to localize (restrict) the conformational sampling space. In protein–protein local docking, the search space can be restricted using user provided restraints. The restraints can be a list of interacting residues of the two proteins, or more specifically be distances between specific amino acids. Such restraints are often extracted from experimental data. Computational methods such as CPORT [95] (*see Note 6*), BIPSPI [96], EVcoupling complex [97] (*see Note 7*) among others, can also be employed to predict the restraints.

Local docking can now be performed using the identified restraints. HADDOCK [53] (<http://milou.science.uu.nl/services/HADDOCK2.2/>) is one of the several software/web servers that perform local docking. The structure of the two target proteins and the restraints are inputs to HADDOCK. The residues that are known to contribute to the interaction but are of limited importance, called passive residues, can also be specified or HADDOCK can automatically select them. HADDOCK samples docking poses and performs clustering based on the pose similarity. All the clusters are provided as output. The best cluster is the one with the lowest HADDOCK and Z-score. The server also provides values for electrostatic, desolvation, Van der Waals, and restraint violation energies (*see Note 8*).

4.3.2 *Blind Docking of Protein–Protein Complexes*

In the absence of reliable restraints, blind docking can be performed. Blind docking involves prediction of the structure of the protein–protein complex without any prior knowledge of interacting residues or restraints. The Z-dock web server [87] (<http://zdock.umassmed.edu/>) performs such blind docking (Refer Table 3 for other methods). It takes the structure of the two target proteins as input. If the structure is provided in the form of PDB IDs, the entire biological assembly or specific chains can be used for docking. It uses rigid body docking to sample conformations of the two targets onto each other. The docking poses are evaluated based on a score that involves shape complementarity, electrostatics, and statistical potential terms. The top “N” docking poses can be further evaluated based on the user’s choice. Z-DOCK also provides the facility to select residues that can be part of the binding site or can be excluded from the binding site.

4.4 *Evaluating Protein–Protein Complexes*

Most, if not all models that are built or predicted are scored based on their in house/known scoring schemes. The complexes can be evaluated by various independent scoring schemes to gain higher confidence in the prediction. The PIZSA [130, 131] web server

(<http://cospi.iiserpune.ac.in/pizsa/>) predicts if the complex is a binder/non-binder using a knowledge-based statistical potential. The predicted complexes can be uploaded on the web server (*see Note 1*). A distance cut-off threshold for interface residue definition can be chosen between 4, 6, and 8 Å. The best results are obtained at 4 Å. A Z-score value of greater than 1.2 indicates a stable association.

Another scoring scheme from the FoldX suite [132] (<http://foldxsuite.crg.eu/>) can be used to assess the interaction by calculating the binding-free energy. FoldX is an empirical force field developed for the fast evaluation of protein complexes. The standalone version can be installed and the protein complex can be evaluated using it. A negative value indicates a feasible interaction.

5 Modelling Protein–Nucleic Acid Complexes

Protein–nucleic acid interactions regulate various processes such as gene expression, DNA repair, replication of the DNA/RNA, and several others [133]. Structures of protein–nucleic acid complexes are hence vital to get insights into the molecular mechanism of these processes. This section describes computational methods for predicting the structures of complexes of proteins with DNA/RNA. As with other sections, protein–nucleic acid modelling also has two broad categories, template-based modelling and docking.

5.1 Template-Based Modelling of Protein–Nucleic Acid Complexes

Template-based modelling is preferred over docking in the presence of a suitable template [134]. The template-based methods are of two types, homology modelling [123] and fragment-based assembly [133].

5.1.1 Homology Modelling of Protein–Nucleic Acid Complexes

In the presence of homologous templates, methods such as TFmodeller [123] (Refer to Table 4 for other methods) can be used to model a protein–DNA complex. TFmodeller takes the FASTA sequence of a protein as input. The template to be used (if known) can be provided as an input. If not, TFmodeller identifies homologous templates using PSI-BLAST. The homologs are searched in a library of protein–DNA complexes obtained from the PDB. Each of the identified templates is used to build a complex of the query protein with the template DNA. The predicted models, a matrix of homologous interface contacts, the alignment used for the creation of the complex and a list of query positions interacting with the nucleotide along with their conservation are presented as output.

For modelling protein–RNA complexes (Refer to Table 4 for various methods of protein–RNA complex modelling), MPRDock

Table 4
Non-exhaustive list of some protein–nucleic acid complex modelling methods

Tool	Algorithm	Type of nucleic acid	URL
MODELLER [98]	Homology modelling	Protein/DNA/RNA	https://salilab.org/modeller/download_installation.html (standalone)
Prime 2.0 [99]		RNA	http://www.rnabinding.com/PRIME/PRIME2.0.html
MPRDock [100]		RNA	http://huanglab.phys.hust.edu.cn/mprdock/
RNA secondary structure prediction			
RNAfold [101]	Dynamic programming	RNA	http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi
Modelling nucleic acid structure			
3DNA [102]	Geometry-based modelling	DNA/RNA	http://web.x3dna.org/index.php/fibermodel
3D-DART [103]		DNA	http://milou.science.uu.nl/services/3DDART/
3D-Nus [104]		DNA/RNA	https://iitb.ac.in/3dnus/DNA%20Mismatch.html
SimRNA [105]		RNA	https://genesilico.pl/SimRNAweb/submit
ModeRNA [106]	Template-based modelling	RNA	http://iimcb.genesilico.pl/modernaserver/submit/model/
Binding site prediction			
DBSI [107]	Structure-based	DNA	https://mitchell-lab.biochem.wisc.edu/DBSI_Server/upload.php
DISPLAR [108]		DNA	https://pipe.rcc.fsu.edu/display.html
DR_Bind [109]		DNA	http://dnasite.limlab.ibms.sinica.edu.tw/
DNABINDPROT [110]		DNA	www.prc.boun.edu.tr/appserv/prc/dnabindprot/
PRNA [111]		RNA	http://doc.aporc.org/wiki/PRNA (standalone)
aaRNA [112]		RNA	https://sysimm.ifrec.osaka-u.ac.jp/aarna/

DRNAPred [113]	Sequence-based	DNA/RNA	http://biomine.cs.vcu.edu/servers/DRNAPred/
DP-Bind [114]		DNA	http://leg.rut.albany.edu/dp-bind/
Pprint [115]		RNA	https://webs.iitd.edu.in/raghava/pprint/submit.html
PRIdictor [116]		RNA/protein	http://bclab.inha.ac.kr/pridictor/pridictor.html
RNApin [117]		RNA	https://webs.iitd.edu.in/raghava/rnapin/submit.php
PROMO [118, 119]		DNA	http://algen.lsi.upc.es/egi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
TFBind [120]		DNA	http://tfbind.hgc.jp/
ConTra v3 [121]		DNA	http://bioit2.irc.ugent.be/contra/v3/#/step/1
CiiDER [122]		DNA	http://www.ciiider.org/ (standalone)
Modelling DNA-protein complex (structure of protein with DNA does not exist)			
TFModeller [123]	Homology modelling	DNA-protein	http://maya.ccg.unam.mx/~tfmodell/index.html
HADDOCK [53]	Knowledge-based docking	Nucleic acid-protein or protein-protein	https://milou.science.uu.nl/services/HADDOCK2.2/haddock.php
[124]	Rigid body docking	Nucleic acid-protein or protein-protein	http://www.sbg.bio.ic.ac.uk/docking/download.html (standalone)
ParaDock [125]		Nucleic acid-protein	http://bioinfo3d.cs.tau.ac.il/ParaDock/php.php
NPDock [126]	Rigid body knowledge-based docking	Nucleic acid-protein	http://genesilico.pl/NPDock
3dRPC [127]		RNA-protein	http://biophy.hust.edu.cn/3dRPC
HDock [128]		Nucleic acid-protein/protein-protein	http://hdock.phys.hust.edu.cn/
PatchDock [129]		Nucleic acid-protein/protein-protein	https://bioinfo3d.cs.tau.ac.il/PatchDock/

(<http://huanglab.phys.hust.edu.cn/mprdock/>) [100] uses a combination of template-based modelling and docking. MPRDock allows flexibility of protein side chains by considering an ensemble of protein structures that are modelled based on homologous templates. The RNA is considered as a rigid entity and is docked on each protein from the ensemble. The docked complexes are evaluated by an inbuilt scoring function. The lower the score, the better is the model. The input for MPRDock is the structure of RNA and structure or FASTA sequence of the protein. The binding interface and distance restraints (between amino acid and nucleotide residues) if known, can also be provided. The output consists of all the modelled protein–RNA complexes along with their energy values.

5.1.2 Fragment-Based Modelling of Protein– Nucleic Acid Complexes

Protein-assisted DNA assembly [133] is a fragment-based method that can be used to predict the DNA–protein complex or DNA–binding site on a protein. It has a library of small fragments of proteins (length of 6–12 amino acids) along with their interacting dsDNA (length of 4–8 base pairs) obtained from the known DNA–protein complexes. An empirical interaction model generator performs docking using this library to build docking models. The models are then scored and filtered using a statistical knowledge-based force field (*see Note 9*).

Similar to protein-assisted DNA assembly, RNAX [135] is a fragment-based method for docking of RNA fragments. Refer to the tutorial <http://modelx.crg.es/PADAI/Tutorial> for details of the commands for both RNAX and protein-assisted DNA assembly.

5.2 Docking of Protein–Nucleic Acid Complexes

Protein–nucleic acid docking methods can be knowledge based or ab initio. Knowledge-based methods can be applied if the information of the interface region is known; otherwise, ab initio methods are used.

5.2.1 Knowledge-Based Docking of Protein–Nucleic Acid Complexes

Knowledge-based docking uses information about the interface residues in the protein. The interface residues can be inferred from experiments or can be predicted computationally. A variety of sequence and structure-based algorithms can be used to predict these interface residues on DNA/RNA and on protein (Refer to Table 4 for the methods). These interface residues can be specified as inputs to HADDOCK (<https://milou.science.uu.nl/services/HADDOCK2.2/haddock.php>) for docking DNA/RNA on protein [53, 136]. HADDOCK takes structures of both, the protein and the DNA/RNA as inputs (*see Note 10*). With the specified input structures and restraints, HADDOCK performs rigid docking followed by semi-flexible and solvent refinements. The docking models are clustered based on structural similarity to one another (RMSD). The final clusters (predicted models) are selected based on the HADDOCK scoring function.

5.2.2 Blind Docking of Protein–Nucleic Acid Complexes

NPDock [126] is an exclusively designed nucleic acid–protein docking method that can be used when the sequence of DNA that binds the protein of interest is known. It has been employed to characterize novel transcription factors such as PvDREB1A [137]. NPDock accepts structures of DNA/RNA and proteins as inputs. In NPDock, DNA/RNA–protein rigid body docking is performed using GRAMM [138]. The docked RNA–protein complexes are scored using statistical potentials DARS-RNP and QUASI-RNP [139], while DNA–protein complexes are scored using a combination of QUASI-DNP, DFIRE [71], and Varani group potential [25] for DNA–protein complexes. The best scoring models are clustered based on structural similarity and refined using a simulated annealing protocol. The predicted models are the best scoring complexes in the three biggest clusters. The clash score of the best model of the biggest cluster is provided along with the plot for the change in score across the duration of the simulation.

To get better confidence in the models generated by various software, the models can be further assessed using the Evaluate-Complex function of ModelX. The command line parameters to be used are mentioned in the ModelX tutorial (<http://modelx.crg.es/PADATutorial>).

6 Modelling Macromolecular Assemblies Containing Various Biomolecules

Macromolecular assemblies are biological structures with dimensions in the range of few nanometers to micrometers. They consist of various proteins, peptides, nucleotides, etc. that together act as a functional unit. Elucidating the 3D structures of these macromolecular assemblies is crucial to understand their mechanism. Experimental methods of determining structures of assemblies are challenging due to the complexity and heterogeneity of the assemblies. Computational methods such as integrative modelling can aid in determining the structure of these assemblies. Integrative modelling uses various inputs obtained from multiple experiments, statistical analysis, etc. to model the structure of the assembly [140–142]. It follows a four-stage process that involves data collection, representation and evaluation of models, sampling conformations, and validation. These four stages are iterated until ensemble(s) of structures that satisfy the input restraints are found. The following sections describe each of these steps.

6.1 Data Collection

This stage involves finding data that describes the assembly. The description involves identifying copy number, shape, and localization of each unique component, shape and symmetry of the overall assembly, relative orientations, envelope surface, and contacts between the components. These data can be obtained from

Table 5

Very few software suites do most/all steps of integrative modelling. This table is a list of methods that could be used for the conversion of experimental data into spatial restraints for macromolecular assembly modelling. References to studies where such methods were utilized is provided in the last column

Experimental technique	Measured data	Structural data	Example reference
Chemical cross-linking	Mass/charge ratio of joint fragments	Upper limit on pair distance between reacted groups	[143–145]
Forster resonance energy transfer (FRET)	The yield of fluorescence energy transfer	Distance between donor–acceptor pairs	[146]
Electron paramagnetic resonance (EPR)	Dipole–dipole coupling between electron spins	Distance between pairs of spin labels	[147, 148]
Small angle X-ray scattering (SAXS)	Scattering intensity as a function of momentum transfer	Pair distribution function or shape envelope	[149, 150]
EM and Cryo-EM	Shape envelope	Volume restraints	[151–154]
Deuterium exchange mass spectroscopy (DXMS)	Rate constant of H/D exchange	Solvent exposure	[155]
Radical footprinting	Rate constant from the dose–response curve	Solvent exposure	[156]
Circular dichroism (CD)	Mean residue ellipticity as a function of wavelength	Secondary structure content	[157]

different independent experiments. For instance, the overall shape and symmetry of the macromolecular assembly can be obtained by Electron Microscopy (EM) or Cryo-Electron Microscopy (Cryo-EM) (Refer Table 5 for data that can be extracted from various experiments). Along with experimental data, computational data such as homology models of individual components, and statistical inferences from bioinformatics data can also be used to model the 3D structure of the assembly. The quality and quantity of the collected data affect the accuracy of the generated models (*see* Notes 11 and 13).

6.2 Data Representation and Model Evaluation

The data collected in the previous stage is represented as spatial restraints for modelling (Refer Table 5). In cases where experimental data are not available, computational techniques play a dominant role in determining the inter-component structural data. Several methods that model protein–protein, protein–DNA/RNA complexes [Refer Subheading 4.3.1 and Subheading 5.2.1] can provide spatial restraints between the interacting components that can be used for macromolecular complex modelling.

The features being restrained include angles, distances, and relative orientations. These restraints are in the form of probability density functions that describe the assembly. All the specified restraints are combined into a scoring scheme and used to evaluate the generated conformations.

Integrative modelling platform (IMP) is one of the earliest software to perform integrative modelling. We use IMP to illustrate the workflow of the integrative modelling method. IMP provides IMP:Model and IMP:Restraint modules [158, 159] to facilitate the representation of experimental data into spatial restraints. These modules can represent different experimental data to a single and compatible platform for representation and scoring.

6.3 Sampling and Optimization

Two different protocols can be followed depending on the symmetric or non-symmetric nature of the macromolecular assemblies. In a symmetric complex, individual components follow a symmetrical pattern such as linear, spiral, circular (e.g., Rad51, Microtubules, Actin filaments). In a non-symmetric complex, the different components do not follow a regular pattern (e.g., Ribosome, Proteasome, Chromatin, Intermediate Filaments).

If the macromolecular complex is symmetric, then the symmetry restraint between the repeating units provide a symmetrical axis. Rigid body transformation of the repeating units around the symmetrical axis can be done using CLICK (a topology-independent structural superimposition program) [160, 161] to create a complete model of the multi-component macromolecular structure.

If the macromolecular complex is not symmetric, then various conformations are sampled followed by optimization (*see Note 12*). The computational assembly starts with sampling a random configuration. The scoring scheme constructed in Subheading 5.2 evaluates the 3D structure/model. An optimizer minimizes the violated restraints, and the final score defines the quality of the optimized models.

Depending on the type of experimental data, several methods exist in the IMP package and other softwares for computational optimization [5]. For instance, the IMP:MultiFit module for multi-component molecular docking and fitting on EM maps [1], IMP:EmageFit module uses available subunit structures and EM class averages [162], IMP:MultiFoXS for multi-state models using SAXS data [52].

6.4 Ensemble Analysis

Models from Subheading 6.3 are clustered based on structural similarity to get ensembles. Analyzing these ensembles allows us to evaluate the quality of the models. The analysis involves the assessment of probability distributions of component properties such as positions, contacts, and localization. Single peak distributions with a small standard deviation indicate precise input information. Lack of such single peak distributions indicates the

possibility of alternate configurations/conformations or inconsistent input data. In such cases, the entire exercise can be repeated leaving fewer or different sets of restraints for validation or alternatively by getting more information about the assembly to get more restraints. If the ensemble analysis shows satisfactory results (*see* **Notes 13** and **14**), then the model can be further validated by experimental testing.

7 Notes

1. The computational techniques are sensitive to clashes and orientations of side chains in the initial input models. The input models should be free of clashes. To remove the clashes chimera can be used. Open the structure in Chimera and go to Tools ->Structure editing ->Energy minimize. A more elaborate energy minimization can also be done to remove clashes and improve interactions using GROMACS [163, 164] (Please follow <http://www.mdtutorials.com/> till energy minimization).
2. Model the protein structure to fill in missing atoms/residues (complete PDB) before predicting the binding pockets. Do not add hydrogens to the structure while predicting the binding pockets using the DEPTH server. Use the complete PDB with no missing atom for docking.
3. It is recommended to pre-check the PDB file for the presence of mutated non-standard residues. The PDB file should have at least 31 and maximum 1000 amino acids. Additional inputs such as the peptide sequence and the residues that are away from the binding site if known can be provided. These additional inputs improve the computational efficiency of the method.
4. The protein-peptide complex for Rosetta FlexPepDock should not contain any heteroatoms.
5. The models built using SPRING only contain C-alpha atoms. The complete models can be further built using the predicted model as a template using MODELLER (Please follow MODELLER tutorial on Basic modelling at <https://salilab.org/modeller/tutorial/basic.html>).
6. CPORT over predicts the interface residues. In cases if a large number of interface residues are predicted, one can just take the interface residue prediction from any of the servers that CPORT uses to make the prediction.
7. The coupling file provided by EVcouplings contains information of both inter- and intra-target protein couplings. It is

important to filter the table to only extract information about inter-protein coupling.

8. HADDOCK server has multiple services based on the type of restraint data. The easy interface is used when the number of interface residues are less and we are confident about them. The prediction interface should be used with tools that over predict the interface such as CPROT. Several restraints such as ambiguous interaction restraints, dipolar coupling restraints, pseudo contact restraints, etc. can also be utilized in the Expert and Guru interface. Details about restraints that can be set up using the Expert and Guru Interface can be found in HADDOCK manual (<https://www.bonvinlab.org/software/haddock2.2/manual/>). Increasing the number of restraints can help reduce docking sampling space and improve prediction accuracy.
9. The MYSQL dumps are extremely large and you may need to install MYSQL in an external device with at least 50GB space available.
10. Important to note that the web server asks for residue numbers. So if the protein has multiple chains and the residue numbers overlap, it can create a problem. So the residues must be renumbered so that they are unique.
11. The amount and quality of the data collected can significantly increase or decrease the accuracy of the models. Thus, the data for building and validating the models should be balanced in terms of quality and quantity.
12. Non-symmetric macromolecular complexes need to be sampled extensively during optimization compared to symmetric complexes. Inappropriate sampling and scoring strategy may present convergence issues to optimizing algorithms and can lead to incorrect models. Thus, obtaining symmetry restraints (if present) can significantly improve the model quality.
13. The clustering of the ensemble can lead to three possible outcomes. (1) A single cluster satisfies all restraints; this implies that the data is sufficient for determining the unique native structure. (2) Two or more clusters satisfy the restraints; this implies that data is insufficient to resolve a unique native structure or there are multiple conformations of the system. (3) No cluster satisfies the restraints; this implies that either the data is wrong or there has been an error in data interpretation.
14. Integrative modelling uses experimental data having different resolutions to construct a 3D model. Thus, different parts of the macromolecular complex have different resolution and accuracy.

References

1. Lasker K, Topf M, Sali A, Wolfson HJ (2009) Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *J Mol Biol* 388:180–194
2. Greenberg CH, Kollman J, Zelter A et al (2016) Structure of γ -tubulin small complex based on a cryo-EM map, chemical cross-links, and a remotely related structure. *J Struct Biol* 194:303–310
3. Carlsson J, Coleman RG, Setola V et al (2011) Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nat Chem Biol* 7:769–778
4. Kuroda D, Shirai H, Jacobson MP, Nakamura H (2012) Computer-aided antibody design. *Protein Eng Des Sel* 25:507–522
5. Soni N, Madhusudhan MS (2017) Computational modeling of protein assemblies. *Curr Opin Struct Biol* 44:179–189
6. <https://www.ebi.ac.uk/training/online/course/introduction-metabolomics/what-metabolomics/no-glossary-small-molecules-no-glossary>. Accessed 30 Jun 2020
7. McFedries A, Schwaid A, Saghatelian A (2013) Methods for the elucidation of protein-small molecule interactions. *Chem Biol* 20:667–673
8. Sen N, Kanitkar TR, Roy AA et al (2019) Predicting and designing therapeutics against the Nipah virus. *PLoS Negl Trop Dis* 13: e0007419
9. Kim S, Thiessen PA, Bolton EE et al (2016) PubChem substance and compound databases. *Nucleic Acids Res* 44:D1202
10. Irwin JJ, Shoichet BK (2005) ZINC – a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45:177–182
11. Wishart DS, Feunang YD, Guo AC et al (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 46:D1074–D1082
12. Gaulton A, Hersey A, Nowotka M et al (2017) The ChEMBL database in 2017. *Nucleic Acids Res* 45:D945
13. Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. *J Chem Educ* 87:1123–1124
14. Kanehisa M, Furumichi M, Tanabe M et al (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 45:D353
15. Hastings J, de Matos P, Dekker A et al (2012) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* 41: D456–D463
16. Feng Z, Chen L, Maddula H et al (2004) Ligand depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* 20:2153–2155
17. Stank A, Kokh DB, Fuller JC, Wade RC (2016) Protein binding pocket dynamics. *Acc Chem Res* 49:809–815
18. Tan KP, Nguyen TB, Patel S et al (2013) Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. *Nucleic Acids Res* 41. <https://doi.org/10.1093/nar/gkt503>
19. Konc J, Miller BT, Štular T et al (2015) ProBiS-CHARMMing: web Interface for prediction and optimization of ligands in protein binding sites. *J Chem Inf Model* 55:2308–2314
20. Jendele L, Krivak R, Skoda P et al (2019) PrankWeb: a web server for ligand binding site prediction and visualization. *Nucleic Acids Res* 47:W345–W349
21. Hussein HA, Borrel A, Geneix C et al (2015) PockDrug-server: a new web server for predicting pocket druggability on holo and apo proteins. *Nucleic Acids Res* 43:W436–W442
22. Xu Y, Wang S, Hu Q et al (2018) CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Res* 46: W374–W379
23. Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* 10:168
24. Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* 15:359–363
25. Xu B, Yang Y, Liang H, Zhou Y (2009) An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. *Proteins* 76:718–730
26. Brady GP Jr, Stouten PFW (2000) Fast prediction and visualization of protein binding

- pockets with PASS. *J Comput Aided Mol Des* 14:383–401
27. Yang J, Roy A, Zhang Y (2013) Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* 29:2588–2595
 28. Wass MN, Kelley LA, Sternberg MJE (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res* 38:W469–W473
 29. Repasky MP, Shelley M, Friesner RA (2007) Flexible ligand docking with Glide. In: *Current protocols in bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ
 30. Lang PT, Brozell SR, Mukherjee S et al (2009) DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA* 15:1219–1230
 31. Verdonk ML, Cole JC, Hartshorn MJ et al (2003) Improved protein–ligand docking using GOLD. *Proteins* 52:609–623
 32. Cross SSJ (2005) Improved FlexX docking using FlexS-determined base fragment placement. <https://doi.org/10.1021/CI050026F>
 33. Venkatachalam CM, Jiang X, Oldfield T, Waldman M (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J Mol Graph Model* 21:289–307
 34. Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein–small molecule docking methods. *J Comput Aided Mol Des* 16:151–166
 35. Morris GM, Huey R, Lindstrom W et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 30:2785–2791
 36. Trott O, Olson AJ (2019) Autodock vina: improving the speed and accuracy of docking. *J Comput Chem* 31:455–461
 37. Allen WJ, Balias TE, Mukherjee S et al (2015) DOCK 6: impact of new features and current docking performance. *J Comput Chem* 36:1132
 38. Grosdidier A, Zoete V, Michielin O (2011) SwissDock, a protein–small molecule docking web service based on EADock DSS. *Nucleic Acids Res* 39:W270
 39. Zhou Y, Elmes MW, Sweeney JM et al (2019) Identification of fatty acid binding protein 5 inhibitors through similarity-based screening. *Biochemistry* 58:4304–4316
 40. Nguyen MN, Sen N, Lin M et al (2019) Discovering putative protein targets of small molecules: a study of the p53 activator Nutlin. *J Chem Inf Model* 59:1529–1546
 41. Krumm BE, Grisshammer R (2015) Peptide ligand recognition by G protein-coupled receptors. *Front Pharmacol* 6:48
 42. Antunes DA, Devaurs D, Moll M et al (2018) General prediction of peptide-MHC binding modes using incremental docking: a proof of concept. *Sci Rep* 8:1–13
 43. Cunningham AD, Qvit N, Mochly-Rosen D (2017) Peptides and peptidomimetics as regulators of protein–protein interactions. *Curr Opin Struct Biol* 44:59–66
 44. Du Q-S, Xie N-Z, Huang R-B (2015) Recent development of peptide drugs and advance on theory and methodology of peptide inhibitor design. *Med Chem (Los Angeles)* 11:235–247
 45. Lau JL, Dunn MK (2018) Therapeutic peptides: historical perspectives, current development trends, and future directions. *Bioorg Med Chem* 26:2700–2707
 46. Yan C, Zou X (2015) Predicting peptide binding sites on protein surfaces by clustering chemical interactions. *J Comput Chem* 36:49–61
 47. Taherzadeh G, Zhou Y, Liew AW-C, Yang Y (2018) Structure-based prediction of protein–peptide binding regions using random forest. *Bioinformatics* 34:477–484
 48. Trabuco LG, Lise S, Petsalaki E, Russell RB (2012) PepSite: prediction of peptide-binding sites from protein surfaces. *Nucleic Acids Res* 40:W423–W427
 49. Lavi A, Ngan CH, Movshovitz-Attias D et al (2013) Detection of peptide-binding sites on protein surfaces: the first step toward the modeling and targeting of peptide-mediated interactions. *Proteins* 81:2096–2105
 50. Litfin T, Yang Y, Zhou Y (2019) SPOT-peptide: template-based prediction of peptide-binding proteins and peptide-binding sites. *J Chem Inf Model* 59:924–930
 51. Lee H, Heo L, Lee MS, Seok C (2015) GalaxyPepDock: a protein–peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Res* 43:W431–W435
 52. London N, Raveh B, Cohen E et al (2011) Rosetta FlexPepDock web server—high resolution modeling of peptide–protein interactions. *Nucleic Acids Res* 39:W249–W253
 53. van Zundert GCP, Rodrigues JPGLM, Trellet M et al (2016) The HADDOCK2.2 Web Server: user-friendly integrative modeling of biomolecular complexes. *J Mol Biol* 428:720–725
 54. Frenkel D, Clark DE, Li J et al (1995) PRO-LIGAND: an approach to de novo molecular

- design. 4. Application to the design of peptides. *J Comput Aided Mol Des* 9:213–225
55. Antes I (2010) DynaDock: a new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. *Proteins* 78:1084–1104
 56. Donsky E, Wolfson HJ (2011) PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors. *Bioinformatics* 27:2836–2842
 57. Trellet M, Melquiond ASJ, Bonvin AMJJ (2013) A unified conformational selection and induced fit approach to protein-peptide docking. *PLoS One* 8:e58769
 58. Lamiable A, Thévenet P, Rey J et al (2016) PEP-FOLD3: faster *de novo* structure prediction for linear peptides in solution and in complex. *Nucleic Acids Res* 44:W449–W454
 59. Antunes DA, Moll M, Devaurs D et al (2017) DINC 2.0: a new protein-peptide docking webserver using an incremental approach. *Cancer Res* 77:e55–e57
 60. Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* 46:499–511
 61. Schindler CEM, de Vries SJ, Zacharias M (2015) Fully blind peptide-protein docking with pepATTRACT. *Structure* 23:1507–1515
 62. Yan C, Xu X, Zou X (2016) Fully blind docking at the atomic level for protein-peptide complex structure prediction. *Structure* 24:1842–1853
 63. Kurcinski M, Jamroz M, Blaszczyk M et al (2015) CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res* 43:W419–W424
 64. Slutzki M, Ben-Shimon A, Niv MY (2017) AnchorDock for blind flexible docking of peptides to proteins. Humana Press, New York, NY, pp 95–108
 65. Porter KA, Xia B, Beglov D et al (2017) ClusPro PeptiDock: efficient global docking of peptide recognition motifs using FFT. *Bioinformatics* 33:3299–3301
 66. Alam N, Goldstein O, Xia B et al (2017) High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. *PLoS Comput Biol* 13:e1005905
 67. Verschuere E, Vanhee P, Rousseau F et al (2013) Protein-peptide complex prediction through fragment interaction patterns. *Structure* 21:789–797
 68. Vanhee P, Stricher F, Baeten L et al (2009) Protein-peptide interactions adopt the same structural motifs as monomeric protein folds. *Structure* 17:1128–1136
 69. Unal EB, Gursoy A, Erman B (2010) VitAL: Viterbi algorithm for *de novo* peptide design. *PLoS One* 5:e10926
 70. Petsalaki E, Stark A, García-Urdiales E, Russell RB (2009) Accurate prediction of peptide binding sites on protein surfaces. *PLoS Comput Biol* 5:e1000335
 71. Zhou H, Zhou Y (2009) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 11:2714–2726
 72. Ko J, Park H, Heo L, Seok C (2012) GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Res* 40:W294–W297
 73. Ko J, Park H, Seok C (2012) GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. *BMC Bioinformatics* 13:198
 74. Lee H, Seok C (2017) Template-based prediction of protein-peptide interactions by using GalaxyPepDock. Humana Press, New York, NY, pp 37–47
 75. Raveh B, London N, Schueler-Furman O (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins* 78:2029–2040
 76. Ben-Shimon A, Niv MY (2015) AnchorDock: blind and flexible anchor-driven peptide docking. *Structure* 23:929–940
 77. Ben-Shimon A, Eisenstein M (2010) Computational mapping of anchoring spots on protein surfaces. *J Mol Biol* 402:259–277
 78. Stein A, Mosca R, Aloy P (2011) Three-dimensional modeling of protein interactions and complexes is going ‘omics. *Curr Opin Struct Biol* 21:200–208
 79. Baspinar A, Cukuroglu E, Nussinov R et al (2014) PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. *Nucleic Acids Res* 42:W285–W289
 80. Kawabata T (2016) HOMCOS: an updated server to search and model complex 3D structures. *J Struct Funct Genom* 17:83–99
 81. Hosur R, Xu J, Bienkowska J, Berger B (2011) IWRAP: an interface threading approach with application to prediction of cancer-related protein-protein interactions. *J Mol Biol* 405:1295–1310

82. Aloy P, Russell RB (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* 19:161–162
83. Guerler A, Govindarajoo B, Zhang Y (2013) Mapping monomeric threading to protein–protein structure prediction. *J Chem Inf Model* 53:717–725
84. Singh R, Park D, Xu J et al (2010) Struct2-Net: a web service to predict protein–protein interactions using a structure-based approach. *Nucleic Acids Res* 38:W508–W515
85. Hosur R, Peng J, Vinayagam A et al (2012) A computational framework for boosting confidence in high-throughput protein–protein interaction datasets. *Genome Biol* 13:R76
86. Mukherjee S, Zhang Y (2011) Protein–protein complex structure predictions by multimeric threading and template recombination. *Structure* 19:955–966
87. Pierce BG, Wiehe K, Hwang H et al (2014) ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics* 30:1771–1773
88. Macindoe G, Mavridis L, Venkatraman V et al (2010) HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Res* 38:W445–W449
89. Kozakov D, Hall DR, Xia B et al (2017) The ClusPro web server for protein–protein docking. *Nat Protoc* 12:255–278
90. Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* 125:1731–1737
91. Quignot C, Rey J, Yu J et al (2018) InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. *Nucleic Acids Res* 46:W408–W416
92. Mashiach E, Nussinov R, Wolfson HJ (2010) FiberDock: flexible induced-fit backbone refinement in molecular docking. *Proteins* 78:1503–1519
93. Tuncbag N, Gursoy A, Nussinov R, Keskin O (2011) Predicting protein–protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc* 6:1341–1354
94. Fukuhara N, Kawabata T (2008) HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Res* 36:W185–W189
95. de Vries SJ, Bonvin AMJJ (2011) CPORT: a consensus Interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One* 6:e17695
96. Sanchez-Garcia R, Sorzano COS, Carazo JM, Segura J (2019) BIPSPI: a method for the prediction of partner-specific protein–protein interfaces. *Bioinformatics* 35:470–477
97. Hopf TA, Schärfe CPI, Rodrigues JPGLM et al (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *elife* 3:03430
98. Eswar N, Webb B, Marti-Renom MA et al (2006) Comparative protein structure modeling using modeller. *Curr Protoc Bioinformatics* 15:5.6.1–5.6.30
99. Zheng J, Xie J, Hong X, Liu S (2019) RMA-align: an RNA structural alignment tool based on a novel scoring function RMscore. *BMC Genomics* 20:276
100. He J, Tao H, Huang S-Y (2019) Protein-ensemble–RNA docking by efficient consideration of protein flexibility through homology models. *Bioinformatics* 35:4994–5002
101. Gruber AR, Lorenz R, Bernhart SH et al (2008) The Vienna RNA websuite. *Nucleic Acids Res* 36:W70–W74
102. Li S, Olson WK, Lu X-J (2019) Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res* 47:W26–W34
103. van Dijk M, Bonvin AMJJ (2009) 3D-DART: a DNA structure modelling server. *Nucleic Acids Res* 37:W235–9. <https://doi.org/10.1093/nar/gkp287>
104. Patro LPP, Kumar A, Kolimi N, Rathinavelan T (2017) 3D-NuS: a web server for automated modeling and visualization of non-canonical 3-dimensional nucleic acid structures. *J Mol Biol* 429:2438–2448
105. Magnus M, Boniecki MJ, Dawson W, Bujnicki JM (2016) SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. *Nucleic Acids Res* 44:W315–W319
106. Rother M, Milanowska K, Puton T et al (2011) ModeRNA server: an online tool for modeling RNA 3D structures. *Bioinformatics* 27:2441–2442
107. Sukumar S, Zhu X, Ericksen SS, Mitchell JC (2016) DBSI server: DNA binding site identifier. *Bioinformatics* 32:2853–2855

108. Tjong H, Zhou H-X (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res* 35:1465–1477
109. Chen YC, Wright JD, Lim C (2012) DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res* 40: W249–W256
110. Ozbek P, Soner S, Erman B, Haliloglu T (2010) DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues. *Nucleic Acids Res* 38:W417–23
111. Liu Z-P, Wu L-Y, Wang Y et al (2010) Prediction of protein–RNA binding sites by a random forest method with combined features. *Bioinformatics* 26:1616–1622
112. Li S, Yamashita K, Amada KM, Standley DM (2014) Quantifying sequence and structural features of protein–RNA interactions. *Nucleic Acids Res* 42:10086–10098
113. Yan J, Kurgan L (2017) DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* 45:e84
114. Hwang S, Gou Z, Kuznetsov IB (2007) DP-bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* 23:634–636
115. Kumar M, Gromiha MM, Raghava GPS (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* 71:189–194
116. Tuvshinjargal N, Lee W, Park B, Han K (2016) PRIdictor: protein–RNA interaction predictor. *Biosystems* 139:17–22
117. Panwar B, Raghava GPS (2015) Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. *Genomics* 105:197–203
118. Messeguer X, Escudero R, Farre D et al (2002) PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics* 18:333–334
119. Farré D, Roset R, Huerta M et al (2003) Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. *Nucleic Acids Res* 31:3651–3653
120. Tsunoda T, Takagi T (1999) Estimating transcription factor bindability on DNA. *Bioinformatics* 15:622–630
121. Kreft Ł, Soete A, Hulpiau P et al (2017) ConTra v3: a tool to identify transcription factor binding sites across species, update 2017. *Nucleic Acids Res* 45:W490–W494
122. Gearing LJ, Cumming HE, Chapman R et al (2019) CiiiDER: a tool for predicting and analysing transcription factor binding sites. *PLoS One* 14:e0215495
123. Contreras-Moreira B, Branger P-A, Collado-Vides J (2007) TFmodeller: comparative modelling of protein–DNA complexes. *Bioinformatics* 23:1694–1696
124. Gabb HA, Jackson RM, Sternberg MJE (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* 272:106–120
125. Banitt I, Wolfson HJ (2011) ParaDock: a flexible non-specific DNA–rigid protein docking algorithm. *Nucleic Acids Res* 39: e135
126. Tuszynska I, Magnus M, Jonak K et al (2015) NPDock: a web server for protein–nucleic acid docking. *Nucleic Acids Res* 43: W425–W430
127. Huang Y, Li H, Xiao Y (2018) 3dRPC: a web server for 3D RNA–protein structure prediction. *Bioinformatics* 34:1238–1240
128. Yan Y, Zhang D, Zhou P et al (2017) HDock: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res* 45: W365–W373
129. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 33: W363–W367
130. Roy AA, Dhawanjewar AS, Sharma P et al (2019) Protein interaction Z score assessment (PIZSA): an empirical scoring scheme for evaluation of protein–protein interactions. *Nucleic Acids Res* 47:W331–W337
131. Dhawanjewar AS, Roy AA, Madhusudhan MS (2019) A knowledge-based scoring function to assess the stability of quaternary protein assemblies. *bioRxiv*:562520. <https://doi.org/10.1101/562520>
132. Schymkowitz J, Borg J, Stricher F et al (2005) The FoldX web server: an online force field. *Nucleic Acids Res* 33:W382–W388

133. Blanco JD, Radusky L, Clemente-González H, Serrano L (2018) FoldX accurate structural protein-DNA binding prediction using PADA1 (protein assisted DNA Assembly 1). *Nucleic Acids Res* 46:3852–3863
134. Xue LC, Rodrigues JPGLM, Dobbs D et al (2016) Template-based protein-protein docking exploiting pairwise interfacial residue restraints. *Brief Bioinform*:bbw027
135. Blanco JD, Radusky LG, Cianferoni D, Serrano L (2019) Protein-assisted RNA fragment docking (RnaX) for modeling RNA-protein interactions using ModelX. *Proc Natl Acad Sci U S A* 116:24568–24573
136. Karaca E, Melquiond ASJ, de Vries SJ et al (2010) Building macromolecular assemblies by information-driven docking. *Mol Cell Proteomics* 9:1784–1794
137. Vatansever R, Uras ME, Sen U et al (2017) Isolation of a transcription factor DREB1A gene from *Phaseolus vulgaris* and computational insights into its characterization: protein modeling, docking and mutagenesis. *J Biomol Struct Dyn* 35:3107–3118
138. Katchalski-Katzir E, Shariv I, Eisenstein M et al (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A* 89:2195–2199
139. Tuszyńska I, Bujnicki JM (2011) DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. *BMC Bioinformatics* 12:348
140. Alber F, Dokudovskaya S, Veenhoff LM et al (2007) Determining the architectures of macromolecular assemblies. *Nature* 450:683–694
141. Rout MP, Sali A (2019) Principles for integrative structural biology studies. *Cell* 177:1384–1403
142. Braitbard M, Schneidman-Duhovny D, Kalisman N (2019) Integrative structure modeling: overview and assessment. *Annu Rev Biochem* 88:113–135
143. Mouradov D, Craven A, Forwood JK et al (2006) Modelling the structure of latexin-carboxypeptidase a complex based on chemical cross-linking and molecular docking. *Protein Eng Des Sel* 19:9–16
144. Mouradov D, King G, Ross IL et al (2008) Protein structure determination using a combination of cross-linking, mass spectrometry, and molecular modeling. *Methods Mol Biol* 426:459–474
145. Forwood JK, Thakur AS, Guncar G et al (2007) Structural basis for recruitment of tandem hotdog domains in acyl-CoA thioesterase 7 and its role in inflammation. *Proc Natl Acad Sci U S A* 104(25):10382–10387
146. Schröder GF, Grubmüller H (2004) FRETsg: biomolecular structure model building from multiple FRET experiments. *Comput Phys Commun* 158:150–157
147. Alexander N, Al-Mestarihi A, Bortolus M et al (2008) De novo high-resolution protein structure determination from sparse spin-labeling EPR data. *Structure* 16:181–195
148. Schmitz C, Vernon R, Otting G et al (2012) Protein structure determination from pseudocontact shifts using ROSETTA. *J Mol Biol* 416:668–677
149. Zheng W, Doniach S (2002) Protein structure prediction constrained by solution X-ray scattering data and structural homology identification. *J Mol Biol* 316:173–187
150. Zheng W, Doniach S (2005) Fold recognition aided by constraints from small angle X-ray scattering data. *Protein Eng Des Sel* 18:209–219
151. De Rosier DJ, Klug A (1968) Reconstruction of three dimensional structures from electron micrographs. *Nature* 217:130–134
152. Nogales E, Scheres SHW (2015) Cryo-EM: a unique tool for the visualization of macromolecular complexity. *Mol Cell* 58:677–689
153. Short JM, Liu Y, Chen S et al (2016) High-resolution structure of the presynaptic RAD51 filament on single-stranded DNA by electron cryo-microscopy. *Nucleic Acids Res* 44:9017–9030
154. Ho C-M, Li X, Lai M et al (2020) Bottom-up structural proteomics: cryoEM of protein complexes enriched from the cellular milieu. *Nat Methods* 17:79–85
155. Hamuro Y, Burns LL, Canaves JM et al (2002) Domain organization of D-AKAP2 revealed by enhanced deuterium exchange-mass spectrometry (DXMS). *J Mol Biol* 321:704–714
156. Kamal JKA, Chance MR (2007) Modeling of protein binary complexes using structural mass spectrometry data. *Protein Sci* 17:79–94
157. Lees JG, Janes RW (2008) Combining sequence-based prediction methods and circular dichroism and infrared spectroscopic data to improve protein secondary structure determinations. *BMC Bioinformatics* 9:24
158. Russel D, Lasker K, Webb B et al (2012) Putting the pieces together: integrative

- modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* 10:e1001244
159. Alber F, Förster F, Korkin D et al (2008) Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev Biochem* 77:443–477
 160. Nguyen MN, Madhusudhan MS (2011) Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Res* 39:e94–e94
 161. Nguyen MN, Tan KP, Madhusudhan MS (2011) CLICK—topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Res* 39:W24–W28
 162. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2016) FoXS, FoXSDock and MultiFoXS: single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res* 44:W424–9
 163. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. *Comput Phys Commun* 91:43–56
 164. Pronk S, Páll S, Schulz R et al (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29:845–854

Part II

Protein Production



High-Level Production of Recombinant Eukaryotic Proteins from Mammalian Cells Using Lentivirus

Ester Behiels and Jonathan Elegheert

Abstract

Mammalian protein expression systems are ideally suited for the high-level production of recombinant eukaryotic secreted and membrane proteins for structural biology applications. Here, we present genetic transduction of HEK293-derived cells using lentivirus as a robust and cost-efficient method for the rapid generation of stable expression cell lines. We describe the features of the lentiviral transfer plasmid pHR-CMV-TetO₂, as well as detailed protocols for production of lentiviral particles, determination of functional lentiviral titer, infection of expression cells, culture and expansion of the resulting stable cell lines, their adaptation to adherent and suspension growth, and constitutive or inducible milligram-scale protein production. The typical lead-time for a full production run is ~3–4 weeks, with an anticipated yield of up to tens of milligrams of protein per liter of expression medium.

Key words Recombinant protein production, Membrane proteins, Lentivirus, HEK293 cells, Stable cell lines, Flow cytometry, Structural biology

1 Introduction

1.1 *Lentiviral Transduction of HEK293 Cells*

Large-scale production of eukaryotic secreted and membrane proteins for biochemical and structural studies crucially depends on the use of expression systems that contain the necessary cellular machinery for protein synthesis, folding and quality control, correct subcellular targeting, and for performing post-translational modifications such as glycosylation. Human embryonic kidney 293 (HEK293) cells and their engineered derivatives [1] have become the mammalian expression hosts of choice because of their robust growth, ease of culture in adherent and suspension formats, and consistent and high yields.

The most widely used and established methods for introducing the gene-of-interest (GOI) into the HEK293 expression host include (i) transient transfection, where non-integrating expression plasmids are introduced into the host cell in high copy numbers using DNA condensing agents [2, 3], (ii) stable transfection, where

after selection the genetic material is either long-term integrated into the cellular genome or maintained as an episomal plasmid [4], and (iii) baculovirus transduction of mammalian cells (BacMam) where a modified insect cell virus is used as a vehicle for transient delivery of the GOI [5, 6]. Disadvantages of transient transfection include the high consumable costs associated with large-scale plasmid preparation kits, with large volumes of expression media and with high numbers of plastic roller bottles and extended-surface culture flasks. Since stable integration of foreign DNA into the genome is a relatively rare event, the time frame for establishing and selecting a high-expressing monoclonal cell line using stable transfection is up to 8–10 weeks per construct, which is not well-suited to achieve a high sample throughput. Finally, in BacMam, large quantities of P1 and P2 virus need to be produced to be able to transiently infect large volumes of expression cells, in a laborious procedure that requires dedicated infrastructure.

In an effort to address these shortcomings, we recently implemented a lentivirus-based approach for the rapid generation of polyclonal stable HEK293 cell lines [7]. The recombinant lentivirus system exploits the transduction principles and genetic components of the human immunodeficiency virus-1 (HIV-1) [8] in a procedure where a specialized HEK293T-based producer cell line is transfected with a transfer, packaging and envelope plasmid mix to yield recombinant lentiviral particles (Fig. 1) that can stably transduce HEK293 expression cells with high efficiency to enable large-scale protein production (Fig. 2).

We constructed the transfer plasmid pHR-CMV-TetO₂ for optimal expression in HEK293 cell lines [7]; it is a second-generation design where the GOI is under control of a major immediate-early human cytomegalovirus enhancer/promoter (CMV-MIE), and is flanked by the HIV-1 5' and 3' long-terminal repeats (LTRs) for viral packaging and integration into the host genome (Fig. 1a). Other components include the minimally necessary psi packaging signal (ψ ; regulates the packaging of the lentiviral RNA genome into the viral capsid), Rev. response element (RRE; for nuclear export of unspliced and partially spliced viral RNA transcripts), polypurine tract (PPT; necessary for priming plus-strand DNA synthesis) as well as the woodchuck hepatitis virus (WHV) post-transcriptional regulatory element (WPRE; for improved transcription termination, transcript stability, and transgene expression) [9] (Fig. 1a).

With the pHR-CMV-TetO₂ transfer plasmid, a second-generation packaging system consisting of two helper plasmids is used; (i) an envelope plasmid that encodes the vesicular stomatitis virus G envelope protein (VSV-G) to yield a pseudotyped lentiviral particle with high infectivity and broad host range, and (ii) a packaging plasmid that contains the HIV genes that encode proteins that are crucial for virus production: Gag (structural precursor

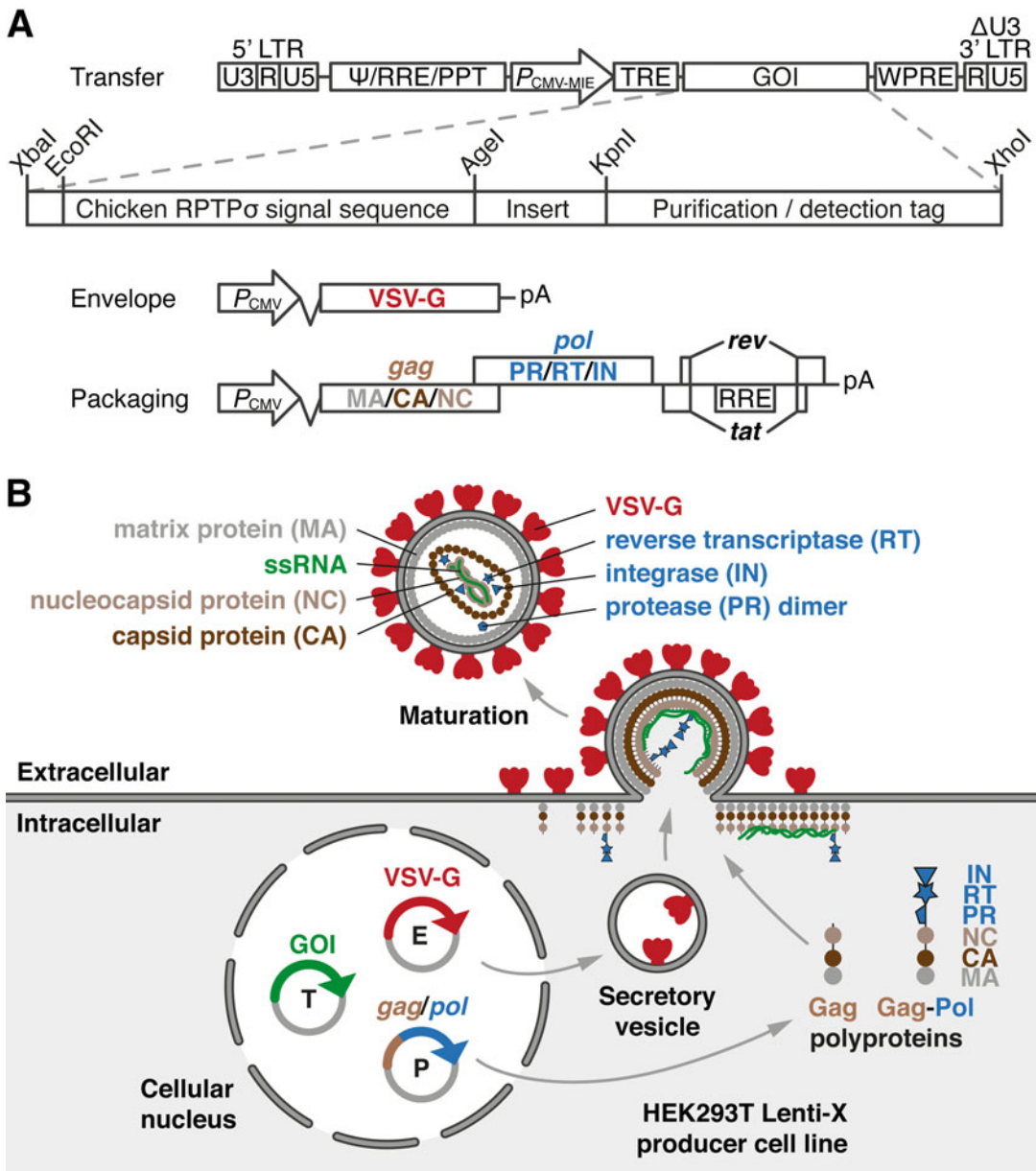


Fig. 1 Lentiviral plasmids and production of recombinant lentiviral particles. (a) Genetic elements of the lentiviral transfer, envelope and packaging plasmids, and layout of the pHR-CMV-TetO₂ multiple cloning site (MCS). *LTR* long-terminal repeat (U3, R and U5 region), ψ psi packaging signal, *RRE* Rev. response element, *PPT* polypurine tract, $P_{CMV-MIE}$ major immediate-early human cytomegalovirus enhancer/promoter, *TRE* tetracycline response element, *GOI* gene-of-interest, *WPRE* woodchuck hepatitis virus (WHV) post-transcriptional regulatory element, *RPTP σ* receptor protein tyrosine phosphatase sigma, *VSV-G* vesicular stomatitis virus G protein, *pA* polyadenylation signal, *gag* group-specific antigen, *pol* polymerase. (b) The HEK293T Lenti-X producer cell line is transfected with a packaging (P), envelope (E), and transfer (T) plasmid mix, leading to expression of viral enzymes, structural proteins, and accessory proteins, and to production of

protein) and Gag-Pol (polymerase) polyproteins, Tat (viral transactivator for activation of transcription from the 5' LTR) and Rev. (facilitates nuclear export of viral RNA transcripts) (Fig. 1a, b).

Commonly used HEK293 cell lines for large-scale protein production include cells expressing the SV40 large-T antigen (HEK293T) [10], the suspension-adapted N-acetylglucosaminyl-transferase I-negative (GnTI[−]) HEK293 (HEK293S GnTI[−]) cells that produce homogeneous high mannose-type (Man₅GlcNAc₂) N-linked glycans that are sensitive to cleavage by endoglycosidase H (EndoH) or F1 (EndoF1) [11], and the HEK293S GnTI[−] TetR cells that additionally express Tet repressor protein (TetR) for inducible expression [12] (Fig. 2). Viral entry into the HEK293 expression cell line is triggered by binding of viral VSV-G to the cell-surface low-density lipoprotein receptor (LDL-R) [13], is then followed by release and reverse transcription of the viral single-stranded RNA (ssRNA), and finally by stable integration of the proviral DNA into the host cell genome (Fig. 2a). Using the pHR-CMV-TetO₂ transfer plasmid, inducible expression is enabled by the presence of two *TetO* operator sequences (forming the tetracycline response element or TRE) downstream of the CMV-MIE enhancer/promoter, which bind the Tet repressor protein (TetR) that is constitutively expressed in HEK293S GnTI[−] TetR cells. Transcription of the GOI is induced by application of the antibiotic doxycycline (Dox) that binds TetR to release it from the *TetO* operator sequences [14] (Fig. 2b).

1.2 Brief Overview of the Procedure

In a first step, a lentivirus producer cell line is transiently co-transfected with the transfer, envelope, and packaging plasmids to produce lentiviral particles that are secreted into the cell culture medium (*see* Subheading 3.4). If desired, after 72 h, the resulting lentiviral titer can be determined using either flow cytometry (*see* Subheading 3.5), or endpoint dilution and fluorescence microscopy (*see* Subheading 3.6). The resulting lentiviral stock solution is then used to infect the expression host cell line of choice (*see* Subheading 3.4). This procedure leads to the rapid (within ~7 days) establishment of polyclonal cell lines that can be selected, expanded, and adapted to adherent (*see* Subheading 3.7) or suspension (*see* Subheading 3.8) protein expression setups.

Fig. 1 (continued) recombinant secreted and pseudotyped (i.e., containing non-native envelope protein) lentiviral particles. VSV-G coats the viral membrane. Structural proteins encoded by *gag* include the matrix protein (MA; coats the inner surface of the viral membrane), capsid protein (CA; coats the viral single-stranded RNA (ssRNA)), and nucleocapsid protein (NC; forms a complex with the viral ssRNA). The *pol* gene encodes the enzymes reverse transcriptase (RT), integrase (IN) and protease (PR). Gag and Gag-Pol polyproteins are cleaved upon maturation of the virion.

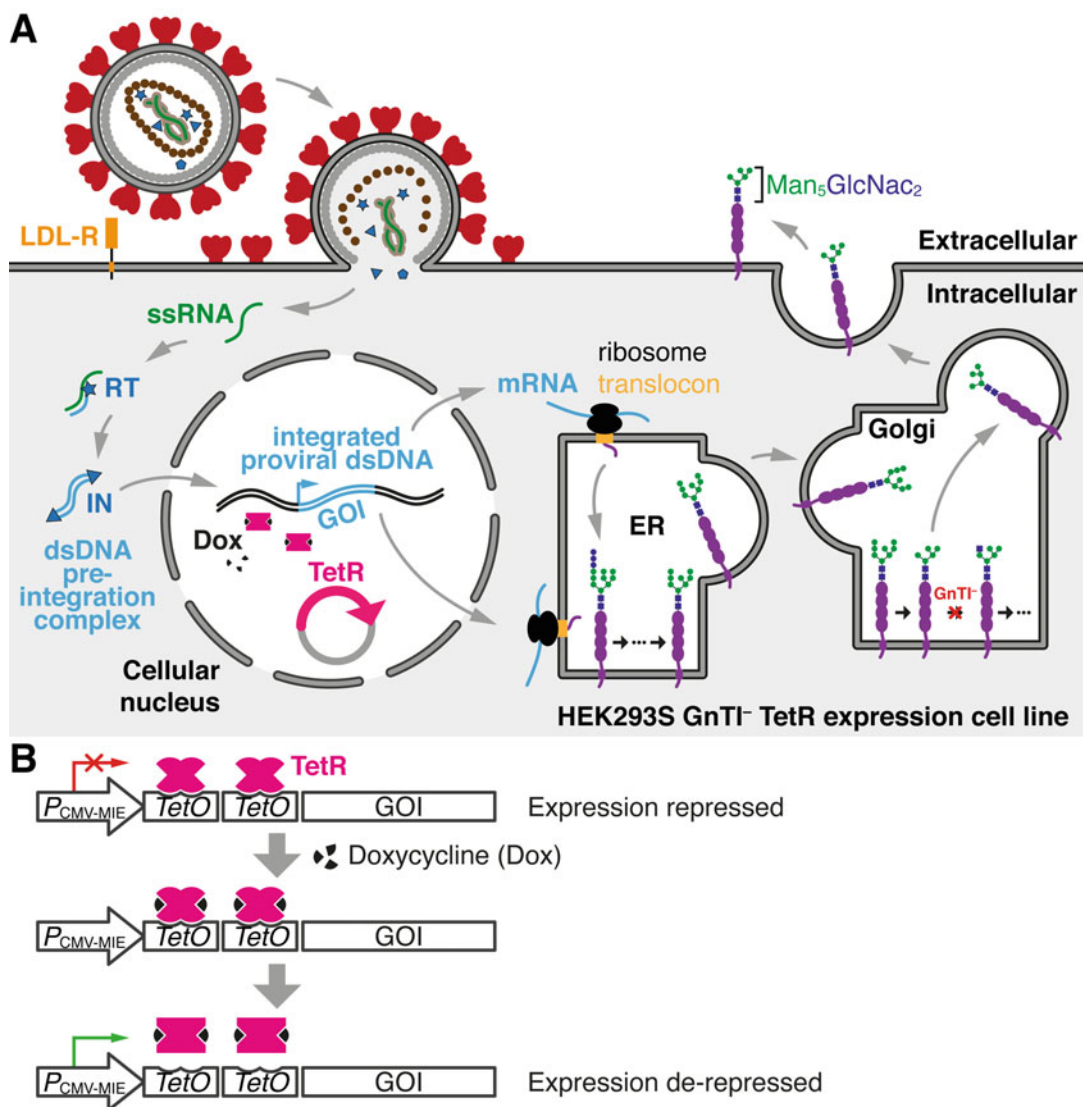


Fig. 2 Integration of proviral DNA and inducible protein expression. **(a)** The lentiviral particle attaches to the host cell membrane through the interaction of the VSV-G glycoprotein with the low-density lipoprotein receptor (LDL-R). After membrane fusion, the viral ssRNA genome is uncoated and released into the cytoplasm. The ssRNA is reverse-transcribed into double-stranded DNA (dsDNA) by the viral reverse transcriptase (RT). A dsDNA-integrase (IN) pre-integration complex is translocated to the cell nucleus where the dsDNA is stably integrated into the host cell genome. HEK293S GnT1⁻ TetR cells stably express Tet repressor protein (TetR) that blocks transcription of the GOI. Binding of Doxycycline (Dox) to TetR de-represses GOI transcription. In co-translational translocation, the ribosome and translocon associate to direct nascent polypeptides containing a signal sequence (SS) into the endoplasmic reticulum (ER) lumen. The biosynthesis of N-linked glycans occurs throughout the ER and the Golgi apparatus. HEK293S GnT1⁻ TetR cells are deficient in the enzyme N-acetylglucosaminyltransferase I (GnT1) and produce homogeneous high mannose-type (Man₅GlcNAc₂) N-linked glycans that are sensitive to cleavage by endoglycosidase. **(b)** Inducible expression using the pHR-CMV-TetO₂ plasmid and HEK293S GnT1⁻ TetR cell line. The GOI is flanked upstream by a CMV-MIE promoter/enhancer and two TetO sequences. Expression of the GOI is repressed by the high affinity binding of TetR homodimers to the TetO sequences in the absence of Dox. Binding of Dox to TetR is followed by un-binding of TetR and de-repression of transcription.

1.3 Biological Safety

Work involving HIV-1-based lentivirus should be carried out in a biosafety level 2 (BSL2) or 2+ (BSL2+) facility (depending on the relevant institutional and governmental biosafety guidelines). The major risks associated with a lentivirus-based expression system are (i) the capability for generation of replication-competent lentivirus (RCL) by recombination events and (ii) the potential for oncogenesis or other deleterious effects upon insertion of the provirus carrying the transgene, either directly, through the intrinsic nature of the transgene, or indirectly, through insertional mutagenesis.

The risk of RCL formation is reduced by the specific design of the second-generation pHR-CMV-TetO₂ vector system; it separates transfer, envelope, and packaging components of the lentivirus over three separate plasmids. The transfer plasmid cannot produce functional viral particles without the genes encoded in the envelope and packaging vectors. Hence, it is not possible for viruses produced from this system to replicate after the initial infection of the expression cell line, unless recombination would occur between the three plasmids and the resulting construct would be packaged into a viral particle. It lacks the accessory virulence factors *vif*, *vpr*, *vpu*, and *nef*, which normally provide a fitness advantage but are non-essential *in vitro* [15]. Finally, the transfer plasmid is self-inactivating (SIN) [16] since it carries a deletion in the enhancer/promoter region ‘U3’ of the 3’ LTR (Δ U3) (while the 5’ LTR U3 is left intact). Upon transduction, this deletion is copied into the 5’ LTR during reverse transcription, resulting in transcriptional inactivation of the provirus in the infected cell. All these modifications significantly reduce the risk and moreover, a recent comprehensive analysis of lentiviral clinical vector sets as well as monitoring of patients throughout their entire gene therapy treatment cycle have highlighted the unlikelihood of RCL development [17].

The risk for insertional oncogenesis by lentiviral vectors relates to their potential of dysregulating cellular proto-oncogenes after random genomic integration. However, compared to γ -retroviral vectors, this genotoxic risk is strongly reduced because of the specific integration pattern of lentiviral vectors and because of the SIN LTR design [18]. Additionally, the potential for oncogenesis is largely based on the nature of the GOI contained within the lentiviral transfer plasmid and should be considered on a case-by-case basis via carrying out a risk assessment (*a.o.* for toxic genes, oncogenes and genes involved in cell growth, cell death, or apoptosis).

We recommend strict adherence to a number of risk reduction measures to work safely with recombinant lentivirus (*see* Subheading 3.1).

2 Materials

2.1 Mammalian Cell Lines and Bacterial Cloning Strains

1. HEK293T Lenti-X cells (Takara Bio, #632180).
2. HEK293T cells (ATCC, #CRL-3216).
3. HEK293S GnTI[−] cells (ATCC, #CRL-3022).
4. HEK293S GnTI[−] TetR cells (available by request from N. Callewaert group, VIB-UGent Center for Medical Biotechnology; nico.callewaert@ugent.vib.be).
5. Stbl3 bacterial cloning strain (Thermo Fisher Scientific, #C737303).
6. NEB stable bacterial cloning strain (New England Biolabs, #C3040).

2.2 Lentiviral Plasmids

1. Lentiviral transfer plasmid pHR-CMV-TetO₂ and variants (Addgene, #113883-#113901, *see Note 1*).
2. Lentiviral packaging plasmid psPAX2 (Addgene, #12260).
3. Lentiviral envelope plasmid pMD2.G (Addgene, #12259).

2.3 Tissue Culture Reagents

1. Complete DMEM/F-12/10% FBS medium: 445 mL DMEM/F-12 (with high glucose, L-Gln, phenol red, sodium pyruvate), 50 mL fetal bovine serum (FBS; 10% vol/vol), 5 mL MEM non-essential amino acids solution (NEAA; 1% vol/vol). Store at 4 °C.
2. Low-serum DMEM/F-12/2% FBS medium: 485 mL DMEM/F-12, 10 mL FBS (2% vol/vol), 5 mL NEAA (1% vol/vol). Store at 4 °C.
3. Serum-free DMEM/F-12 medium: 495 mL DMEM/F-12, 5 mL NEAA (1% vol/vol). Store at 4 °C.
4. Low-serum FreeStyle 293/1% FBS medium: 490 mL FreeStyle 293, 5 mL FBS (1% vol/vol), 5 mL NEAA (1% vol/vol). Store at 4 °C.
5. 1 mg/mL PEI (25 kDa, branched, M_n ~ 10,000) stock: weigh 500 mg PEI liquid in a 50 mL tube (branched PEI is too viscous to pipette). Add 30 mL of warm (~60 °C) ultrapure water and rotate until dissolved. Top up to 50 mL and mix well. Dilute 10× to 1 mg/mL in ultrapure water and adjust the pH to 7.0 with HCl. Filter-sterilize through a 0.2-μm syringe filter inside a biological safety cabinet. Aliquot and store at −20 °C.
6. Trypsin-EDTA (0.05% wt/vol), with and without phenol red.
7. 1 mg/mL soybean trypsin inhibitor (SBTI) stock: dissolve 10 mg of SBTI in PBS to a final volume of 10 mL. Filter-sterilize using a 0.2-μm syringe filter unit inside a biological safety cabinet. Aliquot and store at −20 °C.

8. 10 mg/mL polybrene infection reagent.
9. 10 mg/mL doxycycline hydrochloride (Dox): dissolve 50 mg of Dox in 100% (vol/vol) ethanol to a final volume of 5 mL. Filter-sterilize using a 0.2- μ m syringe filter unit inside a biological safety cabinet. Aliquot and store at -20°C .
10. 2 mg/mL blasticidin hydrochloride (1000 \times stock): dissolve 20 mg of blasticidin HCl in ultrapure water to a final volume of 10 mL. Filter-sterilize using a 0.2- μ m syringe filter unit. Aliquot and store at -20°C .
11. 500 mM sodium butyrate: dissolve 1.1 g of sodium butyrate in ultrapure water to a final volume of 20 mL. Filter-sterilize using a 0.2- μ m syringe filter unit inside a biological safety cabinet. Aliquot and store at -20°C .
12. 500 mM valproic acid (VPA): dissolve 1.66 g of VPA in ultrapure water to a final volume of 20 mL. Filter-sterilize using a 0.2- μ m syringe filter unit inside a biological safety cabinet. Aliquot and store at -20°C .
13. 500 μ M kifunensine: dissolve 11.6 mg of kifunensine in ultrapure water to a final volume of 100 mL. Filter-sterilize using a 0.2- μ m syringe filter unit inside a biological safety cabinet. Aliquot and store at -20°C .
14. Rely+On Virkon disinfectant powder.

2.4 Equipment

1. Class II microbiological safety cabinet.
2. Disposable long-cuff gloves.
3. Stationary tissue culture incubator with CO_2 , temperature, and humidity control.
4. Shaking mammalian cell culture incubator with CO_2 , temperature, and humidity control.
5. Roller bottle apparatus and roll-in incubator with CO_2 , temperature, and humidity control.
6. Ribbed-surface polystyrene roller bottles (2125 cm^2) with filter cap.
7. Black 96-well cell culture microplate with clear bottom.
8. Sterile 30-mL Luer-lock plastic syringes.
9. Sterile 0.45- μ m polyethersulfone (PES) syringe filter units.
10. 500-mL and 2-L baffled polycarbonate Erlenmeyer flasks and corresponding filter caps.
11. Automated cell counter or hemocytometer.
12. Flow cytometer or cell sorter.
13. Wide-field fluorescence microscope.

3 Methods

3.1 Safety Guidelines

In our laboratory, the following measures are taken to reduce the risks (also *see* Table 1):

1. Class II viral work is carried out by specially trained staff in a dedicated, restricted-access room with dedicated equipment, wearing protective equipment at all times (laboratory coat, safety eyewear, and disposable long-cuff gloves).
2. All manipulations involving lentivirus are carried out in a Class II microbiological safety cabinet (MSC), and working areas are decontaminated with 1% (wt/vol) Virkon, both before and after.
3. No sharps or glassware are used, and care is taken to avoid aerosol generation (e.g., upon centrifugation of virus-containing supernatant).
4. All accidental spillage has to be reported to the responsible safety officer. Affected surfaces and items are thoroughly decontaminated using 1% (wt/vol) Virkon.
5. All waste has to be disposed of as biohazardous waste, according to the specific institutional and governmental guidelines. In our laboratory, all materials exposed to viral supernatant (including tips, pipettes, etc.) are decontaminated with 1% (wt/vol) Virkon and subsequently disposed of in autoclave bags, which are put in dedicated biological waste containers to be autoclaved.

3.2 General Cell Culture

All mammalian cell culture is performed in a Class II microbiological safety cabinet and because the protocols described here do not use penicillin-streptomycin (“Pen-Strep”) to prevent bacterial growth, it is important to implement a strict sterile technique in a well-maintained tissue culture infrastructure. HEK293

Table 1
Exposure risks and precautions

Risk	PPE and precautions
Direct contact via skin	Long-cuff gloves, lab coat, long trousers, and closed shoes
Injection	No sharps, no glassware
Exposure to mucous membranes (eyes, nose, mouth)	Safety goggles and face mask
Aerosols	No centrifugation

All work is carried out in a Class II microbiological safety cabinet (MSC)

All working areas are decontaminated with 1% (wt/vol) Virkon, both before and after

Table 2
HEK293 cell growth conditions

Growth	Cells	Medium	Incubator
Adherent	HEK293T	DMEM/F-12/10%	Humidified CO ₂ incubator at 37 °C
	HEK293S GnTI [−]	FBS	and 6% CO ₂
	HEK293S GnTI [−]	(for expansion).	Stationary type for vented flasks
	TetR	DMEM/F-12/2% FBS	Roll-in type with roller bottle apparatus
	HEK293T Lenti-X	(for expression).	for vented roller bottles
Suspension	HEK293S GnTI [−]	FreeStyle 293/1%FBS	Shaking CO ₂ incubator at 37 °C,
	HEK293S GnTI [−]	medium	8% CO ₂ , and 130 rpm for baffled
	TetR		Erlenmeyer flasks with filter cap

(HEK293T, HEK293S GnTI[−], HEK293S GnTI[−] TetR, or HEK293T Lenti-X) cells should be split regularly and should not be used beyond passage P20 (P10 for HEK293S GnTI[−] TetR; *see Note 2*) to ensure maximum cell viability and maximum protein or virus yield. Growth conditions are summarized in Table 2. It should be noted that the required CO₂ level depends on the type of medium (*see Note 3*). In case of the HEK293S GnTI[−] TetR cell line, blasticidin (2μg/mL; to retain selective pressure on the pcDNA6/TR genetic elements that direct TetR expression) should be added to the growth medium, while both blasticidin (2μg/mL) and doxycycline (0.1–10μg/mL; to induce transcription of the GOI) should be added to the expression medium.

Adherent cells are split at 90–95% confluency (*see Note 4*). We recommend splitting them 1/6 to 1/10 (*see Note 5*) twice a week at fixed times, e.g., Monday morning and Thursday afternoon, for a maximum of ~10 weeks (20 passages). Suspension cells are usually split at a density of ~2.0 × 10⁶ cells/mL by diluting with FreeStyle 293/1% FBS medium (*see Note 6*) to a final cell density of ~0.5 × 10⁶ cells/mL. The exact moment of splitting has to be determined by monitoring the cell density on a daily basis by performing cell counts; in general, this is after 2–3 days. All PBS, FBS, trypsin-EDTA, and cell culture media should be pre-warmed (37 °C) for these procedures.

Procedure for splitting and expanding adherent HEK293 cells, grown to 90–95% confluency, in a T75 or T175 flask (as used in Subheadings 3.7 and 3.8):

1. Remove the DMEM/F-12/10% FBS medium and gently wash the cells with 5 mL (T75) or 10 mL (T175) PBS (*see Note 7*).
2. Dissociate the cells by incubating them with 2 mL (T75) or 5 mL (T175) trypsin-EDTA for 3 min in a humidified incubator at 37 °C and 6% CO₂, followed by gently tapping the flask. Quench the trypsin by adding 10 mL (T75) or 25 mL (T175) complete DMEM/F-12/10% FBS medium and break up any

clumped cells by pipetting up and down with a sterile serological 10 mL pipette.

3. Transfer the required number of cells (1/10th to 1/6th of the total cell slurry, *see* **Note 5**) to a new T75 or T175 flask and top up the DMEM/F-12/10% FBS medium to the recommended volume, i.e., 12 mL (T75) or 30 mL (T175). Place the flasks in the incubator.

3.3 Plasmid Preparation

1. “Design and cloning”—The multiple cloning site (MCS) of the transfer plasmid pHR-CMV-TetO₂ has a modular design which is fully compatible with that of the popular pHLsec plasmid for transient transfection [2], hence allowing for the easy transfer of inserts and tags. The layout of the MCS is as follows: *EcoRI*–chicken RPTP σ signal sequence–*AgeI*–target gene–*KpnI*–tag and stop codons–*XhoI*. The GOI should be inserted between the *AgeI* and *KpnI* restriction sites using general cloning procedures, in frame with the RPTP σ signal sequence (*see* **Note 8**), and in frame with the desired tag, which can be conveniently inserted between the *KpnI* and *XhoI* sites (Fig. 1a).
2. “Production”—10 μ g of each of the three lentiviral plasmids (pHR-CMV-TetO₂; psPAX2; pMD2.G) is required for transfection of one T75 flask of HEK293 Lenti-X producer cells, which will then serve for transduction of one T75 flask of expression cells. One to two minipreps should give enough yield when replicating the plasmid DNA in a suitable strain like Stbl3 or NEB stable (*see* **Note 9**) and using a commercial endotoxin-free plasmid DNA miniprep kit for plasmid isolation (including the necessary wash steps for endotoxin removal according to the manufacturer’s recommendations).

3.4 Lentivirus Production and Transduction of Expression Cells

3.4.1 Production of Lentiviral Particles

1. *Day 1*. Seed 9×10^6 HEK293T Lenti-X cells (*see* **Note 10**) in 12 mL of DMEM/F-12/10% FBS medium in a T75 flask to achieve ~50% confluency. Place the flask in the incubator.
2. *Day 2*. Prepare the plasmid DNA transfection mix in a sterile 1.5-mL tube: 10 μ g pHR-CMV-TetO₂ transfer plasmid, 10 μ g psPAX2 packaging plasmid, 10 μ g pMD2.G envelope plasmid (*see* **Note 10**). Add serum-free DMEM/F-12 medium to a final volume of 0.25 mL. Mix gently by pipetting.
3. Prepare 75 μ L of PEI (1 mg/mL; 1:2.5 (wt/wt) DNA:PEI ratio) in a sterile 1.5-mL tube. Top up with serum-free DMEM/F-12 medium to a total volume of 0.25 mL. Mix gently by pipetting.
4. Combine the 0.25 mL plasmid DNA transfection mix and 0.25 mL PEI mix in a 1.5-mL tube.

5. Vortex the 1.5-mL tube gently for 10 s, then briefly centrifuge the tube at low speed (100 g, 22–24 °C, 30 s) to collect the liquid at the bottom of the tube.
6. Incubate the DNA:PEI mix in the flow cabinet for 15–20 min.
7. Remove and discard the DMEM/F-12/10% FBS medium from the HEK293T Lenti-X T75 flask, which is now >90% confluent (continued from **step 1**, Subheading 3.4.1). Wash with 10 mL of PBS and add 11.5 mL of fresh DMEM/F-12/2% FBS medium.
8. Add the DNA:PEI mix and gently tilt the T75 flask to cover all cells and place back in the incubator.

3.4.2 Transduction of Expression Cells

1. *Day 4.* Seed 9×10^6 expression cells (e.g., HEK293T, HEK293S GnTI[−] or HEK293S GnTI[−] TetR) in 12 mL of DMEM/F-12/10% FBS medium in a T75 flask to achieve ~50% confluency. Place the flask in the incubator.
2. *Day 5.* 3 days after transfection, collect the 12 mL of lentivirus-containing conditioned medium from the HEK293T Lenti-X T75 flask (from **step 8**, Subheading 3.4.1) into a sterile 50-mL tube.
3. Add 6 mL of fresh DMEM/F-12/10% FBS medium to the 50-mL tube.
4. Filter the resulting 18 mL conditioned medium through a 0.45- μ m filter unit attached to a Luer-lock syringe and into a new sterile 50-mL tube.
5. Add 18 μ L of polybrene (from a 10-mg/mL 1000 \times stock solution) to the 18 mL conditioned medium and mix gently (see **Note 11**).
6. Take the 90–95% confluent T75 flask with HEK293 expression cells that was seeded at ~50% confluency the day before (from **step 1**, Subheading 3.4.2). Remove the DMEM/F-12/10% FBS medium and gently wash the cells with 10 mL PBS.
7. Add the 18 mL of filtered, lentivirus- and polybrene-containing medium.
8. Place the flask in the incubator for 72 h, during which the lentiviral particles will infect the cells and stably integrate their genetic material into the host cell genome.

At the point of harvesting the lentivirus-containing supernatant (Subheading 3.4.2, **step 5**), the functional lentiviral titer (the number of “functional” infectious particles present per volume unit of conditioned medium), can be determined. Determination of the lentiviral titer and of the multiplicity of infection (MOI), which is defined as the ratio of the number of applied viral particles to the number of target cells at the time of infection, is not mandatory but

allows better control of the transduction process (*see* **Note 12**). To enable titer determination using either flow cytometry (Subheading 3.5) or fluorescent microscopy (Subheading 3.6), the GOI must be cloned into the appropriate pHR-CMV-TetO₂ transfer plasmid that either directs co-expression of a fluorescent protein marker from a bicistronic transcript, or as a direct fusion with the target protein [7] (*see* **Note 1**). These vector variants can also be used for enrichment of the polyclonal stable cell line or for the isolation of single clones, using fluorescence-activated cell sorting (FACS) [7].

For expansion into adherent cell culture, which is typically used for secreted proteins, follow the protocol in Subheading 3.7. For expansion into suspension cell culture, follow the protocol in Subheading 3.8; this strategy is used for cell-surface-bound proteins and membrane proteins but is equally suitable for secreted and intracellular proteins.

3.5 Determination of Functional Lentiviral Titer by Flow Cytometry

The following titration procedure is for one flat-bottom 12-well cell culture plate. Each of the first ten positions (wells no. 1–10) is used to titer one lentiviral particle-containing supernatant dilution, for a total of ten dilutions (10^0 to 10^9). Well no. 11 will be used to obtain a negative control sample for flow cytometry analysis, and well no. 12 will be used to perform a cell count at the time of infection.

1. *Day 4*. Seed each well of the 12-well plate with $\sim 4.5 \times 10^5$ HEK293T cells ($\sim 1.25 \times 10^5$ cells/cm²) in DMEM/F-12/10% FBS medium, bringing them to $\sim 50\%$ confluency.
2. Grow the cells overnight at 37 °C in a humidified incubator operated at 6% CO₂.
3. *Day 5* (continued from Subheading 3.4.2, step 5). Make ten-fold dilution stocks of lentivirus from 10^0 to 10^9 (ten dilutions in total) in DMEM/F-12/2% FBS medium in sterile 2-mL tubes to a final volume of 1.5 mL.
4. For well no. 12 of the 12-well plate, trypsinize the cells by removing the DMEM/F-12/10% FBS medium, washing the cells with PBS, and adding 250 μ L trypsin-EDTA (0.05% wt/vol). Incubate for 3 min in a humidified incubator operated at 37 °C with 6% CO₂. Add 750 μ L DMEM/F-12/10% FBS to inactivate the trypsin. Perform a cell count to determine the total number of cells in one well, using an automated cell counter or alternatively manually using a hemocytometer.
5. For wells no. 1–10 of the 12-well plate, remove the DMEM/F-12/10% FBS medium, wash the cells gently with PBS and add 1 mL of every virus-containing dilution (10^0 to 10^9) to the corresponding well position.

6. For well no. 11 of the 12-well plate, remove the DMEM/F-12/10% FBS medium, wash the cells gently with PBS and replace with 1 mL of fresh DMEM/F-12/2% FBS medium.
7. Incubate the plate for 72 h at 37 °C in a humidified incubator operated at 6% CO₂.
8. *Day 8.* After 72 h, trypsinize the cells of wells no. 1-11 by removing the DMEM/F-12/2% FBS medium, washing the cells three times with PBS, and adding 250 µL trypsin-EDTA (0.05% wt/vol without phenol red) to every well. Incubate for 3 min in a humidified incubator operated at 37 °C with 6% CO₂. Add 250 µL SBTI (1-mg/mL stock) to every well to inactivate the trypsin (*see Note 13*).
9. Using flow cytometry, determine the percentage of fluorescence-positive cells for all 10 infected samples after establishing an appropriate gating strategy to remove debris and dead cells, with the cells from well no. 11 serving as negative control.
10. To calculate titer of the lentiviral stock solution, use the following equation for those dilutions where 1-10% of fluorescence-positive cells are observed:

Transduction Units (TU)/mL =

$$\frac{\text{fraction of positive cells} \times \text{cell count at infection} \times \text{dilution factor}}{\text{volume of dilution stock solution (mL)}}$$

For example, if 1 mL of virus-containing solution was added to 8×10^5 HEK293T cells, and 5% of fluorescence-positive cells are observed in well no. 3 (dilution 10^2), it follows that;

$$\begin{aligned} \text{Transduction Units (TU)/mL} &= \frac{0.05 \times 800,000 \times 100}{1 \text{ mL}} \\ &= 4 \times 10^6 \end{aligned}$$

3.6 Determination of Functional Lentiviral Titer by Endpoint Dilution and Fluorescence Microscopy

The following titration procedure is for one black 96-well cell culture microplate with clear bottom. Columns 1–3, 4–6, 7–9, and 10–12 can each be used to titer one lentivirus preparation in triplicate; the eight dilutions (10^0 to 10^7) fit into rows A to H.

1. *Day 4.* Seed each well of the plate with $\sim 4 \times 10^4$ HEK293T cells ($\sim 1.25 \times 10^5$ cells/cm²) in DMEM/F-12/10% FBS medium, bringing them to $\sim 50\%$ confluency.
2. Grow the cells overnight at 37 °C in a humidified incubator operated at 6% CO₂.
3. *Day 5* (continued from Subheading 3.4.2, step 5). Make ten-fold dilution stocks of lentivirus from 10^0 to 10^7 (eight

dilutions in total) in DMEM/F-12/2% FBS medium in sterile 1.5-mL tubes to a final volume of 1 mL.

4. Remove the DMEM/F-12/10% FBS medium from the 96-well plate, wash the cells with PBS and add 100 μ L of every virus-containing dilution (10^0 to 10^7) to the corresponding well position. Infect the HEK293T cells in triplicate for each dilution.
5. Incubate the plate for 72 h at 37 °C in a humidified incubator operated at 6% CO₂.
6. *Day 8*. After 72 h, count the number of fluorescent cells in the dilution that contains less than 10 fluorescent cells.
7. To calculate lentiviral titer, use the following equation:

$$\text{Transduction Units(TU)/mL} = \text{averaged number of fluorescent cells} \times \text{dilution factor} \times 10$$

For example, if an average number of 5 fluorescent cells is observed in row F (dilution 10^5), it follows that;

$$\text{Transduction Units(TU)/mL} = 5 \times 100,000 \times 10 = 5 \times 10^6$$

3.7 Protein Expression in Adherent Polyclonal Stable Cell Lines

This subheading describes soluble secreted protein expression in HEK293T or HEK293S GnTI[−] cells grown in six polycarbonate roller bottles (1.5 L expression medium in total). The number of roller bottles can be scaled according to the expression level and desired yield of the target protein. Timelines are only indicative (*see Note 5*).

1. *Day 8* (continued from Subheading 3.4.2, step 8). Wash the T75 flask containing the polyclonal stable cell line three times with 5 mL PBS and split it into two new T175 flasks. Incubate for ~3 days (until ~95% confluency).
2. *Day ~11*. Split each of the two T175 flasks into three new T175 flasks. Incubate for ~2 d (until ~95% confluency).
3. *Day ~13*. Split each of the six T175 flasks into one new ribbed-surface polystyrene roller bottle (2125 cm²) with filter cap and top up with DMEM/F-12/10% FBS medium to a final volume of 250 mL.
4. Place the roller bottles into a dedicated roller bottle apparatus and roll-in incubator (e.g., Wheaton R2P or Schuett-Biotec Incudrive-90) at 37 °C and 6% CO₂ (*see Note 14*). Rotate the bottles at 0.8–1.0 rpm and visually monitor attachment of the cells to the plastic surface on a daily basis.
5. *Day ~19*. The roller bottles should be 90–95% confluent. Remove the DMEM/F-12/10% FBS medium and replace with 250 mL of DMEM/F-12/2% FBS expression medium.

Optionally at this point, various chemicals can be added to influence protein expression or to manipulate protein post-translation modification, most notable N-linked glycosylation in HEK293T cells (*see Note 15*). Place the roller bottles back into a dedicated roller bottle apparatus at the desired expression temperature (37 or 30 °C; *see Note 16*).

6. Monitor cell viability throughout the experiment to decide on the appropriate time for collection of the conditioned medium. Typically, this is 4–5 days for HEK293T cells and 7–10 days for HEK293S GnTI[−] cells.

3.8 Protein Expression in Suspension-adapted Polyclonal Stable Cell Lines

This subheading describes membrane protein expression in suspension-adapted HEK293S GnTI[−] TetR cells grown in two 2-L polycarbonate Erlenmeyer flasks (1.6 L expression medium in total). The number of flasks can be scaled according to the expression level and desired yield of the target protein. Timelines are only indicative (*see Note 5*), and when diluting cell suspensions, keep in mind that the final culture volume should not exceed 40% of the vessel's total volume.

1. *Day 8* (continued from Subheading 3.4.2, step 8). Wash the T75 flask containing the polyclonal stable cell line three times with 5 mL PBS and split it into two new T175 flasks. Incubate for ~3 days (until ~95% confluency).
2. *Day ~11*. Remove the DMEM/F-12/10% FBS medium from both flasks and gently wash the cells with 10 mL PBS. Add 30 mL low-serum FreeStyle 293/1% FBS medium and detach the cells by gently pipetting the medium against the flask surface, using a sterile serological 10 mL pipette (*see Note 17*). Break up any cell clumps by gently pipetting up and down.
3. Pool the cells from the two T175 flasks and dispense them into a single 500-mL baffled polycarbonate Erlenmeyer flask with filter cap (*see Note 18*).
4. Perform a cell count of this cell suspension (anticipated cell density is $\sim 1.5 \times 10^6$ cells/mL).
5. Dilute with low-serum FreeStyle 293/1% FBS medium to a final cell density of $\sim 0.5 \times 10^6$ cells/mL (this usually corresponds to ~200 mL). Add blasticidin to a final concentration of 2 µg/mL. Grow the cells in a shaking incubator operated at 37 °C with 8% CO₂ and a shaking speed of 130 rpm, until they reach a density of $\sim 2.0 \times 10^6$ cells/mL (*see Note 19*). Monitor cell density at daily intervals by performing a cell count.
6. *Day ~14*. Transfer the cells into a single 2-L baffled polycarbonate Erlenmeyer flask with filter cap and again dilute the suspension with low-serum FreeStyle 293/1% FBS medium to a final cell density of $\sim 0.5 \times 10^6$ cells/mL (~800 mL final volume). Maintain blasticidin at a final concentration of 2 µg/

mL. Grow the cells until they reach a density of $\sim 2.0 \times 10^6$ cells/mL. Monitor cell density at daily intervals.

7. *Day ~17.* Distribute the ~ 800 mL of cells into two 2-L baffled polycarbonate Erlenmeyer flasks and dilute the suspension with low-serum FreeStyle 293/1% FBS medium to a final cell density of $\sim 1.0 \times 10^6$ cells/mL (~ 800 mL final volume). Maintain blasticidin at a final concentration of $2 \mu\text{g/mL}$. Grow the cells until they reach a density of $\sim 3.0 \times 10^6$ cells/mL. Monitor cell density at daily intervals (*see* **Note 20**).
8. Add Dox to induce protein expression, at a final concentration of $0.1\text{--}10 \mu\text{g/mL}$. Place the flasks back in the shaking incubator at the desired expression temperature (37 or 30°C , *see* **Note 16**).
9. *Day ~19–20.* At the desired time of collection ($24\text{--}72$ h post-induction; *see* **Note 16**), centrifuge the cell suspension for 10 min at 1500 g at 4°C and discard the supernatant. Snap-freeze the pellets in liquid nitrogen (LN_2) and store them at -80°C indefinitely for future use.

Similarly to expression in adherent format (Subheading 3.7, **step 5**), the iHDACs VPA and sodium butyrate can be added (Subheading 3.8, **step 8**) to enhance stable transgene expression; add them to a final concentration of $1\text{--}10$ mM [19] (*see* **Note 15**).

4 Notes

1. Plasmids obtained from Addgene (<https://www.addgene.org/>), a non-profit plasmid repository, are subject to a Uniform Biological Material Transfer Agreement (UBMTA). A total of 19 transfer plasmids are made available, encoding for multiple co-expressed selection markers, as well as for different purification and detection tags [7].
2. TetR in the HEK293S GnTI[−] TetR cell line is expressed under control of a CMV promoter, which is prone to transcriptional silencing after prolonged activity. Hence, to avoid losing expression of TetR, cells should not be passaged beyond P10.
3. The pH of the growth medium depends on the concentration of CO_2 , the growth medium used (concentration of buffering agent and salinity), and on the temperature. It can be calculated according to the following modified Henderson-Hasselbalch equation:

$$\text{pH} = \text{pK}_a + \log_{10} \left(\left(52 \times \frac{[\text{NaHCO}_3 \text{ (g/L)}]}{\% \text{CO}_2} \right) - 1 \right)$$

The pK_{a1} of carbonic acid (H_2CO_3) at $37^\circ C$ in an aqueous solution with physiological ionic strength in equilibrium with CO_2 , is 6.1. To ensure a physiological pH, open culture vessels (i.e., with filter cap) should be used in a properly configured CO_2 incubator. The sodium bicarbonate ($NaHCO_3$) concentration in DMEM-F12 is 2.438 g/L; hence, a pH of 7.40 at $37^\circ C$ is achieved at a CO_2 concentration of 6.0%. For FreeStyle 293 expression medium, a CO_2 concentration of 8.0% is recommended by the manufacturer (Thermo Fisher Scientific). We recommend the following technical bulletin (Thermo Fisher Scientific) for a detailed treatment of pH and pressure in open and closed tissue culture vessels: <https://assets.fishersci.com/TFS-Assets/LSG/Application-Notes/D19558.pdf>.

4. Confluency is defined as the percentage of the flask area covered by adherent cells. For adherent HEK293 cells, 100% confluency corresponds to a cell density of $\sim 250,000$ cells/cm², as we determined by flow cytometry [7]. This is a useful reference number for calculating seeding densities of plates, flasks, and bottles.
5. The HEK293 cell lines generally have a doubling time of 24–36 h. For adherent cells, this corresponds to splitting the population 1/10 to 1/5, respectively, to reach confluency after 3.5 days of growth. The exact doubling time is however influenced by many factors: the cell line, the culturing method, whether or not cells are transfected or transduced, the nature of the GOI, as well as other variables such as inaccuracies in incubator temperature and CO_2 level. Therefore, the suggested timelines are only indicative; viability, growth and cell density should always be monitored on a daily basis, and the timings should be adjusted accordingly.
6. FreeStyle 293 expression medium is a chemically defined medium that does not normally require supplementation with FBS and that is optimized for culturing adapted 293-F, 293-H, and FreeStyle 293-F cells (Thermo Fisher Scientific). We still add 1% FBS to facilitate the adaptation of the HEK293S GnTI[−] TetR cells to FreeStyle 293 medium.
7. To decant growth and conditioned medium from a flask without disturbing the adherent cells, gently rotate the flask vertically such that the medium flows towards the bottom. Then, decant the medium so that it flows over the side of the flask opposite to the cells, and into the collection or waste vessel. To wash cells with PBS, to add Trypsin-EDTA, or to add fresh medium, hold the flask at a 45° angle and pipette the liquid onto the side of the flask opposite to the cells. Then, gently

rotate the flask horizontally such that the liquid fully covers the cells .

8. The RPTP σ signal sequence targets the nascent protein to the secretory pathway. In some cases, it could be desirable to remove this sequence. These cases include intracellular target proteins, membrane proteins that use transmembrane segments for targeting to the secretory pathway, or better performance of the native signal sequence.
9. Due to plasmid instability (presence of LTRs, AT-rich regions, etc.), lentiviral plasmids should be replicated in a suitable bacterial strain with a reduced frequency of homologous recombination [20]. Minipreps from lentiviral plasmids typically give a lower yield ($\sim 10\mu\text{g}/\text{miniprep}$), with a greater variability depending on multiple factors including the insert, the backbone, the bacterial strain, and the growth conditions.
10. The lentiviral titer produced by the producer cell line depends on multiple factors such as the identity of the lentiviral transfer vector, the efficiency with which the producer cell line is transfected, and the molar ratio of the transfer plasmid to the envelope and packaging plasmids. Using the 293 T Lenti-X cell line, titers up to 10^8 infectious units per mL (IFU/mL) can be achieved, which is ~ 30 -fold higher than what can be achieved using regular HEK293T cells, according to the manufacturer (Takara Bio).
11. Polybrene (hexadimethrine bromide) is a cationic polymer that reduces charge repulsion between viral particles and the cell membrane to promote virus-host cell fusion [21].
12. Insert size strongly correlates with viral titer [22], and viral titer subsequently determines the MOI. Hence, if the titer of the lentivirus-containing supernatant from the producer cell line is not determined, transduction is performed at an unknown MOI [7]. The probability $P(n)$ that a given cell will be infected by n virus particles when inoculated with an MOI of m for a given population can be calculated using a Poisson distribution [23]:

$$P(n) = \frac{m^n \times e^{-m}}{n!}$$

For example, to calculate the fraction of non-infected cells at an MOI of 1:

$$P(0) = \frac{1^0 \times e^{-1}}{0!} = 0.368$$

The average fraction of infected cells as a result of inoculation with a given MOI m can be obtained by:

$$P(n > 0) = 1 - P(0) = 1 - e^{-m}$$

In these formulas, the number of “functional” infectious particles (TU) per cell can substitute for a strict physical particle count per cell (MOI) since viral particles that are defective or that fail to infect their target cell will not produce a transduction and genomic integration event.

13. For flow cytometry and cell sorting applications, we recommend using soybean trypsin inhibitor (SBTI) instead of complete medium to inactivate trypsin and prevent cell clumping. We use trypsin-EDTA solution without phenol red to avoid background fluorescence.
14. Although we recommend using roller bottles with filter cap in a 6% CO₂ atmosphere, this is not a strict necessity and we routinely also use roller bottles with closed cap as previously described [24].
15. Chemicals that can be added at the point of protein expression: the histone deacetylase inhibitors (iHDACs) VPA and sodium butyrate can enhance stable transgene expression; add them to a working concentration of 1–10 mM [19]. In the case of HEK293T cells, add kifunensine to a final concentration of 5 μ M (from a 500- μ M 100 \times stock solution) to inhibit class I α -mannosidases and obtain protein carrying EndoH-sensitive N-linked glycans [25].
16. The goal is to balance all these expression parameters (temperature, histone deacetylase inhibitor and Dox concentration, collection time) such that they lead to maximal production of secreted, correctly folded protein. They are best first determined in small-scale culture and optimized for each target protein [7].
17. As an alternative, cells can be trypsinized, although adaptation to suspension growth may take longer.
18. The polycarbonate Erlenmeyer flasks for suspension cell culture can be washed, autoclaved, and reused. When the filter caps deteriorate after rounds of autoclaving, the caps can be separately purchased and replaced. This provides a cost-effective alternative to disposable plasticware. As an alternative, the plastic bottles in which the culture media are delivered can be used as culture vessels as previously described [24]. In any case, make sure that the cell culture volumes do not exceed 40% of the vessel’s total volume.
19. Suspension-adapted HEK293S GnTI[−] (TetR) cells adapt readily from an adherent monolayer to suspension culture. When grown in low-serum FreeStyle 293/1% FBS medium and seeded at a minimal density of $\sim 0.5 \times 10^6$ cells/mL, a doubling time of ~ 24 – 36 h can be expected.

20. Suspension cells can be grown up to $\sim 4.0\text{--}5.0 \times 10^6$ cells/mL to further increase total biomass at harvest, which is the maximum viable cell density when using traditional HEK293 culture media. We strongly advise monitoring cell viability throughout the experiment to decide on the appropriate time for induction of protein expression or harvesting of cells.

References

1. Lin YC, Boone M, Meuris L et al (2014) Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat Commun* 5:4767
2. Aricescu AR, Lu W, Jones EY (2006) A time- and cost-efficient system for high-level protein production in mammalian cells. *Acta Crystallogr D Biol Crystallogr* 62:1243–1250
3. Durocher Y, Perret S, Kamen A (2002) High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. *Nucleic Acids Res* 30:E9
4. Chaudhary S, Pak JE, Gruswitz F et al (2012) Overexpressing human membrane proteins in stably transfected and clonal human embryonic kidney 293S cells. *Nat Protoc* 7:453–466
5. Goehring A, Lee CH, Wang KH et al (2014) Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. *Nat Protoc* 9(11):2574–2585
6. Dukkupati A, Park HH, Waghay D et al (2008) BacMam system for high-level expression of recombinant soluble and membrane glycoproteins for structural studies. *Protein Expr Purif* 62:160–170
7. Elegheert J, Behiels E, Bishop B et al (2018) Lentiviral transduction of mammalian cells for fast, scalable and high-level production of soluble and membrane proteins. *Nat Protoc* 13:2991–3017
8. Naldini L, Blomer U, Gallay P et al (1996) In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* 272:263–267
9. Zufferey R, Donello JE, Trono D, Hope TJ (1999) Woodchuck hepatitis virus posttranscriptional regulatory element enhances expression of transgenes delivered by retroviral vectors. *J Virol* 73:2886–2892
10. DuBridge RB, Tang P, Hsia HC et al (1987) Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Mol Cell Biol* 7:379–387
11. Reeves PJ, Callewaert N, Contreras R, Khorana HG (2002) Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proc Natl Acad Sci U S A* 99:13419–13424
12. Reeves PJ, Kim JM, Khorana HG (2002) Structure and function in rhodopsin: a tetracycline-inducible system in stable mammalian cell lines for high-level expression of opsin mutants. *Proc Natl Acad Sci U S A* 99:13413–13418
13. Finkelshtein D, Werman A, Novick D et al (2013) LDL receptor and its family members serve as the cellular receptors for vesicular stomatitis virus. *Proc Natl Acad Sci U S A* 110:7306–7311
14. Yao F, Svensjo T, Winkler T et al (1998) Tetracycline repressor, tetR, rather than the tetR-mammalian cell transcription factor fusion derivatives, regulates inducible gene expression in mammalian cells. *Hum Gene Ther* 9:1939–1950
15. Zufferey R, Nagy D, Mandel RJ et al (1997) Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. *Nat Biotechnol* 15:871–875
16. Zufferey R, Dull T, Mandel RJ et al (1998) Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery. *J Virol* 72(12):9873–9880
17. Marcucci KT, Jadowsky JK, Hwang WT et al (2018) Retroviral and lentiviral safety analysis of gene-modified T cell products and infused HIV and oncology patients. *Mol Ther* 26(1):269–279
18. Montini E, Cesana D, Schmidt M et al (2009) The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of HSC gene therapy. *J Clin Invest* 119(4):964–975
19. Backliwal G, Hildinger M, Kuettel I et al (2008) Valproic acid: a viable alternative to sodium butyrate for enhancing protein expression in mammalian cell cultures. *Biotechnol Bioeng* 101:182–189

20. Al-Allaf FA, Tolmachov OE, Zambetti LP et al (2013) Remarkable stability of an instability-prone lentiviral vector plasmid in *Escherichia coli* Stbl3. *Biotech* 3:61–70
21. Davis HE, Rosinski M, Morgan JR, Yarmush ML (2004) Charged polymers modulate retrovirus transduction via membrane charge neutralization and virus aggregation. *Biophys J* 86:1234–1242
22. Kumar M, Keller B, Makalou N, Sutton RE (2001) Systematic determination of the packaging limit of lentiviral vectors. *Hum Gene Ther* 12:1893–1905
23. Ellis EL, Delbruck M (1939) The growth of bacteriophage. *J Gen Physiol* 22:365–384
24. Seiradake E, Zhao Y, Lu W, Aricescu AR, Jones EY (2015) Production of cell surface and secreted glycoproteins in mammalian cells. *Methods Mol Biol* 1261:115–127
25. Chang VT, Crispin M, Aricescu AR et al (2007) Glycoprotein structural genomics: solving the glycosylation problem. *Structure* 15:267–273



Chapter 5

Transient Transfection and Expression of Eukaryotic Membrane Proteins in Expi293F Cells and Their Screening on a Small Scale: Application for Structural Studies

Ganna O. Krasnoselska, Maud Dumoux, Nadisha Gamage, Harish Cheruvara, James Birch, Andrew Quigley, and Raymond J. Owens

Abstract

Cancers, neurodegenerative and infectious diseases remain some of the leading causes of deaths worldwide. The structure-guided drug design is essential to advance drug development for these important diseases. One of the key challenges in the structure determination workflow is the production of eukaryotic membrane proteins (drug targets) of high quality. A number of expression systems have been developed for the production of eukaryotic membrane proteins. In this chapter, an optimized detailed protocol for transient transfection and expression of eukaryotic membrane proteins in Expi293F cells is presented. Testing expression and purification on a small scale allow optimizing conditions for sample preparation for downstream structural (cryo-EM) elucidation.

Key words Protein purification, Mammalian expression system, Expi293F cells, Transient expression, Membrane proteins, GPCR, Detergent screen, Small-scale tests, FSEC

1 Introduction

Within the last decade, there was an obvious increase in the number of deposited structures of human membrane proteins linked to developments in both, cryo-EM imaging (see for reviews [1, 2]) and recombinant protein production technologies [3] (Fig. 1a).

Membrane proteins account for up to one third of proteins encoded in genomes [4, 5] and for humans, their number exceeds 6000 [5]. They perform a wide array of vital functions in the cell and remain the major category of targets for approved drugs [6–8]. While being attractive as pharmaceutical targets, proteins with transmembrane segments and large hydrophobic surface remain problematic for production, extraction, and purification. The classical studies of integral membrane proteins remain laborious and

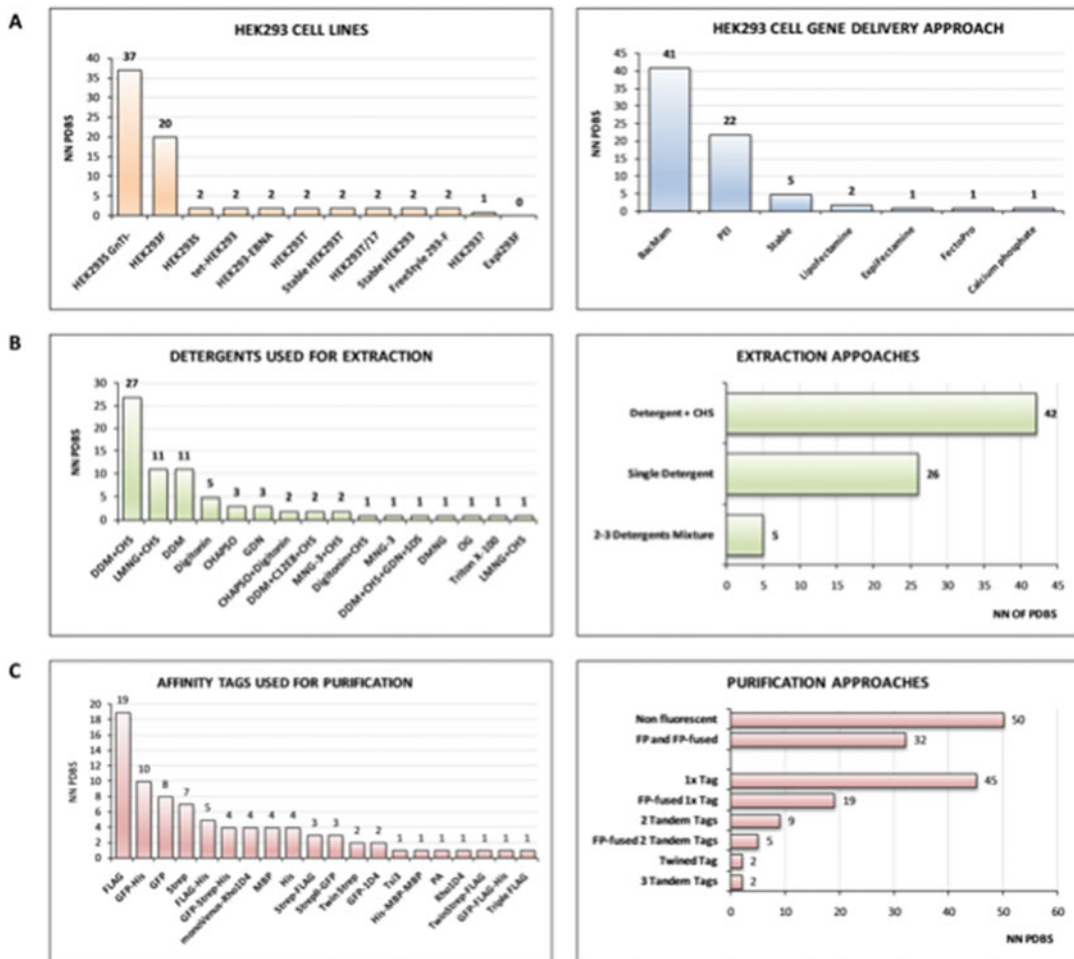


Fig. 1 Annual deposition of the human membrane protein structures 1997–2019 and expression hosts used for protein production. **(a)** A number of human membrane protein (MPs) structures released annually (state on 5.01.2020). In the last 2 years, more than 120 structures of human MPs were deposited to the Protein Data Bank (PDB) archive. **(b)** Expression hosts used to produce human MPs for downstream structural analysis. The majority of targets were produced in insect and HEK293 expression systems. Up to now, a high prevalence of using insect cells system to express GPCR family members is observed

involve their overexpression in heterologous systems, detergent-mediated extraction from the cell and purification.

To get good quality samples for structural studies, usually multiple screenings are required. Nowadays, the quality of the samples used in research (e.g., yield, physical characteristics, functional activity, applicability for structural studies) largely depends on parameters of the protein production system [3, 9–13] (Figs. 1 and 2). Such parameters as expression system (cell line, expression medium, expression method), the design of expression construct (fusion tag and its position, codon optimization; truncations and point mutations), and the applied conditions for protein extraction

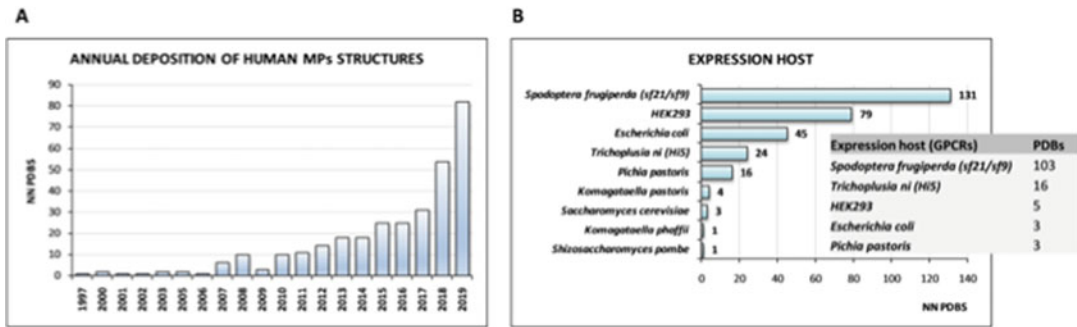


Fig. 2 Approaches used for the production of human MPs in HEK293-derived cell lines and protein purification strategies. Within the last decade (2010–2019) numerous human MPs were successfully produced in HEK293 cells, purified and used for atomic-level structural determination (over 75 structures are currently deposited in PDB archive). The protein production and extraction strategies for individual proteins within this group differ. **(a)** Statistics on HEK293-derived cell lines and DNA delivery systems used for protein production. The HEK293S GnTI[−] cell line is the most commonly used host cell line and BacMam system-based transduction of HEK293 cells remains prevalent in human MPs production pipelines. **(b)** Detergents used for the extraction of human MPs. At the initial step of protein solubilization, when protein extraction in the native oligomeric and functional state is essential, in most cases detergents of choice (DDM, LMNG, MNG-3, Digitonin) were supplemented with cholesteryl hemisuccinate (CHS) to increase protein stability in solution. The use of single detergents (DDM, Digitonin, CHAPS, GDN, MNG-3, DMNG, OG, and Triton X-100) or mixed micelles also leads to the extraction of multiple stable proteins. **(c)** Statistics on affinity tags used to purify human MPs. Different single and tandem tags were used for the purification of target proteins. To enable in-cell detection of expression and for further purification, numerous protein targets have been expressed fused to fluorescent proteins (FP)

and purification (buffer additives, ligands, lipids, detergents) are all important variables.

While the major class of human membrane proteins with deposited structures (~40%), GPCR family members, were mostly produced in insect cells with the use of viruses, members of other protein families (immune receptors, ion channels, and transporters) were preferentially produced in mammalian cells (Fig. 1b). Human Embryonic Kidney (HEK293) and Chinese Hamster Ovary (CHO) cells are the most popular mammalian expression systems for the production of target proteins [14]: CHO cells remain the main cell factory to produce pharmaceutical targets [15] and HEK293 cells are largely exploited to produce targets for research. The typical yields of human membrane proteins are in the range of 3–200 µg for every liter of HEK293 cell culture [16–18]. In a very few cases, higher yields (~0.5–1 mg/L) were reported [21, 22].

With respect to the overexpression of human membrane proteins for structural studies, human cell lines (ATCC collection contains >2000 human cell lines and hybridomas of different cell origin, cell type, and application) excels in the production of complex proteins as they provide a native microenvironment for proper folding, processing, and post-translational modifications all

essential for the functional integrity of proteins. The multiple available HEK293-derived cell lines can grow in both, adherent and suspension cultures, and they display high susceptibility to both, virus transduction (baculoviruses and lentiviruses) and transient transfection achieved with the aid of chemical compounds (cationic liposomes and cationic polymers) [21–23]. Therefore, the protein production design space for HEK293 cells is immense. While baculovirus transduction of mammalian cells (BacMam) is an established approach (e.g., BacMam transduction of suspension-adapted HEK293S GnT1[−] in [24]) used to produce numerous human membrane proteins for structural studies, the transient transfection and expression approach remains much less used (Fig. 2a).

The recent progress in the field of recombinant protein production makes HEK293 system more attractive for protein production and may facilitate the production of human membrane proteins by transient approach. As such, new-engineered cell lines (Expi293F and Expi293E) are capable of growing to higher density in suspension (healthy to up to 6×10^6 cells/mL) and perform required post-translational modifications. To obtain the aberrant protein glycosylation phenotype, the HEK293S GnT1[−] cell line is available [25]. Cheaper alternatives to commonly used transfection agents can be purchased (e.g., cationic polyethylenimines, PEI) [22, 26]. Much work was done on improving promoters and their regulatory elements and in-cell fluorescence detection approach for faster and better detection of produced proteins.

Despite there is currently no universal approach appropriate for eukaryotic protein production in HEK293 cells, there are few step-by-step protocols and guides on establishing expression in HEK293 cells and performing screening for proteins with a different application for research [27–31].

In this chapter, we detail an optimized step-by-step protocol to produce full-length eukaryotic membrane proteins using transient transfection and expression in small suspension cultures of Expi293F cells for higher yields and better quality of proteins at a lower cost. The most notable advantages of Expi293F cell line are high-density growth in suspension culture, ease of cell transfection, simple scale-up, versatile protein expression, and production of proteins with required post-translational modifications. To enable the production of full-length toxic to cell proteins, we modified the cell pre-treatment procedure and used conditions of low hypothermia for cell growth after transfection. In our system, in-cell detection of expressed proteins with the use of fluorescence detection techniques allows fast analysis of targets and constructs at a low scale.

Table 1
Overview of tested eukaryotic membrane proteins

Gene name	Localization in cell	Source	TM	kDa
<i>Ntsr1</i>	Plasma membrane	<i>Rattus norvegicus</i>	7	47
<i>SLC10A1</i>	Plasma membrane	<i>Bos taurus</i>	9	41
<i>ADORA2A</i>	Plasma membrane	<i>Homo sapiens</i>	7	45
<i>SLC6A1</i>	Plasma membrane	<i>Homo sapiens</i>	12	67
<i>SLC35D1</i>	ER membrane	<i>Homo sapiens</i>	8	39
<i>SLC35D2</i>	Golgi apparatus membrane	<i>Homo sapiens</i>	10	37

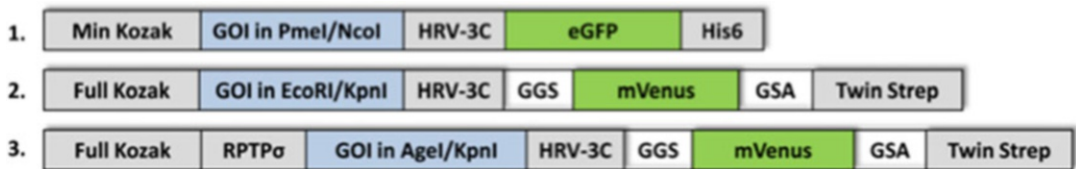


Fig. 3 Overview of used expression constructs. Three different C-terminal protein construct variants were used for expression and purification of selected targets. In all cloned constructs, the gene of interest (GOI) was fused to a fluorescent protein (eGFP or mVenus) and affinity tag (His6 or Twin Strep) through cleavable linkage. With one of the constructs (N3), the effect of the signal sequence (RPTPσ) on the expression of membrane proteins was assessed

The protocol was established by a case study of six eukaryotic membrane proteins (Table 1). The selected targets belong to different protein families, including two members of the GPCR family, and have different subcellular localization. To test whether different choices of promoters (CAG and CMV) and fusion tags (eGFP-His and mVenus-Strep) could affect the transient expression in Expi293F cells, we used two different vector systems and three expression construct variants (Fig. 3). Namely, the protocol was tested with pOPINE 3C-eGFP-His6 vector (one of the pOPIN multi-target vectors available from our lab (<https://www.oppf.rc-harwell.ac.uk/OPPF/>) and pHR-CMV-TetO2 vector (transfer vectors suitable for both, transient expression in HEK293 cell lines and generation of lentiviruses for transduction-mediated expression and production of stable HEK293S GnT1⁻ and HEK293S GnT1⁻ TetR cell lines [32]). We show that for some particular targets (e.g., *Ntsr1* GPCR target) the N-terminal fusion of protein sequence to secretion signal peptide (in this case, RPTPσ—MPALLSLVSLLSVLLMGCV) can improve protein expression in Expi293F cells.

In a few recent publications, the successful use of mixed detergents for protein extraction for downstream structural analysis was reported. Namely, DDM, GDN, and SDS were used for extraction of voltage-gated sodium channel $\text{Na}_v1.2$ [33]; (ii) CHAPSO and Digitonin was used for extraction of intracellular protease γ -secretase [34, 35]; and (iii) DDM and C12E8 was used for extraction of ABC transporter Pgp [36, 37]. In our work, we also tested a set of mixed and single detergents for protein extraction and purification and found that CYMAL-6 and mixed micelles can be a good alternative to the most commonly used DDM detergent.

2 Materials

2.1 Sub-cloning Genes of Interest in pOPIN Vectors Using In-Fusion Cloning Technique

1. Gene of interest (cDNA, synthetic gene, or other).
2. pOPINE 3C-eGFP-His6 vector digested with PmeI/NcoI restriction enzymes.
3. Primers with 15 bp extensions overlapping with in-fusion entry sites.
4. Phusion[®] High-Fidelity PCR kit (New England BioLabs).
5. DpnI restriction enzyme.
6. AMPure XB Beckman Coulter magnetic beads.
7. 96-well Magnetic Separator.
8. As Alternative: NucleoSpin[®] Gel and PCR Clean-up kit (MACHEREY-NAGEL).
9. Elution TE Buffer (10 mM Tris-Acetate pH 8.0, 1 mM EDTA).
10. Vazyme ClonExpress II One Step Cloning kit.
11. 37 °C incubator.
12. Stellar competent cells (Takara Bio).
13. Sterile (multi-well, not TC) plates with lids.
14. Autoclaved LB medium.
15. 50 mg/mL sterile-filtered carbenicillin stock.
16. LB-agar plates supplemented with 50 μ g/mL carbenicillin, 1 mM IPTG, and 20 μ g/mL X-gal.
17. QIAGEN Miniprep kit.
18. Autoclaved 100% Glycerol stock.

2.2 Preparation of Transfection-Grade Plasmids

1. QIAGEN Plasmid Plus Midi kit.
2. Vacuum manifold (such as Promega Vac-Man[™]).
3. 1.5 mL sterile Eppendorf tubes.
4. Microvolume spectrophotometer such as NanoDrop[™].

**2.3 Transient
Transfection
and Expression
in Expi293F Cells:
General Procedure**

1. Laminar flow hood.
2. Purified plasmid DNA of interest.
3. Expi293F cells (cell line catalog number A14527).
4. Gibco Expi293™ Express medium.
5. 125 and 500/1000 mL sterile plain bottom flasks with vented closure (ThermoFisher Scientific).
6. CO₂ orbital shaker.
7. Tabletop centrifuge suitable for 50 mL Falcon tubes (Sorvall Legend RT Plus).
8. Trypan blue stain (4% solution, Gibco).
9. Countess Automated Cell Counter (Invitrogen).
10. Countess cell counting chamber slides (Invitrogen).
11. 0.22μM syringe sterile filters (Fisher scientific).
12. Sterile filtered 1 mg/mL Polyethylenimine PEI MAX 40 K (water solution, pH titrated to 7.0 with NaOH) (Polysciences Inc., 24,765-1).
13. Gibco OPTI-MEM reduced serum medium.
14. Sterile-filtered stock solutions of enhancers prepared in Expi293™ Express medium: 45% glucose, 0.3 M valproic acid, and 1 M sodium propionate.
15. 50 mL sterile Falcon tubes.
16. Automatic pipette filler.
17. Sterile serological pipettes (1, 5, 10, 25, and 50 mL).
18. Rainin filter tips.

**2.4 Screening
of Expression
Parameters on a Small
Scale: 1 mL/12-Well
Plate Expression**

1. 12-well tissue culture-treated plates with lid (Greiner CELLSTAR®).

**2.5 Up-Scaled
Expression**

1. 500/1000 mL sterile plain bottom flasks with vented closure (ThermoFisher Scientific) or 1000 mL roller bottle (BIOFILL).

**2.6 Visualization
of Expression by
Fluorescence
Microscopy**

1. EVOS fluorescent microscope with 20× objective lens and EVOS™ Light Cube for GFP detection (Excitation/Emission wavelength = 470/525 nm).
2. Compatible and calibrated for imaging TC plates (96-, 24-, 12-, and 6-well plates can be used).

2.7 Analysis of Expression with Tali Imaging System

1. Tali™ Image-Based Cytometer (Invitrogen).
2. Tali™ Cellular Analysis slides (Invitrogen).

2.8 Analysis of Expression Using In-Gel GFP Fluorescence of Cell Probes

1. 2× Loading Dye (100 mM Tris/HCl pH 6.8, 20% glycerol (v/v), 200 mM DTT, 4% SDS, 0.2% bromophenol blue).
2. Inhibitors cocktail for mammals (P8340, Sigma).
3. DNase I (SLBW0018, Sigma).
4. Tabletop centrifuge suitable for 1.5 mL Eppendorf tubes (such as Beckman Coulter Microfuge® 16).
5. Optional: Sonic bath.
6. Vertical rotating platform suitable for 1.5 mL Eppendorf tubes (such as HulaMixer® Sample Mixer, Invitrogen).
7. 10% Bis-Tris NuPAGE™ Midi protein Gels (1 mm, 26-well).
8. 1× NuPAGE MOPS or 1× NuPAGE MES buffer.
9. Bio-Rad Precision Plus Protein™ Dual Color Marker.
10. Novex BenchMark™ Fluorescent Protein Marker.
11. Bio-Rad PowerPac™ Basic Power Supply.
12. Bio-Rad ChemiDoc MP Imaging system with excitation (485) and emission (525) filters.

2.9 Membranes Preparation

1. Low spin centrifuge (such as Beckman Coulter Allegra X-15R).
2. Ice-cold Buffer1 containing 20 mM Tris/HCl pH 8.0, 100 mM NaCl, 10 mM MgSO₄, 5% glycerol (all required Buffers are summarized in Table 2).

Table 2
Composition of buffers

NN	Composition
Buffer1	20 mM Tris/HCl pH 8.0, 100 mM NaCl, 10 mM MgSO ₄ , 5% glycerol
Buffer2	20 mM Tris/HCl pH 8.0, 100 mM NaCl, 10 mM MgSO ₄ , 25 mM imidazole pH 8.0
Buffer3	20 mM Tris/HCl pH 8.0, 100 mM NaCl, 10 mM MgSO ₄ , 25 mM imidazole pH 8.0, 0.05% DDM
Buffer4	20 mM Tris/HCl pH 8.0, 100 mM NaCl, 10 mM MgSO ₄ , 50 mM imidazole pH 8.0, 0.05% DDM
Buffer5	20 mM Tris/HCl pH 8.0, 100 mM NaCl, 10 mM MgSO ₄ , 400 mM imidazole pH 8.0, 0.05% DDM
Buffer6	100 mM Tris pH 8.0, 150 mM NaCl, 10 mM MgSO ₄ , 0.05% DDM
Buffer7	100 mM Tris pH 8.0, 150 mM NaCl, 10 mM MgSO ₄ , 50 mM biotin, 0.05% DDM
Buffer8	10 mM Tris pH 8.0, 150 mM NaCl, 0.05% DDM

3. Benchtop ultrasonic disintegrator (MSE Soniprep150 Plus Ultrasonic Disintegrator).
4. Floor standing Beckman Coulter Optima L-100 XP ultracentrifuge.
5. Ti 45 rotor type and compatible tubes.
6. Douncer homogenizer.
7. Ice-cold Buffer2 containing 20 mM Tris/HCl pH 8.0, 100 mM NaCl, 10 mM MgSO₄, 25 mM imidazole pH 8.0.

2.10 Screening of Detergents for Protein Extraction from Membranes on a Small Scale

1. 100–200× cmc or/and 10% detergent stocks (water solutions).
2. Benchtop Beckman Coulter MAX-XP ultracentrifuge.
3. TLA 55 rotor.
4. Beckman Coulter Microcentrifuge 1.5 mL tubes.

2.11 Purification of His-Tagged Targets in Different Detergents on a Small Scale

1. Buffer 3 containing 20 mM Tris/HCl pH 8.0, 100 mM NaCl, 10 mM MgSO₄, 25 mM imidazole pH 8.0, 0.05% DDM.
2. Ni-NTA agarose pre-equilibrated in Buffer3.
3. Washing Buffer 4 containing 20 mM Tris/HCl pH 8.0, 100 mM NaCl, 10 mM MgSO₄, 50 mM imidazole pH 8.0, 0.05% DDM.
4. Elution Buffer 5 containing 20 mM Tris/HCl pH 8.0, 100 mM NaCl, 10 mM MgSO₄, 400 mM imidazole pH 8.0, 0.05% DDM.

2.12 Purification of Strep-Tagged Targets in Different Detergents on a Small Scale

1. MagStrep “type3” XT Beads 5% suspension (IBA Lifesciences).
2. Biotin.
3. Buffer 6 containing 100 mM Tris pH 8.0, 150 mM NaCl, 10 mM MgSO₄, 0.05% DDM.
4. Buffer 7 containing 100 mM Tris pH 8.0, 150 mM NaCl, 10 mM MgSO₄, 50 mM biotin, 0.05% DDM.
5. 24-well deep-well plates.
6. 24-well magnetic separator.

2.13 Quality Control of His- and Strep-Tag-Purified Samples: FSEC

1. SRT-C-300 HPLC system column (20 mL).
2. HPLC system.
3. Chromacol 0.3 mL Screw Top Fixed Insert Vial (Thermo-Fisher, 03-FISV) and Thermo Scientific™ 9 mm Autosampler Vial Screw Thread Caps.
4. Freshly filtered and degassed running Buffer8 (20 mL for each sample) containing 10 mM Tris pH 8.0, 150 mM NaCl, 0.05% DDM.

- 2.14

Purification from Small-Scale Expression Tests (3 mL/6-Well Plate)

1. 6-well tissue culture-treated plates with lid (Greiner CELLSTAR®).

2. 24 Tip Horn for use with ultrasonic disintegrator (if multiple samples will be analyzed).

3. Ammonium sulfate solution saturated in 50 mM Tris/HCl pH 8.0.
- 2.15

Large-Scale Affinity Purification

Consumables will depend on results of test expressions and analytical purifications.

3 Methods

- 3.1

Sub-cloning Genes of Interest in pOPIN Vectors Using In-Fusion Cloning Technique

The full list of multi-target pOPIN vectors is available at <https://www.oppf.rc-harwell.ac.uk/OPPF/protocols/cloning.jsp>. In our studies, the good level of expression of full-length targets was achieved with pOPINE 3C-eGFP-His6 vector.

1. Design pairs of primers with 15 bp extensions overlapping with in-fusion entry sites (example in Table 3).

2. Amplify target gene in 50μL PCR reaction using recommended for DNA polymerase settings.

3. PCR reaction must be followed by 1-h DpnI digestion.

4. Purify resulting PCR fragments using AMPure XB Beckman Coulter magnetic beads (80μL for each PCR reaction) and elute in 20μL of TE buffer. As alternative spin column-based purification of PCR products can be used.
- Table 3
Primers for In-Fusion cloning of targets in PmeI/NcoI restriction sites of pOPINE 3C-eGFP-His6 vector
- | Gene name | Oligonucleotide sequence 5'→3'
(fwd primer) | Oligonucleotide sequence 5'→3'
(rev primer) |
|----------------|---|---|
| <i>Ntsr1</i> | <u>AGGAGATATACCATG</u>
CACCTCAACAGCTCCGTGC | <u>CAGAACTTCCAGTTT</u>
AGGACAAAGGCAGGCCAGCG |
| <i>SLC10A1</i> | <u>AGGAGATATACCATG</u>
GAGGCCTTCAACGAATCTTCC | <u>CAGAACTTCCAGTTT</u>
GTTTGCCATGTTGAGTTGCTC |
| <i>ADORA2A</i> | <u>AGGAGATATACCATG</u>
CCCATCATGGGCTCCTCG | <u>CAGAACTTCCAGTTT</u>
GTCCGTGGCGTAGGTCTGG |
| <i>SLC6A1</i> | <u>AGGAGATATACCATG</u>
GCGACCAACGGCAGCAA | <u>CAGAACTTCCAGTTT</u>
GATGTAGGCCTCCTTGCTGG |
| <i>SLC35D1</i> | <u>AGGAGATATACCATG</u>
GCGGAAGTTCATAGACG | <u>CAGAACTTCCAGTTT</u>
CAACACTGCTCCTTTCCCCT |
| <i>SLC35D2</i> | <u>AGGAGATATACCATG</u>
ACGGCCGGCGGCCAGGC | <u>CAGAACTTCCAGTTT</u>
GCTCTTCAAATCCAAACAGA |

5. For 10 μ L of In-Fusion reaction, mix 50–100 ng (1–3 μ L) of purified PCR fragments, 100 ng (1–2 μ L) of PmeI/NcoI double-digested pOPINE 3C-eGFP-His6 vector, 1 μ L Exnase II, and 2 μ L optimized buffer supplied with the cloning kit.
6. Incubate reaction 25 min at 37 °C and then stop immediately by adding 20 μ L of ice-cold TE buffer.
7. Use 5 μ L of the resulting reaction mixture to transform 20 μ L Stellar competent cells using standard heat shock transformation protocol [38].
8. To clone multiple constructs prepare and use sterile 2 mL LB-agar/24-well plates with lids (not tissue culture treated).
9. For X-gal blue/white screening of recombinant plasmids use LB-agar plates supplemented with 50 μ g/mL carbenicillin, 1 mM IPTG, and 20 μ g/mL X-gal. As positive clones pick only white colonies and culture them overnight in 10 mL of LB medium freshly supplemented with 50 μ g/mL carbenicillin (*see Note 1*).
10. Use cell pellets from overnight cultures for plasmid preparation using QIAGEN Miniprep kit.
11. Confirm obtained DNA constructs by sequencing obtained clones with T7 fwd and GFP rev primers.
12. Produce 50% glycerol stocks and use them to grow larger scale cultures for transfection-grade plasmid preparation.

3.2 Preparation of Transfection-Grade Plasmids

1. Purify transfection-grade plasmids (0.5–1 mg) from 150 mL overnight LB culture using QIAGEN Plasmid Plus Midi kit and if required, store plasmids at –20 °C in sterile 1.5 mL Eppendorf tubes (*see Note 2*).
2. Measure the purity and concentration of obtained DNA using a NanoDrop spectrophotometer. Plasmid DNA used for transfections should be of high purity. Good quality DNA with no protein and chemical contaminations should have the ratios of absorbance 260/280 between 1.8–2.0 and 260/230 between 2.0 and 2.2.
3. Calculate the overall amount of the DNA required for the transfection. Use 1 μ g DNA per each one million of transfected cells (*see Note 3*).

3.3 Transient Transfection and Expression in Expi293F Cells: General Procedure

1. Perform all manipulations with Expi293F cells (subculture/expand/transfect/enhance/feed) in a laminar flow hood.
2. Aspirate and dispense cells using sterile serological pipettes and automatic pipette filler. Pipettes should be discarded after a single use. Avoid vigorous mixing and pipetting of cells. Use the slow dispensing mode of pipette filler for handling cells and high-speed mode for dispensing medium.

3. Record passage number of cells and determine cell viability and total cells count during maintenance culture (*see Note 4*).
4. To check the cell number and viability by Trypan blue exclusion, take fresh 10 μ L cells aliquots, and mix them with 10 μ L Trypan blue stain. Apply 10 μ L of the resulting mixture on the cell chamber slide. Insert chamber slide in Countess Automated Cell Counter, focus the image and run the “Count” program. Only cells showing $\geq 95\%$ viability can be used for further transfection.
5. Maintain the suspension culture of Expi293F cells for at least three passages after defrosting (passage numbers 3–30 can be used in experiments) in a humidified (80%) incubator with 5–8% CO₂ at 37 °C with 120 rpm in Gibco Expi293™ Express medium at a cell density between 0.5 and 5.0×10^6 cells/mL. Use 125 mL flask to maintain 30 mL Expi293F cells. To up-scale the culture, use 500 mL flask to maintain 100 mL Expi293F cells and 1000 mL flask to maintain 300 mL cells.
6. One day before transfection seed Expi293F cells at a cell density of 1×10^6 cells/mL.
7. On the day of transfection transfer cells in sterile 50 mL Falcon tubes and shortly pellet (500 *g*, 10 min, RT), discard the supernatant and re-suspend cells in fresh pre-warmed expression medium by gentle pipetting to a final density of $2.0\text{--}2.5 \times 10^6$ cells/mL (*see Note 5*).
8. In the transfection mixture, dilute each 1 μ g DNA with 100 μ L OPTI-MEM serum-free medium and add 8 μ g polyethylenimine PEI MAX 40 K. After thorough mixing, incubate the mixture 10 min at RT and add gently (dropwise) to Expi293F cells (*see Note 6*).
9. Place cells immediately in shaking incubator and grow at 30 °C, 125–150 rpm, 5–8% CO₂, 80% humidity.
10. To boost protein expression within 20 h of post-transfection supplement the culture with the following final concentrations of enhancers: 5 mM valproic acid, 6.5 mM sodium propionate, and 0.9% glucose (*see Note 7*).
11. After transfection, grow cells for another 1–6 days (we recommend to grow cultures 6 days).
12. High cell viability ($\geq 80\%$) at the end of expression (Day 6) must be observed.

3.4 Screening of Expression Parameters on a Small Scale: 1 mL/12-Well Plate Expression

1. Run expression on a small scale (1 mL cultures in 12-well plate) to pre-screen expression conditions (e.g., two temperatures (30° and 37 °C), exchange of the medium before the transfection and expression time), and expression construct variants (affinity tags and their location) (Fig. 4).

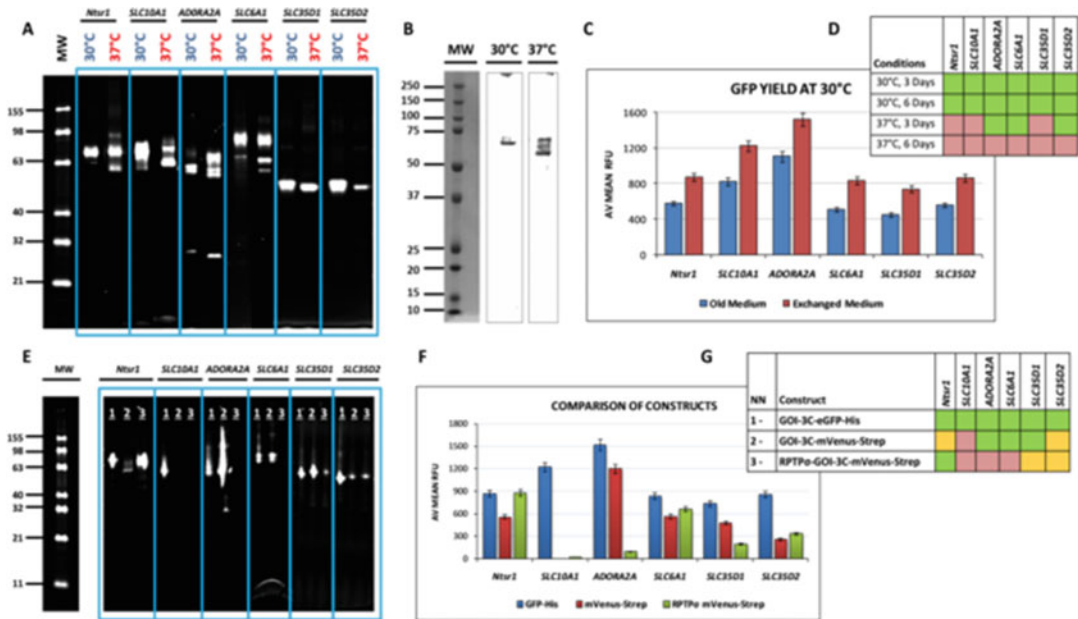


Fig. 4 Optimization of the transient expression of targets on a small-scale. Miniaturizing test expressions using 1 mL/12-well plate allows parallel processing of greater breadth of variables. (a) The in-gel fluorescent protein signals at 48 h post-transfection. Under conditions of mild hypothermia (30 °C), the quality of expressed proteins improves for most of the targets (no fragmentation). (b) Representative anti-His western blot of *SLC10A1* target expressed at two different temperatures. (c) Comparison of in-cell GFP yields of targets produced in old and freshly exchanged expression medium. A full exchange of the medium shortly before transfection increases protein yield at both temperatures, 37 and 30 °C, and allows express proteins longer (summarized in (d)). (e) The in-gel fluorescent protein signals obtained for three different construct variants expressed under optimal conditions (30 °C, fresh expression medium, and 6 days). All targets fused to GFP-His and some of the targets fused to mVenus-Strep are amenable to expression in Expi293F cells. (f) In-cell fluorescence signals indicate higher versatility of GFP-His tag and better expression of targets fused to GFP-His (summarized in (g))

2. One day before transfection seed cells at 1×10^6 cells/mL.
3. On the day of transfection exchange medium and adjust cell density as it is described above.
4. Plate freshly suspended cells in 12-well tissue culture-treated plates (1 mL in each well).
5. Transfect each well with 2µg of DNA diluted in 200µL OPTI-MEM medium and supplement with 16µg PEI MAX 40 K.
6. In case protein production is toxic to cells, fast (within 2 days) cell proteolysis and massive protein fragmentation can be observed (example in Fig. 4). These conditions must be excluded from further experiments.
7. Results of the expression tests can be analyzed using one of the cell imaging systems (EVOS microscope or/and Tali imaging cytometer) and In-Gel GFP fluorescence of probes as it is described in Subsections 3.6–3.8.

Table 4
Optimized conditions for up-scaled transient expression of targets in Expi293F cells

Step	Parameter	Recommendation	
I. Transfection (10 min)	Cell volume	100 mL	300 mL
	Amount of DNA	200–250µg	600–750µg
	Transfection agent	PEI MAX 40 K	
	Transfection medium	OPTI-MEM Serum Reduced	
	Duration	10 min	
	Temperature	RT	
II. Expression (3–6 days)	Starting cell density	$2.0\text{--}2.5 \times 10^6$	
	Cells viability	≥95%	
	Expression medium	Freshly Exchanged Gibco Expi293™	
	Flask type	Vented, 500 mL	Vented, 1000 mL
	Temperature	30 °C	
	Shake speed	125 rpm	150 rpm
	Duration	3–6 days	
	Detection	(GFP)-fluorescence	
III. Supplements	Time of addition	≤20 h post-transfection	
	Enhancer1 (Valproic acid)	1.7 mL	5.0 mL
	Enhancer2 (Sodium propionate)	0.65 mL	2.0 mL
	Feed (Glucose)	2.0 mL	5.5 mL

3.5 Up-Scaled Expression

1. The optimal transfection and expression conditions were determined using six target eukaryotic membrane proteins and are summarized in Table 4.
2. Grow transfected cells in 500 mL (100 mL cells) or 1000 mL (300 mL cells) sterile flask with vented closure. As an alternative, 1000 mL roller bottle can be used to grow 100–300 mL transfected cells.

3.6 Visualization of Expression by Fluorescence Microscopy

1. Use 12-well plate from small-scale experiments or collect fresh 1 mL cell probes from up-scaled expression and pipet them in a new 12-well plate.
2. Focus image using the white and green light detection options. Capture three fluorescent images from three randomly chosen locations under a 20× objective lens of EVOS fluorescent microscope.
3. In the case of successful transfection and expression, on average, a fluorescent image will contain several hundred green cells (Fig. 5).

3.7 Analysis of Expression with Tali Imaging System

1. Quantify cell viability and protein expression in GFP-containing cells (% of cells expressing GFP and GFP yield) in suspension cell-based assay using Tali™ Image-Based Cytometer.

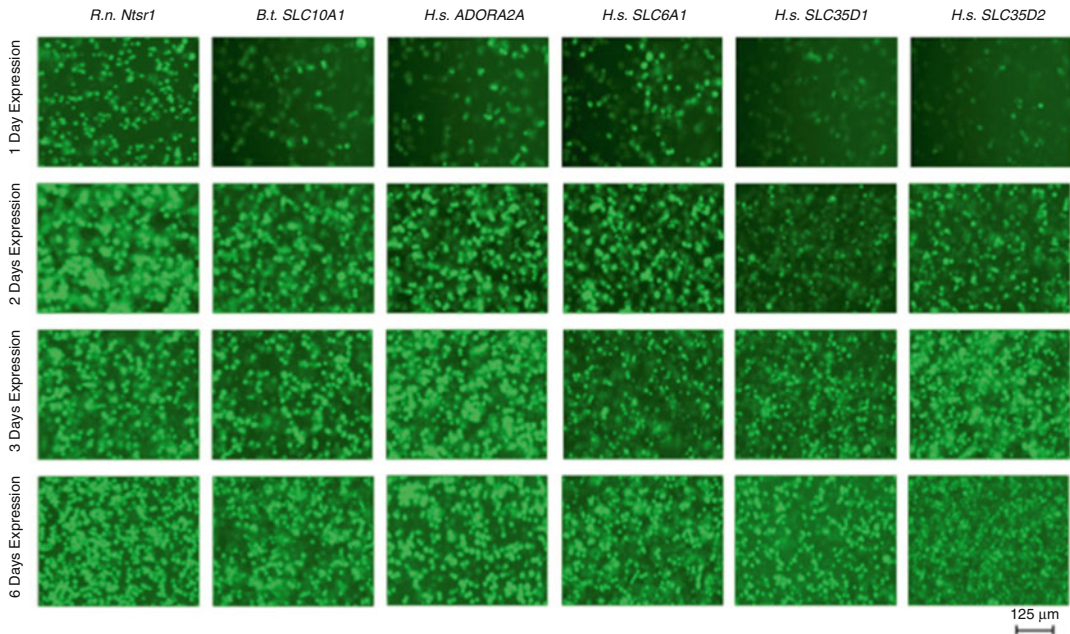


Fig. 5 Visualization of GFP-fused protein expression. In-cell GFP fluorescence signals images of targets expressed in Expi293F cells at 30 °C at different time points captured with EVOS fluorescent microscope (scale bar, 125μm). Already at 24 h post-transfection, the GFP signals for target proteins are detected. With enhancers and after longer expression (2–6 days in total) the progress in GFP fluorescence is observed

2. Pipet 20μL of freshly taken cell probes on slides supported by Tali™ Image-Based Cytometer.
3. Image cells using ≥ 9 fields.
4. Apply the RFU threshold to quantify the number of cells expressing GFP.
5. Plot and analyze data for % of cells expressing GFP and mean RFU signals of cells (Fig. 6a and b).
6. To correct data for cells and media background autofluorescence, use the negative control (cells transfected with construct without reporter gene).

3.8 Analysis of Expression Using In-Gel GFP Fluorescence of Cell Probes

1. Perform all work with probes on ice or in a cold room to preserve GFP fluorescence of targets.
2. Pipet 1 mL cell probes in 1.5 ml Eppendorf tubes.
3. Pellet cells (12,500 *g*, 10 min, 4 °C) and discard the supernatant media.
4. Suspend pellets in 150μL of 2× Loading Dye freshly supplemented with DNaseI and mixture of protease inhibitors for mammals. Mix with a pipette vigorously to get a homogeneous solution.

5. Optional: In addition, sonicate probes for 10 min in a sonic bath.
6. Mix probes 20 min at the vertical rotating platform in a cold room.
7. Spin down probes (12,500 *g*, 10 min, 4 °C) and load 5 μ L aliquots of the supernatants containing target proteins on 10% Bis-Tris gels. Do not boil samples before loading on SDS-PAGE. As a protein standard can be used 2 μ L Bench-Mark Fluorescent Protein Marker or 2 μ L Precision Plus Protein™ Dual Color Marker. Run the gel at 4 °C in 1 \times MOPS or 1 \times MES buffer for 3.5 h at 90 V.
8. Visualize In-Gel GFP fluorescence of target proteins using imaging system supplied with excitation (485) and emission (525) filters (e.g., Bio-Rad ChemiDoc MP Imaging system or other) (Fig. 6c and d).

3.9 Membranes Preparation

1. Prepare membranes from ≥ 100 mL up-scaled expression of target proteins.
2. After expression transfer cells in two sterile 50 mL Falcon tubes and collect cell pellets by short (10 min) centrifugation at 3000 *g*, 4 °C.
3. Suspend obtained pellets in 20 mL of ice-cold Buffer1 containing a freshly added mixture of protease inhibitors for mammals and DNase I.
4. Break cells on ice by 5 min sonication in 50% duty cycle with 10% amplitude and sonication pulse duration of 10 s.
5. Remove unbroken cell debris by 35 min centrifugation at 3000 *g*, 4 °C.
6. Subject obtained supernatant to 2 h ultracentrifugation at 230,000 *g*, 2 h, and 4 °C.
7. Mechanically re-suspend membrane pellets in 20 mL Buffer2 with Dounce homogenizer using 10–20 passes with a pestle.
8. Use obtained membranes for (i) small-scale detergent/buffer screening or (ii) directly for large-scale purification.

3.10 Screening of Detergents for Protein Extraction from Membranes on a Small Scale

1. Both, single detergents from different classes and mixed micelles can be used (an example of the detergent screen is provided in Fig. 7e).
2. Make sure that equal volumes of detergents are added to each probe. To do so, prepare 100 μ L stock solutions of detergents: (i) To compare extraction efficiencies of detergents according to their cmc values, prepare 100–200 \times cmc stock solutions of detergents of choice; (ii) To compare extraction efficiencies of 1% detergents, prepare 10% stock solutions of detergents to be tested.

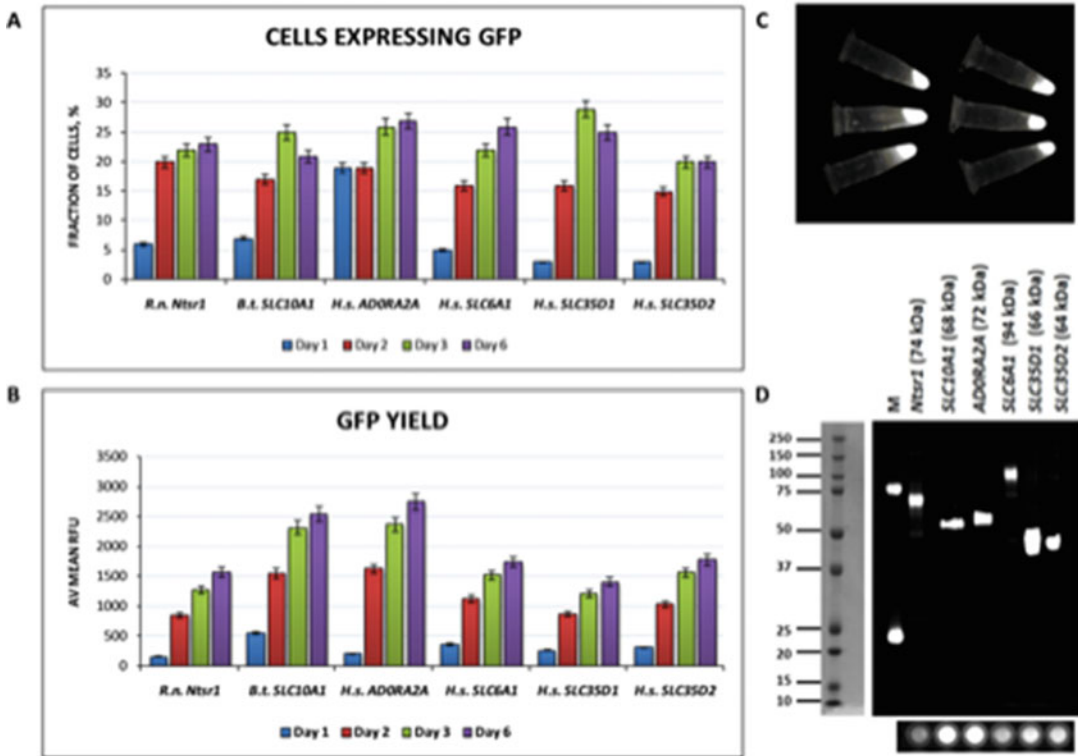


Fig. 6 Analysis of GFP-fused protein expression. Left panel: Quantification of GFP fluorescence in living cells using Tali imaging system. (a) Progression of the number of cells expressing GFP-fused targets and (b) GFP yield over days are shown. The highest GFP signals are observed at the end of expression (day 6 after transfection). Right panel: Checking GFP fluorescence of target proteins after 6 days of expression. (c) GFP fluorescence of harvested cell pellets. (d) GFP fluorescence of treated cell probes before and after SDS-PAGE analysis

- For each detergent probe, mix 0.9 mL of suspended membranes with 100 μ L of prepared detergent stock.
- Solubilize probes 1 h on the vertically rotating platform in a cold room.
- Transfer probes in 1.5 mL microcentrifuge tubes compatible with benchtop ultracentrifuge rotor.
- Clear solubilizate by 1-h centrifugation at 130,000 g , 4 $^{\circ}$ C in a benchtop ultracentrifuge.
- From the obtained supernatant, load 10 μ L of each probe on a gel.
- Evaluate the efficiency of protein extraction from membranes based on the intensity of In-Gel GFP fluorescence signals of protein bands (Fig. 6) using the Imaging system supplied with excitation (485) and emission (525) filters (Fig. 7).

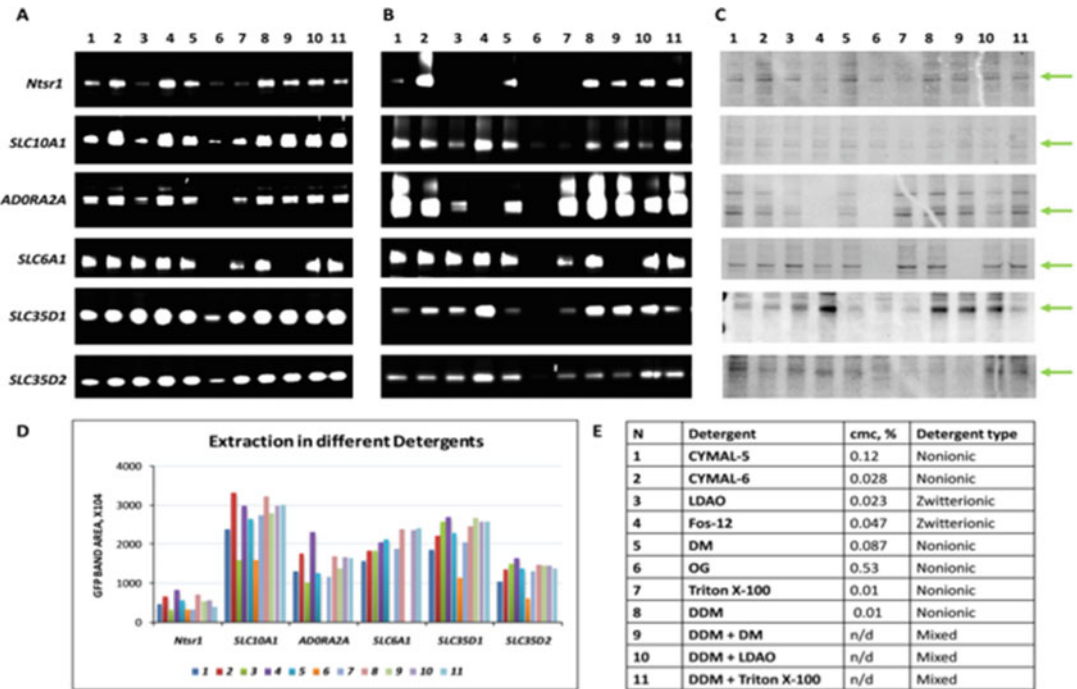


Fig. 7 Screening of detergents for protein solubilization and purification. **(a)** In-Gel GFP fluorescence signals of target proteins solubilized in different detergents (20× cmc). **(b)** In-Gel GFP fluorescence signals and **(c)** Coomassie-stained protein bands of target proteins purified in different detergents via Ni-NTA beads. **(d)** Plotted GFP intensities of targets extracted in different detergents. Despite CYMAL-6 did not provide the best solubilization for most of the protein targets, CYMAL-6 was the only detergent that extracted all targets with good efficiency. Mixed micelles provided very good solubilization for most of the non-GPCR targets. Most of the detergents that provided good solubilization for targets were also good for protein purification via Ni-NTA. **(e)** Example of detergent screen composition used in our studies

9. Use imager integrated software to measure the area of GFP-fused protein bands in each sample lane of the gel. Plot data and compare extraction efficiencies of different detergents (Fig. 7d).
10. In our screen, we compared the extraction efficiency of the most commonly used detergent, DDM, to several single and mixed detergents (Fig. 7d).

3.11 Purification of His-Tagged Targets from Membranes in Different Detergents on a Small Scale

1. Pre-equilibrate Ni-NTA agarose (100μL resin for each probe) in Buffer 3.
2. Apply solubilizate from the previous step on Ni-NTA agarose.
3. Bind proteins O/N at 4 °C using a vertical rotating platform.
4. Do all subsequent purification steps in a batch mode on ice or in a cold room.
5. After binding sediment resin by gravity flow and discard the supernatant.

6. Wash resin three times in 1 mL Buffer 3 and one time in 1 mL Buffer4.
7. For elution, add 150 μ L of Buffer 5 and incubate resin with gentle agitation 1 h before collecting elution.
8. Load purified samples on (i) NuPAGE to assess protein purity and on (ii) FSEC column to assess homogeneity.
9. With the best detergent do large-scale extraction and purification.

**3.12 Purification
of Strep-Tagged
Targets from
Membranes
in Different Detergents
on a Small Scale**

1. Aspirate required volume of MgStrep beads suspension (100 μ L 5% suspension = 10 μ L beads and is used for each 1 mL probe).
2. Separate MgStrep beads on Magnetic Separator and discard storage solution.
3. Equilibrate MgStrep beads suspension in 1 mL of Buffer6.
4. Apply solubilize from the previous step on pre-equilibrated MgStrep beads. For more than three samples use 24-well deep-well block.
5. Shake plate at 400 rpm for ≥ 2 h. For better results, leave suspension for overnight binding in a cold room.
6. Place a plate on Magnetic Separator and discard the supernatant. Rinse beads two times in 1 mL of Buffer6.
7. For elution apply 60 μ L of Buffer7 and shake plate 1 h before collecting the samples.
10. Load purified samples on (i) NuPAGE to assess protein purity and on (ii) FSEC column to assess homogeneity.
8. With the best detergent do large-scale extraction and purification.

**3.13 Quality Control
of His- and Strep-Tag
Purified
Samples: FSEC**

1. Monitor the monodispersity and stability of the purified target proteins in different detergents/buffers using FSEC (Fig. 8).
2. Centrifuge His-tag or Strep-tag purified samples (12,500 g , 5 min, 4 $^{\circ}$ C) and transfer 20–110 μ L probes in 0.3 mL insert vials with a rubber closure. The injection of samples (10–100 μ L) on SRT-C-300 HPLC system column (20 mL) can be done automatically using a high-throughput auto-sampler.
3. Run samples in Buffer8 at 0.5 mL/min flowrate.
4. Record both, GFP and tryptophan fluorescence.
5. Analyze FSEC traces in terms of peak area, elution profile, and volume to get information on (i) expression level, (ii) the degree of monodispersity, and (iii) the approximate molecular mass.

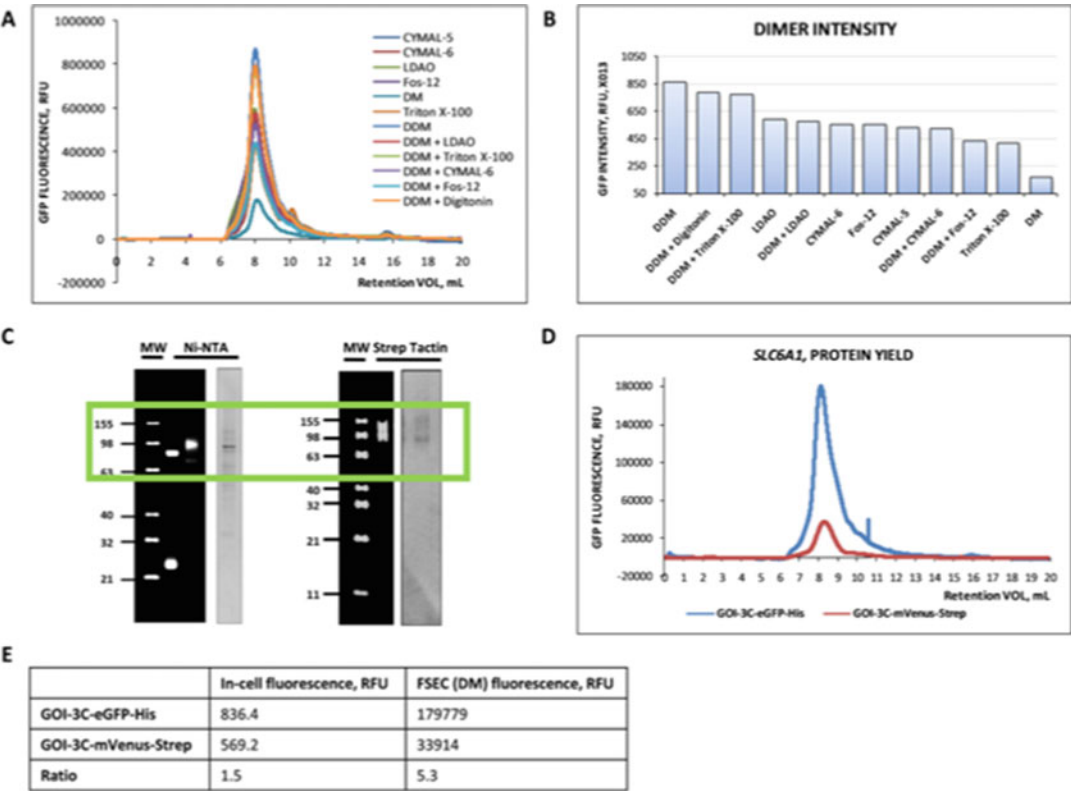


Fig. 8 FSEC analysis of protein stability in different detergents. **(a)** FSEC analysis of *SLC6A1* target purified in different detergents via Ni-NTA beads. Few detergents and detergent mixtures from our list provide good extraction and maintain protein stable across purification, including commonly used for membrane protein DDM + Digitonin mixture. **(b)** Comparison of intensities of probes obtained in different detergents. **(c)** Representative In-gel fluorescence and Coomassie staining of *SCL6A1* samples purified via Ni-NTA and Strep Tactin beads indicate the high quality of samples at the end of the purification. **(d)** According to FSEC traces, both constructs (GFP-His and mVenus-Strep C-terminal fusions) are stable dimers at the end of the purification. **(e)** The yield of GFP-His and mVenus-Strep-fused target differs: While GFP-His-fused target is better expressed and more protein is obtained at the end of the purification, the mVenus-Strep fusion provides higher specificity for binding

6. Monodisperse and folded proteins will yield a single symmetrical peak and polydisperse, unstable, or unfolded proteins will yield multiple asymmetric peaks.

3.14 Purification from Small-Scale Expression Tests (3 mL/6-Well Plate)

1. To test protein construct variants (e.g., His, Strep, or other tags fusions) run small-scale purifications using test plate expressions.
2. Expression volumes such as 3 mL (6-well plate experiment) can be used.
3. Spin down cells in 15 mL Falcon tubes (3000 *g*, 10 min, 4 °C).

4. Discard the supernatant and suspend cell pellets in 2 mL of ice-cold Buffer2 (suitable for His-tagged targets) or Buffer6 (suitable for Strep-tagged targets).
5. Transfer cell suspensions in 24-well deep-well block.
6. Use 24 Tip Horn for the sonicator to process numerous samples simultaneously.
7. Supplement broken cells with 1% DDM (or any other detergent of choice) and solubilize 1 h at a vertical rotator in a cold room.
8. Run short centrifugation (12,500 *g*, 10 min, 4 °C).
9. **Optional:** To concentrate sample for SDS-PAGE analysis and to reduce the detergent concentration before affinity purification, do short ammonium sulfate precipitation of probes. To do so, (i) collect supernatant after low spin centrifugation and measure its precise volume, (ii) add slowly an equal volume of saturated ammonium sulfate solution and mix 2–3 min at RT, (iii) for better precipitation leave on the bench for another 5 min; (iv) spin down 20 min at 12,500 *g*. Pellet will contain target protein and can be re-suspended in $\geq 150\mu\text{L}$ of Buffer3 or Buffer6.
10. Load sample aliquots from **steps 7 and 8** on SDS-PAGE and analyze in-gel GFP fluorescence signals after run is completed (Fig. 9).
11. With remaining supernatant do purifications using magnetic beads as it is described above in Subsections 3.11 and 3.12.

3.15 Large-Scale Affinity Purification

1. Use the best detergent to run large-scale extraction.
2. Large-scale purification must include overnight on-column removal of recombinants fusion and reverse IMAC.

4 Notes

1. Multi-target pOPIN vectors carry lacZ gene upstream of the reading frame allowing blue-white selection in *E. coli*.
2. A suitable mammalian expression vector with an appropriate expression promoter and translational signal (minimal (ACCATG) or full (GCCACCATG) Kozak consensus sequence) should be used for this protocol.
3. Our protocol is suitable for any scale of expression: 1–3 mL plate experiments and 30–300 mL up-scaled expression in flasks. Scale provided volumes and quantities of reagents proportionally to the used volume of transfected cells.

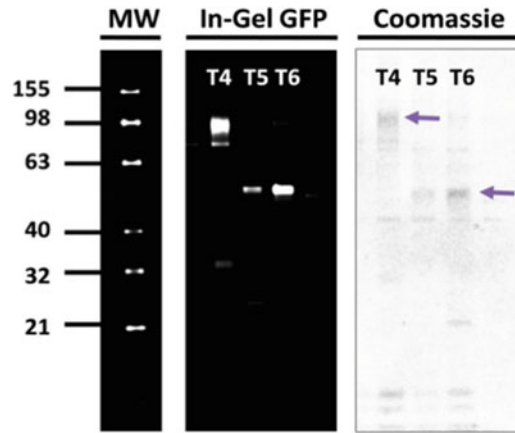


Fig. 9 Small-scale analytical protein purification. In-gel GFP fluorescence signals and Coomassie staining of samples purified using cell pellets from 3 mL/6-well plate test expressions. In this set of experiments, broken cells were solubilized in DDM + LDAO mixture, precipitated with 50% AS and re-suspended material was used for affinity purification on MagStrep beads. Full-length proteins were detected for all three human MPs targets: T4 (*SLC6A1*), T5 (*SLC35D1*), and T6 (*SLC35D2*)

4. Do not use high-density cells ($>5-6 \times 10^6$) for routing sub-culturing as it may reduce protein titer.
5. Higher starting cell density is essential, as at 30 °C the proliferation of cells will be reduced.
6. Do not mix DNA and PEI directly as they will precipitate immediately.
7. Valproic acid and sodium propionate are known as histone deacetylase inhibitors (iHDACs) [39, 40] and are used to cope with transcriptional repression of transfected plasmids. The use of valproic acid, sodium propionate, and glucose feed in combination helps substantially enhance gene expression.

Acknowledgments

This work was funded by Instruct-ULTRA (Coordination and Support Action Number ID 731005) funded by the EU H2020 framework to further develop the services of Instruct-ERIC and the Wellcome Trust Grants 202892/Z/16/Z (Membrane Protein Laboratory) and 090532/Z/09/Z (Wellcome Human Genetics Centre).

References

1. Lyumkis D (2019) Challenges and opportunities in cryo-EM single-particle analysis. *J Biol Chem* 294:5181–5197
2. Rivera-Calzada A, Carroni M (2019) Editorial: technical advances in cryo-electron microscopy. *Front Mol Biosci* 6:72
3. Cuozzo JW, Soutter HH (2014) Overview of recent progress in protein-expression technologies for small-molecule screening. *J Biomol Screen* 19:1000–1013
4. Wallin E, von Heijne G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* 7:1029–1038
5. Almen MS, Nordstrom Fredriksson R et al (2009) Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol* 7:50
6. Rask-Andersen M, Masuram S, Schioth HB (2014) The druggable genome: evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annu Rev Pharmacol Toxicol* 54:9–26
7. Yin H, Flynn AD (2016) Drugging membrane protein interactions. *Annu Rev Biomed Eng* 18:51–76
8. Santos R, Ursu O, Gaulton A et al (2017) A comprehensive map of molecular drug targets. *Nat Rev Drug Discov* 16:19–34
9. Kawate T, Gouaux E (2006) Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* 14:673–681
10. Lee CH, MacKinnon R (2017) Structures of the human HCN1 hyperpolarization-activated channel. *Cell* 168(111–120):e11
11. Noreng S, Bharadwaj A, Posert R et al (2018) Structure of the human epithelial sodium channel by cryo-electron microscopy. *elife* 2018(7): e39340
12. Singh AK, Saotome K, Luke L et al (2018) Structural bases of TRP channel TRPV6 allosteric modulation by 2-APB. *Nat Commun* 9:2465
13. Hunter M, Yuan P, Vavilala D et al (2019) Optimization of protein expression in mammalian cells. *Curr Protoc Protein Sci* 9:e77
14. Estes S, Melville M (2014) Mammalian cell line developments in speed and efficiency. *Adv Biochem Eng Biotechnol* 139:11–33
15. Walsh G (2014) Biopharmaceutical benchmarks. *Nat Biotechnol* 32:992–1000
16. Lu P, Bai X-C, Ma D et al (2014) Three-dimensional structure of human gamma-secretase. *Nature* 512:166–170
17. Miller PS, Aricescu AR (2014) Crystal structure of a human GABAA receptor. *Nature* 512:270–275
18. Su Q, Hu F, Ge X et al (2018) Structure of the human PKD1-PKD2 complex. *Science* 361: eaat9819
19. Li X, Wang J, Coutavas E et al (2016) Structure of human Niemann-Pick C1 protein. *Proc Natl Acad Sci U S A* 113:8212–8217
20. Inoue M, Sakuta N, Watanabe S et al (2019) Structural basis of sarco/endoplasmic reticulum Ca(2+)-ATPase 2b regulation via transmembrane helix interplay. *Cell Res* 27:1221–1230. e3
21. Fraley R, Subramani P, Berg P et al (1980) Introduction of liposome-encapsulated SV40 DNA into cells. *J Biol Chem* 255 (21):10431–10435
22. Boussif O, Lezoualc'h F, Znata MA et al (1995) A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine. *Proc Natl Acad Sci U S A* 92:7297–7301
23. Longo PA, Kavran JM, Kim M-S et al (2013) Transient mammalian cell transfection with polyethylenimine (PEI). *Methods Enzymol* 529:227–240
24. Goehring A, Lee C-H, Wang KH et al (2014) Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. *Nat Protoc* 9:2574–2585
25. Reeves PJ, Callewaert N, Contreras R et al (2002) Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proc Natl Acad Sci U S A* 99:13419–13424
26. Tom R, Bisson L, Durocher Y (2008) Transfection of adherent HEK293-EBNA1 cells in a six-well plate with branched PEI for production of recombinant proteins. *CSH Protoc* pdb prot4978
27. Nettleship JE, Watson PJ, Rhaman-Huq N et al (2015) Transient expression in HEK 293 cells: an alternative to *E. coli* for the production of secreted and intracellular mammalian proteins. *Methods Mol Biol* 1258:209–222
28. Chaudhary S, Pak JE, Pedersen BP et al (2011) Efficient expression screening of human membrane proteins in transiently transfected Human Embryonic Kidney 293S cells. *Methods* 55:273–280

29. Ooi A, Wong A, Esau L et al (2016) A guide to transient expression of membrane proteins in HEK-293 cells for functional characterization. *Front Physiol* 7:300
30. Portolano N, Watson PJ, Firall L et al (2014) Recombinant protein expression for structural biology in HEK 293F suspension cells: a novel and accessible approach. *J Vis Exp* 92:e51897
31. Subedi GP, Watson RW, Moniz H et al (2015) High yield expression of recombinant human proteins with the transient transfection of HEK293 cells in suspension. *J Vis Exp* 106: e53568
32. Elegheert J, Behiels E, Bishop B et al (2018) Lentiviral transduction of mammalian cells for fast, scalable and high-level production of soluble and membrane proteins. *Nat Protoc* 13:2991–3017
33. Pan X, Li Z, Huang X et al (2019) Molecular basis for pore blockade of human Na(+) channel Nav1.2 by the mu-conotoxin KIIIA. *Science* 363:1309–1313
34. Yang G, Zhou R, Zhou Q et al (2019) Structural basis of Notch recognition by human gamma-secretase. *Nature* 565:192–197
35. Zhou R, Yang G, Guo X et al (2019) Recognition of the amyloid precursor protein by human gamma-secretase. *Science* 363: eaaw0930
36. Alam A, Kung R, Kowal J et al (2018) Structure of a zosuquidar and UIC2-bound human-mouse chimeric ABCB1. *Proc Natl Acad Sci U S A* 115:E1973–E1982
37. Alam A, Kowal J, Broude E et al (2019) Structural insight into substrate and inhibitor discrimination by human P-glycoprotein. *Science* 363:753–756
38. Sambrook JF, Russell DW (2001) Molecular cloning: a laboratory manual, 3rd ed, Vols 1, 2 and 3, Cold Spring Harbor Laboratory Press. 1, 2 and 3: p 1–2100
39. Fan S, Maguire CA, Ramirez SH et al (2005) Valproic acid enhances gene expression from viral gene transfer vectors. *J Virol Methods* 125:23–33
40. Lin MY, de Zeote MR, van Putten JP et al (2015) Redirection of epithelial immune responses by short-chain fatty acids through inhibition of histone deacetylases. *Front Immunol* 6:554



Chapter 6

Reproducible and Easy Production of Mammalian Proteins by Transient Gene Expression in High Five Insect Cells

Maren Schubert, Manfred Nimtz, Federico Bertoglio, Stefan Schmelz, Peer Lukat, and Joop van den Heuvel

Abstract

The expression of mammalian recombinant proteins in insect cell lines using transient-plasmid-based gene expression enables the production of high-quality protein samples. Here, the procedure for virus-free transient gene expression (TGE) in High Five insect cells is described in detail. The parameters that determine the efficiency and reproducibility of the method are presented in a robust protocol for easy implementation and set-up of the method. The applicability of the TGE method in High Five cells for proteomic, structural, and functional analysis of the expressed proteins is shown.

Key words Transient gene expression, TGE, High five, Insect cells, Expression vector

1 Introduction

High-quality protein samples are essential for structural, proteomic, and functional analysis of biological processes [1–3]. Especially, the current 2019-CoV pandemic shows the importance of reliable recombinant expression systems that are able to produce ample amounts of correctly folded viral and host proteins. These proteins may be used as tools in diagnostic screening, establishing assays for drug-screening, vaccinology, structural analysis at atomic level using crystallization or cryo-EM [4]. For functional biologic studies, it is essential to produce these proteins in their native state. Therefore, the choice of the appropriate expression system is of upmost importance [5, 6].

Many viral and mammalian proteins required for host–pathogen interaction studies depend on specific post-translational modifications to be biologically active [5, 6]. Others form multimers or assemble as part of multi-protein complexes for full functionality. Proper assembly and folding of the target proteins is only possible using sophisticated eukaryotic expression systems (yeast, insect,

mammalian, and plant), which all have their specific advantages and disadvantages [1].

Recombinant protein expression requires a template for the target gene, an expression vector and a suitable method to introduce and maintain the recombinant expression vector in the producer cell line [5]. Most of the available expression vectors require substantial (re)cloning of the desired target gene into individual vectors, each specific for a particular host system or versatile vectors which can be used in multiple expression systems [2]. The current cloning procedures like, e.g., Golden Gate [7], Molecular Cloning [8], and “SLIC-Fusion” [9] are highly efficient. In combination with the available commercial synthesis of custom-made genes, there are almost no limitations to generate required expression constructs [5]. Therefore, the bottleneck has changed from cloning to fast expression and screening systems.

In industry, the requirement for a GMP controlled process and an optimized yield both determine the choice of stable cell lines as the expression strategy. However, in contrast to transient expression systems, development of stable cell lines is very time-consuming and cost-intensive, which is not suited for high-throughput expression analysis [2, 5]. The viral and plasmid-based transient expression systems in mammalian HEK cells or High Five insect cells (Hi5 cells) both allow scale-down to 2 mL cultures for automation and high-throughput screening [10, 11]. This is essential to develop the initial optimal construct for expression.

The transient gene expression (TGE) in either HEK293-6E or Expi293F cell lines is well established but requires a license and/or expensive transfection materials as well as specific growth media, which make this system expensive and difficult to implement. Since 2015, virus-free transient gene expression in Hi5 cells was optimized and improved substantially [11–15]. This system uses affordable media, is easy to establish, robust and reproducible in performance. Growth is possible in simple incubators without the need of special CO₂ aeration. In comparison with HEK293 and CHO TGE systems, we have shown for many tested proteins that the yield in the Hi5 TGE was more than sufficient to provide the required amounts of high-quality protein.

In this chapter, we present an optimized and robust TGE protocol for Hi5 insect cells exemplified by the production and characterization of the S1 fragment of the SARS-CoV-2 Spike surface protein. The TGE method is applicable to both single molecule expression from a single plasmid and multi-protein complex expression using a large set of vectors in parallel. TGE in Hi5 cells is especially suited for fast, inexpensive, and simple screening using multi-well or chamber bioreactors, as well as for large-scale production of recombinant mammalian proteins in shake flasks or bioreactors. In conclusion, plasmid-based transient expression in

Hi5 insect cells simplifies eukaryotic protein expression to a point where it is superior to using prokaryotic systems [2].

2 Materials

2.1 Expression Vectors and Cell Lines

1. pOpIE2-C series (C-terminal tagged) and pOpIE2 N series (N-terminal tags) are available from the authors on request.
2. The High Five (Hi5) insect cell line (officially called BTI-Tn-5B1-4) was isolated by the *Boyce Thompson Institute* for Plant Research, Ithaca, USA. The cell line can be acquired from Thermo Fisher Scientific (*see Note 1*).

2.2 Cell Culture

1. 125 mL up to 5 L vented polycarbonate shake flasks (Corning).
2. Orbitron™ platform shaker with 50 mm orbit (Infors) in a 27 °C climatized room with 50% humidity (*see Note 2*).
3. Complete Cultivation Medium for Hi5 insect cells: EX-CELL 405 (Sigma) (*see Note 3*).

2.3 Transfection Reagents and Additional Chemicals for Protein Production

Prepare all solutions using ultrapure water and analytical grade reagents. All reagents will be sterile filtered and stored at 4 °C (unless otherwise mentioned).

1. 1 mg/mL Polyethylenimine 40 kDa, linear (Polysciences): Dissolve 0.05 g of PEI in 50 mL of MilliQ pH 7 (*see Note 4*).
2. Expression plasmid (*see Note 5*) best at a concentration of 500–1000 ng/μL highly pure in MilliQ or TE buffer.

3 Methods

3.1 Construction of Expression Vectors

The efficiency of TGE depends on a highly active promoter that can promote transcription by the RNA polymerase II. The immediate early promoter OpIE2 from the baculovirus *Orgyia pseudotugata* was identified as the currently strongest promoter of this type in Hi5 insect cells [10, 11]. This promoter was cloned into the backbone of pIEX/bac5. We generated a series of variants of this expression vector pOpIE2-C1-C5 for easy fusion of the gene of interest (GOI) to purification tags. An overview of the constructs is shown in Fig. 1. We preferably use the restriction sites: NheI, SpeI, XbaI, and AvrII, which all have the same overhang of nucleotides for generating C-terminal fusion. This allows easy recombineering of the cassettes into new variants of the available elements as well as integrating synthetic genes with individual preferred tag sequences. For example, version C5 was generated from C2 by removal of the GFP11 β-strand sequence by simple digestion with XbaI and AvrII followed by ligation of the vector (*see Note 6*). The same strategy is

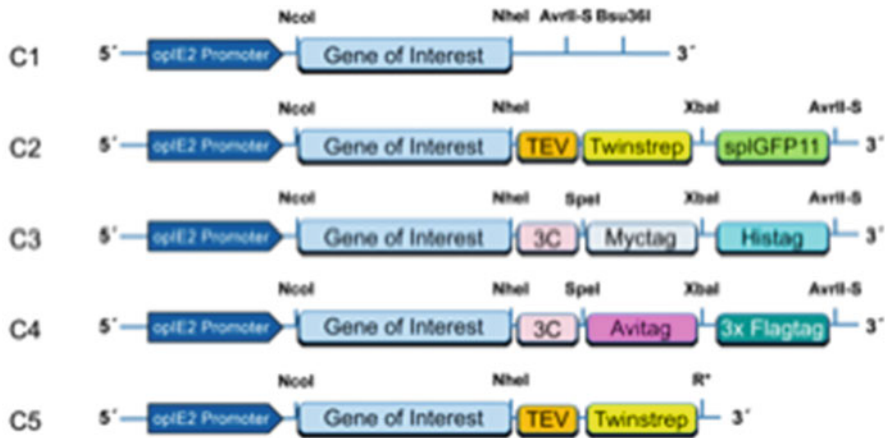


Fig. 1 Schematic representation of the pOpIE2-C-series of constructs. The C represent the series of C-terminal tags which are cloned into the backbone and can be directly chosen to generate the required fusion construct. Optional constructs containing two possible protease sites (TEV and Rhinovirus 3C) in line with three optional tag combinations having either of the purification tag sequences (Twinstrep, Histag, Avitag, and Flagtag) are available. For antibody detection by western blot, two variants carry the sensitive detection sequences (MycTag, Flagtag). The specific *GFP11* B-stand can be used as part of the split*GFP* detection system (GFP fluorescence) [11]. The Avitag can be specifically labeled with biotin using the BIR ligase. This allows immobilization of the GOI to streptavidin linked materials (Western blotting, SPR, BLI, and other biophysical analytic techniques)

also applied for N-terminal fusion constructs (pOpIE2-N-series, data not shown). For cloning of the N-terminal sequence of the GOI, we use the NcoI site that contains the AUG start codon as well as an upstream BamHI site for easy cloning into the vector. Additionally, seamless new fusion tags can be cloned using Golden Gate cloning or SLIC-fusion method.

3.2 Transient Expression in Hi5 Cells

The described protocol is for a 120 mL expression in Hi5 cells. The employed vectors have an important impact on the yield. The scale of the transfection experiment can be easily adapted by linear decrease of the components to a final volume of 2 ml as well as increasing to large scale (2 L).

1. [Day -3/-2, e.g., Friday] Prepare a 40 mL culture containing 0.4×10^6 c/mL 72 h prior transfection or prepare a 40 mL culture containing 0.5×10^6 c/mL 48 h before transfection and incubate the culture 72 h at 27 °C and 90 rpm (see Note 7).
2. [Day 0, e.g., Monday] Count the cells (viability should be above 95%) and prepare a 30 mL culture containing 4×10^6 c/mL by centrifuging the required volume of the cell suspension at $180 \times g$ for 4 min. Discard the supernatant and resolve the cell pellet in 30 mL fresh EX-CELL 405.
3. Pipette 120 µg of your DNA directly to the prepared cells and mix gently (see Note 8).

4. Immediately pipette 480 μL PEI of the 1 mg/mL 40 kDa PEI stock solution to the cells and mix gently.
5. Incubate the culture at 27 °C and 90 rpm for 4 up to 20 h (*see Note 9*).
6. Add 90 mL fresh EX-CELL 405 media.
7. [**Day 2, e.g., Tuesday**] 48 h after transfection feed the cells with 120 mL fresh EX-CELL 405 media.
8. [**Day 3–5, e.g., Thursday**] Take samples daily, count the cells, and determine transfection efficiency in the cytometer or/and determine target protein expression by a suitable technique (SDS-PAGE, slot blot or western blot). If viability of the cells or quality of the recombinant protein starts to drop, harvest the culture.
9. For intracellular proteins, carefully centrifuge the cells at 180 (up to max. 500) $\times g$ for 4 (up to max. 10) min and freeze the cell pellet at -20 °C until cell lysis and purification. Secreted target proteins are first centrifuged at 180 $\times g$ for 4 min, followed by a centrifugation of the supernatant at 2000 $\times g$ for 20 min. Afterwards, the supernatant is filtered with 0.2 μm filters and stored at 4 °C until purification (*see Note 10*).

3.3 Production of the SARS-CoV-2 S1-Opt-delFurin-hFc

The example presented here is the S1 fragment of the SARS-CoV-2 Spike surface protein. It represents an antigen that can be used as optional candidate for generation of vaccines and a protein highly relevant to get insight into the viral infection mechanism. Viral surface proteins often have higher order structures (homo- or hetero-multimers) and are substantially glycosylated. Here we show the result of the protein purification, binding activity and the subsequent analysis of the glycosylation that shows to be specific for lepidopteran insect cells.

3.3.1 Expression and Purification of S1-Opt-delFurin-hFc

The synthetic gene for S1-Opt-delFurin-hFc was codon optimized for mammalian expression and designed according to Wrapp et al. [4]. The sequence was fused to an hFc tag or His tag for easy purification by protein A/Ni-NTA chromatography. The expression of the protein S1-Opt-delFurin-hFc was done for up to 72–96 h. The transfection efficiency determined by the fraction of GFP fluorescent cells reached up to 60% at a vitality of 98%. The supernatant of the hFc tagged protein was purified on a 1 ml rProtA Hitrap column using the standard protocol of the supplier (Cytiva). The eluted S1-Opt-delFurin-hFc protein was pooled and concentrated before loading on a HiLoad 16/600 Superdex 200 pg column (Cytiva). A homogenous peak was isolated after size-exclusion chromatography using TBS (20 mM Tris pH 7.4, 150 mM NaCl) as equilibration buffer. The eluted fractions were analyzed by SDS-PAGE (Fig. 2). Homogeneous samples were pooled and concentrated to 1 mg/ml and stored at -80 °C after snap freezing.



Fig. 2 SDS-PAGE analysis of the samples purified by size-exclusion chromatography. Samples of fractions separated by SEC on a HiLoad 16/600 Superdex 200 pg were analyzed on a Biorad Any kD Gel using denaturing sample buffer. The gel was stained with Instant Blue. Lane 1 Pageruler Plus prestained Molecular Weight Standard (Thermo Fisher Scientific), Lane 2–10 represent successive fractions separated by SEC. The S1-Opt-delFurin-hFc with an estimated size of 130 kD is indicated by the arrow

This sample was further analyzed by SEC-MALS to determine the conformation of the protein and the glycan content of the protein sample. The S1-Opt-delFur-His tagged protein used for ELISA analysis was purified in a similar way using a 1 ml HisTrap Excel column, followed by SEC using PBS (10 mM phosphate pH 7.4, 2.7 mM KCl and 135 mM NaCl) as equilibration buffer.

3.3.2 SEC-MALS Analysis of the Glycosylated S1-Opt-delFurin-hFc Protein

The protein conformation and amount of glycosylation was analyzed by analytical size-exclusion chromatography in combination with multi-angle light scattering (SEC-MALS) [18] using an Agilent 1260 Infinity II system with an UV detector connected to a Wyatt TREOS II MALS detector and an Optilab 505-rEX refractive index (RI) detector. 100 µg of the S1-Opt-delFurin-hFc protein sample was separated on a Superdex 200 increase 10/300 column (Cytiva) with TBS running buffer (20 mM Tris pH 7.4, 150 mM NaCl, 0.1 µm microfiltered). The data were analyzed using the Protein Conjugate Analysis method with the ASTRA 7.3.2.19 software. The calculations are dependent on a good estimate of the refractive index increment (dn/dc) of the sample. For proteins, this value is near 0.185 mL/g. Glycan modifications have an index of 0.14–0.15 mL/g. The fractional mass of the protein and the glycan were deconvoluted using the signal of the UV detector and the RI detector simultaneously recorded to the MALS data.

SEC-MALS analysis of S1-hOpt-delFurin-hFc (Fig. 3) shows that the protein migrates as a homodimer with an overall molecular mass or 235.0 kDa. The protein content is 201.9 kDa with a glycan

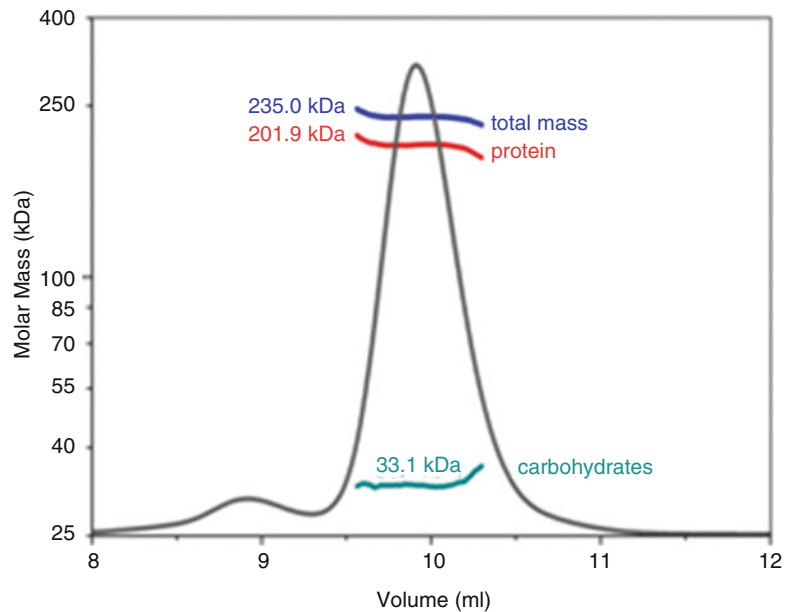


Fig. 3 SEC-MALS Analysis of the S1-Opt-delFurin-hFc protein. The SEC-MALS analysis of S1-Opt-delFurin-hFc shows a main peak fraction with an overall molecular mass or 235.0 kDa (blue line as determined from the RI detector). The protein content was 201.9 kDa (red line as determined from the UV signal) with a glycan composition of 33.1 kDa (carbohydrates). This correlates to a homodimer of the protein with a calculated mass of 101 kDa. The void volume of the column is 8 ml

composition of 33.1 kDa. This correlates to a homodimer of the protein with a calculated mass of 101 kDa. (Table 1b).

3.3.3 Enzyme-Linked Immuno Sorbent Assay (ELISA)

The major function of the SARS-CoV-2 Spike protein is the recognition and binding to the angiotensin-converting enzyme hACE2, priming the internalization of the virus into the human host cell. The functional binding activity of the purified S1-Opt-delFurin-His fragment of the SARS-CoV-2 spike protein to the extracellular domain of hACE2 was tested by ELISA (Fig. 4). Hereto, S1-Opt-delFurin-His was immobilized on a Costar high binding 96-well plate (200 ng/well, blocked with 2% skimmed milk powder in PBST (PBS 1× with 0.05% Tween20) followed by incubation with the indicated concentrations of its binding partner hACE2-mFc. ACE2-mFc binding was detected using goat-anti-mIgG(Fc)-HRP (1:42000, A0168, Sigma) antibody and visualized by tetramethylbenzidine (TMB) substrate. After stopping the reaction by addition of 1 N H₂SO₄, absorbance at 450 nm with a 620 nm reference was measured in an ELISA plate reader (Epoch, BioTek). EC₅₀ value was calculated using GraphPad Prism Version 6.1, fitting to a four-parameter logistic curve, resulting in an EC₅₀ value of 2.7 nM.

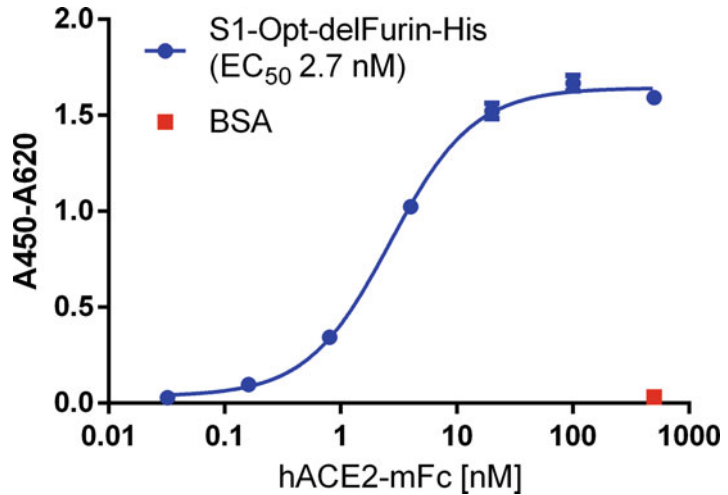


Fig. 4 ELISA analysis of the interaction of S1-Opt-delFurin-His with hACE2. The S1-Opt-delFurin-His-tagged protein was coated to ELISA plates and after blocking non-specific binding sites the wells were incubated with different concentrations of purified hACE2-mFc. The A_{450} and $A_{620\text{nm}}$ (reference) were measured after incubation with anti-mFc-conjugated with HRP and staining with the substrate (TMB). The calculated EC_{50} value was 2.7 nM

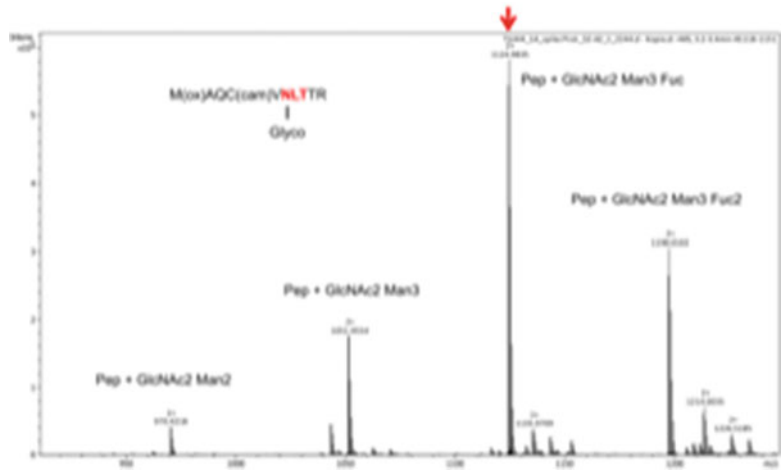


Fig. 5 Representative MS analysis of the glycosylation of isolated tryptic fragment of S1-Opt-delFurin-hFc. The isolated and characterized fragments all carry an equal distribution of three major glycans GlcNAc2 Man3 + GlcNAc2 Man3 Fuc + GlcNAc2 Man3 Fuc2 as shown in this example of the first N-terminal glycosylation site at position 6 of the sequence presented in Table 1b

3.4 Analysis of Glycosylation

The type of glycosylation was determined by mass spectroscopic analysis of peptides isolated from the protein bands after separation by SDS-PAGE and subsequent tryptic digestion as shown in Fig. 5. The (glyco-)peptides extracted after tryptic in-gel digestion of the

Table 1
Characterized glycosylation sites of S1-Opt-delFurin-hFc

A: Analyzed tryptic peptides

The isolated and characterized fragments all carry equal distribution of three major glycans as shown by the analysis presented in **Figure 5**: GlcNAc2 Man3 + GlcNAc2 Man3 Fuc + GlcNAc2 Man3 Fuc2.

Bold: modified amino acids; Bold and red: glycosylated amino acids

Site I: **1**MAQCV**N**LTTR₁₀

1MAQCV**N**LTTRTQLPPAYTNSFTR

Site VI: ₁₄₈VYSSAN**NCT**FEYVSQPFLMDLEGK₁₇₁

Site VII: ₂₀₄DLPQGFSALEPLVDLP**IGIN**ITR₂₂₆

Site VIII: ₂₆₈YNE**NG**ITITDAVDCALDPLSETK₂₈₉

₂₆₃TFLKKYNE**NG**ITITDAVDCALDPLSETK₂₈₉

Site XII: ₆₀₁YQDV**NCT**EVPAIHADQLTPTWR₆₂₃ Semitryptic peptide!

Site XIII: ₆₃₆AGCLIGAEHV**NN**₆₄₇ Semitryptic peptide!

Site XIV: ₇₄₇TKPREEQY**N**STYR₇₅₉

B: S1-Opt-delFurin-hFc protein sequence

The underlined amino sequence represents the tryptic fragment. Bold amino acids represent the glycosylation recognition site N-X-S/T. The glycosylation N-residue is marked in bold and red. The italic sequence represents the hFc tag.

MAQCV**N**LTTRTQLPPAYTNSFTRGVVYDPKVFSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFAST
EKSNIIRGWIIFGTTLDSTQSLIVNNATNVVIVKVEFQFCNDPFLGVYHKNNKSWMESEFRVYSSAN**NCT**FEYVSQPFLMDL
EGKQGNFKNLREFVFNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLP**IGIN**ITR**FQ**TLLALHRSYLTPGDSSSGWTAGAA
AAYVGYLQPR**TFL**LYNE**NG**ITITDAVDCALDPLSETKCTLSFTVEKGIYQTSNFRVQPTESIVRFNITNLCPFGEVFNATRF
ASVYAWNRKRISNCVADYSVLVNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVQRQIAPGQTGKIADYNYKLPPDFTG
CVIAWNSNNLDSKVGNGNYLYRLFRKSNLKPFPERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVLSF
ELLHAPATVCGPKKSTNLVKNKCVNFNGLTGTGVLTESNKKFLPFQGFGRDIADTTDAVRDPQ**TLE**ILDITPCSFGGVSVIT
PGTNTSNQVAVLYQDV**NCT**EVPAIHADQLTPTWRVYSTGNSVVFQTRAGCLIGAEHV**NN**SYECDIPIGAGICASYQTQTNSPGS
ASAAASDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQY**NS**
TVRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWE
SNGQPENNYKTTTPVLDSDGSFFLYSKLTVDKSRWQQGNVFCSCVMHEALHNHYTQKSLSLSPGK

relevant bands from gel bands stained with Instant Blue were analyzed on an Evosep LC system coupled to a tims/TOF Pro mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany).

The acquisition of mass and tandem mass spectra was done with an average resolution of 60,000. To enable the parallel accumulation-serial fragmentation (PASEF) method, precursor m/z and mobility information was first derived from full scan TIMS-MS experiments. Singly charged precursors were excluded by their position in the m/z-ion mobility plane. The collision energy for fragmentation varied between 31 and 52 eV depending on precursor mass and charge [19]. Protein identification was performed against the CoV-2 Spike protein sequence (S1-Opt-del-Fur-hFc, Table 1) using the Peaks 10.5 software (BSI, Toronto, Canada). Variable observed amino acid modifications were: oxidized methionine and the main N-glycans GlcNAc2Man3Fuc and GlcNAc2Man3Fuc2 typical paucimannose-type glycosylation for insect host cells. Additionally, carbamidomethylation of cysteine was selected as a fixed modification. Trypsin was selected as the proteolytic enzyme, with a maximum of two potential missed

cleavages. To ensure optimal identification of glycopeptides, the data were additionally searched manually for sets of precursors ions producing the N-glycan-specific fragment at m/z 204.0872 [GlcNAc+H] + upon high collision energy induced dissociation (HCD) (Fig. 4 and Table 1).

In total 6 out of 13 glycosylation sites could be identified compared to the analysis of the glycan shield of the CoV-2 spike protein by Watanabe et al. [17]. The glycosylation in HEK293 shows three different types of glycosylation from oligomannose, hybrid to complex glycosylation. In contrast, all glycopeptides analyzed from the S1 protein expression in Hi5 insect cells have a comparable distribution with mainly mono or bi-fucosylated GlcNAc2-Man3 representing the typical paucimannose type of glycosylation of lepidopteran cells. This type of glycosylation is advantageous for crystallographic structural analysis of membrane-bound or secreted mammalian and viral proteins.

4 Notes

1. Important to note is that it seems to improve the yield if the cells have been passaged in the EX-CELL 405 medium over a hundred times [14].
2. Cultivation with humidity is preferred but not absolute necessary. Smaller volumes than 15 mL are best cultivated in 50 mL TPP TubeSpin bioreactor vented tubes (Merck) at a higher shaking speed (120 rpm).
3. Other media might not work for the described method and inhibit the transient plasmid transfection [14].
4. Linear 40 kDa PEI proved to be more reliable than linear 25 kDa PEI. It is soluble in water and can be stored at 4 °C.
5. The optimal expression plasmid in our facility comprises the OpIE2 promoter, the IE1 terminator, and a FlashBac compatible backbone. The expression cassette is flanked by baculoviral sequences (orf 603 and orf 1629) which can be used for integration into the baculovirus from the Flashbac system [16]. These sequences are used to enhance transient gene expression [13].
6. Re-ligation will result in removal of both the XbaI and SpeI site. This results in a final construct having just the TEV protease site and the Twinstrep affinity tag. This cloning strategy requires that the synthetic genes will be designed without further internal NheI, SpeI, XbaI, and AvrII sites. The optimal expression plasmid in our facility comprises the OpIE2 promoter, the IE1 terminator, and a FlashBac compatible backbone. The expression cassette is flanked by baculoviral

sequences (orf 603 and orf 1629) that can be used for integration into the baculovirus from the Flashbac system [16]. These sequences are used to enhance transient gene expression [13].

7. This step ensures that the cells are in the optimal growth phase. Overgrown cells ($\sim 5\text{--}6 \times 10^6$ c/mL) or cells not yet in the exponential growth phase (e.g., passaged only a few hours before) do not reach maximum transfection yields.
8. The DNA concentration should be in a range of 0.2–1.0 $\mu\text{g}/\mu\text{L}$. Replacing 5% of the total DNA amount with a control plasmid expressing, e.g., eGFP will help to monitor transfection efficacy. The overall yield is only slightly affected by eGFP-expression. Additionally, one can perform co-expression of a multi-protein complex by distributing the amount of DNA among the expression vectors of the individual subunits.
9. Incubating the cells at high density for 4 h ensures higher transfection rates. Feeding after 20 h leads to a decreased viability and thereby low transfection rate, as the cells will suffer from depletion of medium components by that time.
10. The two-step centrifugation ensures that the cells are removed and not disrupted, preventing to get a lot of intracellular protein contaminating the supernatant. For His-Tag purification, it is important to add 0.5 M NaCl to the supernatant to prevent unspecific binding. His TrapTM Excel resin material (Cytiva) can be used to purify His-tag proteins directly from the supernatant. Other resin requires a re-buffering to other buffer and/or adjustment of the pH as EX-CELL 405 media has a pH of 6.0–6.4.

References

1. Meyer S, Lorenz C, Baser B et al (2013) Multi-host expression system for recombinant production of challenging proteins. *PLoS One* 8: e68674
2. Karste K, Bleckmann M, van den Heuvel J (2017) Not limited to *E. coli*: versatile expression vectors for mammalian protein expression. *Methods Mol Biol Clifton NJ* 1586: 313–324
3. Braun P, LaBaer J (2003) High throughput protein production for functional proteomics. *Trends Biotechnol* 21:383–388
4. Wrapp D, Wang N, Corbet KS et al (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367:1260–1263
5. Hunter M, Yuan P, Vavilala D et al (2018) Optimization of protein expression in mammalian cells. *Curr Protoc Protein Sci* 95:e77
6. Durocher Y, Butler M (2009) Expression systems for therapeutic glycoprotein production. *Curr Opin Biotechnol* 20:700–707
7. Maillonnet S, Grützner R (2020) Synthetic DNA assembly using golden gate cloning and the hierarchical modular cloning pipeline. *Current Protocols in Mol Biol* 130:e115
8. Weber E, Birkenfeld J, Franz J et al (2017) Modular Protein Expression Toolbox (MoPET) a standardized assembly system for defined expression constructs and expression optimization libraries. *PLoS One* 12: e0176314
9. Scholz J, Besir H, Strasser C et al (2013) A new method to customize protein expression vectors for fast, efficient and background free parallel cloning. *BMC Biotechnol* 13:12
10. Bleckmann M, Fritz MH-Y, Bhujji S et al (2015) Genomic analysis and isolation of RNA polymerase II dependent promoters

- from *Spodoptera frugiperda*. PLoS One 10: e0132898
11. Bleckmann M, Schmelz S, Schinkowski C et al (2016) Fast plasmid-based protein expression analysis in insect cells using an automated SplitGFP screen. Biotechnol Bioeng 113:1975–1983
12. Shen X, Pitol A, Bachmann V et al (2015) A simple plasmid-based transient gene expression method using High Five cells. J Biotechnol 216:67–75
13. Puente-Massaguer E, Lecina M, Gòdia F (2018) Nanoscale characterization coupled to multi-parametric optimization of Hi5 cell transient gene expression. Appl Microbiol Biotechnol 103:10495–10510
14. Bleckmann M, Schuerig M, Endres M et al (2019) Identifying parameters to improve the reproducibility of transient gene expression in High Five cells. PLoS One 14:e0217878
15. Puente-Massaguer E, Strobl G, Grabherr R et al (2020) PEI-Mediated transient transfection of High Five cells at bioreactor scale for HIV-1 VLP production. Nano 10:1580
16. Hitchman RB, Posse RD, King LA (2012) High-throughput baculovirus expression in insect cell. Methods Mol Biol Clifton NJ 824:609–627
17. Watanabe Y, Allen JD, Wrapp D et al (2020) Site-specific glycan analysis of the SARS-CoV-2 spike. Science 369:330–333
18. Some D, Amartely H, Tsadok A et al (2019) Characterization of proteins by size-exclusion chromatography couples to multi-angle light scattering (SEC-MALS). J Vis Exp 148. <https://doi.org/10.3791/59615>
19. Roth G, Vanz AL, Lünsdorf H et al (2018) Fate of the UPR marker protein Kar2/Bip and autophagic processes in fed-batch cultures of secretory insulin precursor producing *Pichia pastoris*. Microb Cell Factories 17:123



SynBac: Enhanced Baculovirus Genomes by Iterative Recombineering

Hannah Crocker, Barbara Gorda, Martin Pelosse,
Deepak Balaji Thimiri Govinda Raj, and Imre Berger

Abstract

Baculovirus expression vector systems (BEVS) are widely used to produce heterologous proteins for a wide range of applications. Developed more than 30 years ago, BEVS have been constantly modified to improve product quality and ease-of-use. Plasmid reagents were tailored and engineered to facilitate introduction of heterologous genes into baculoviral genomes. At the same time, detrimental modalities such as genes encoding proteases or apoptotic factors were removed to improve protein yield. Advances in DNA synthesis and manipulation now enable the engineering of part or whole synthetic baculovirus genomes, opening up new avenues to redesign and tailor the system to specific applications. Here, we describe a simple protocol for designing and constructing baculovirus genomes comprising segments of synthetic DNA through the use of iterative Red/ET homologous recombination reactions.

Key words *Autographa californica* multiple nucleopolyhedrosis virus, AcMNPV, Baculovirus expression vector system, Red/ET homologous recombination, Genome engineering

1 Introduction

The baculovirus expression vector system (BEVS) is a time-tested technology to produce heterologous proteins at high yields [1–3]. The use of eukaryotic insect cells as a host often affords authentic translational modifications, which can be instrumental for subsequent investigations of mammalian, notably human, proteins [3–7]. Initially, insertion of heterologous genes in the baculoviral genome relied on in vivo homologous recombination in insect cells co-transfected with purified baculoviral genomic DNA and transfer plasmids comprising the heterologous gene of choice [8, 9]. A significant advance came with the introduction of baculoviral genomes in the form of a bacterial artificial chromosome (BAC), which could be manipulated with comparative ease in *E. coli* cells [10], resulting in BEVS as a convenient eukaryotic protein production platform.

Current BEVS are mostly derived from the *Autographa californica* multiple nucleopolyhedrosis virus (AcMNPV) baculoviral genome, comprising approximately 140 kb genomic DNA [11], including a large set of genes regulating the life cycle of the virus in its natural habitat [12]. Many genes, however, will be unnecessary and sometimes even detrimental for applications in cell culture in the laboratory. For example, the gene encoding for the protease, V-CATH [13] and its molecular chaperone, the chitinase, ChiA, [14] were found to compromise heterologous expression and thus were eliminated in baculovirus genomes used for high-quality protein production including the MultiBac system we developed [15–17]. Deep data mining and comparative genome analyses by our group into the MultiBac baculovirus genome suggested that many more genes and other DNA elements present within AcMNPV could be unnecessary for laboratory culture, and thus could be conceivably disposed of in a redesigned baculoviral genome comprising extensive gene deletions [18].

Our approach of choice to implement this redesign consists of rewiring segments of the baculoviral genome by iteratively replacing wild-type sequences with synthetic DNAs devoid of the genes and DNA regions that we identified for deletion, resulting in baculoviral genome variants of increasingly smaller size and containing increasingly more synthetic DNA [19]. This approach requires an efficient method for engineering large constructs of DNA in a repetitive manner. An efficient method to enable genetic engineering of large DNA constructs is homologous recombination (HR) [20, 21] using the Red/ET system (Gene Bridges GmbH, Germany). This technique relies on linear DNA substrates that contain two approximately 50 base pair (bp) regions of homology that correspond to the target site flanking the genetic material required for the desired modification. Red/ET affords facile generation of mutations, deletions, insertions, gene replacements, or inversions [22]. Selection pressure to incorporate a synthetic fragment of choice into the target DNA is typically exerted by an antibiotic selection marker supplied within the synthetic fragment. To enable an iterative process, we chose a selection of distinct homing endonucleases and site-specific recombinases to remove the selection markers of choice after each step. In this way, the same set of selection markers can be used and subsequently eliminated in each reaction cycle.

In our protocol, firstly, the native DNA sequence from the baculovirus genome is replaced by a selection marker (called here Res1) using an initial Red/ET catalyzed homologous recombination reaction in *E. coli* cells harboring the baculoviral genome in the form of a BAC. Next, a second homologous recombination reaction using identical homology regions is carried out to replace Res1 with a synthetic fragment of DNA containing the desired rewired DNA sequence devoid of detrimental genes and other undesired

DNA regions, and a second, distinct selection marker (Res2). Res2 is flanked by recognition sites of an enzyme (e.g., a homing endonuclease or a site-specific recombinase) for its subsequent removal. Once Res2 is removed, a hybrid (part wild-type, part synthetic) genome comprising the synthetic segment of choice is obtained and can be experimentally validated. The process is then repeated iteratively until the entire set of alterations is introduced in an increasingly synthetic, increasingly smaller baculoviral genome (Fig. 1).

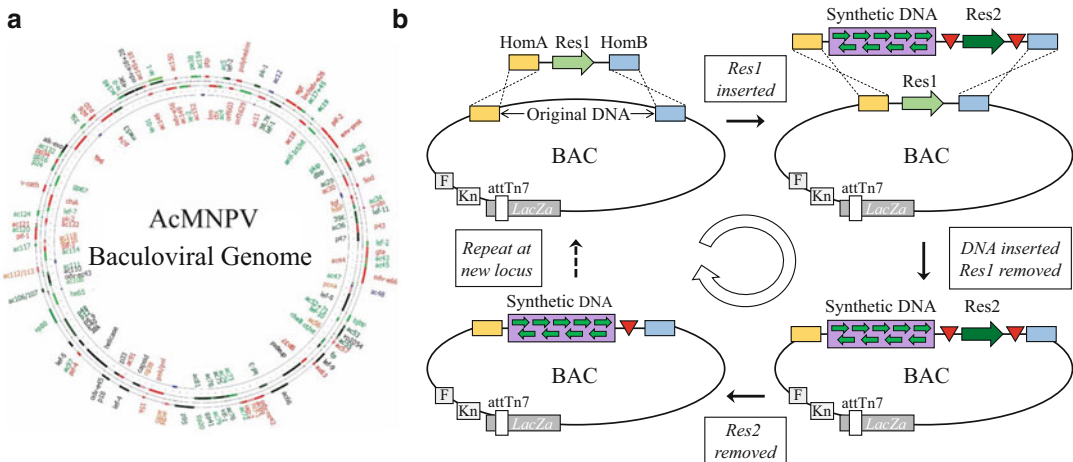


Fig. 1 Baculoviral genome modification by iterative homologous recombination. **(a)** The AcMNPV baculoviral genome (approximately 140 kb) is depicted schematically. Many genes (colored in red) and other DNA regions representing altogether more than 30% of the wild-type sequence can be conceivably eliminated by replacing segments of the wild-type genome with condensed, required synthetic DNA segments [18]. **(b)** The iterative process to create such synthetic baculovirus variants is shown. First, the native sequence within the baculovirus, present as a bacterial artificial chromosome (BAC) in *E. coli* cells, is replaced with a synthetic fragment of DNA containing homology arms (HomA and HomB, yellow and blue rectangles) representing short (approximately 50 bp) sequences flanking the part to be replaced. A gene encoding for antibiotic resistance (Res1, light green arrow) is contained in between HomA and HomB. Replacement of the wild-type sequence with this synthetic DNA is carried out by using homologous recombination with Red/ET enzymes expressed from a plasmid (pRed/ET, Gene Bridges GmbH) transformed into the *E. coli* cells. A second Red/ET HR reaction then replaces Res1 with a synthetic DNA fragment of choice comprising the same homology regions (HomA and HomB) flanking the custom-designed synthetic DNA segment (green arrows in purple box) with a gene encoding for a second antibiotic resistance (Res2, dark green arrow). The Res2 selection marker gene is flanked by recognition sites (red triangles) for an enzyme (e.g., a homing endonuclease or a recombinase) for the subsequent removal of Res2. The process can be repeated iteratively. BAC-specific DNA elements (Kn, kanamycin resistance marker; F, F replicon; attTn7, Tn7 transposon attachment site; LacZα gene for blue/white screening) are shown as boxes colored in gray

2 Materials

All reagents should be prepared using ultra-pure water (Millipore Milli-Q system or equivalent, with a sensitivity of 18.2 M Ω ·cm at 25 °C). DNA, antibiotics, enzymes, and their corresponding buffers should be stored at –20 °C and bacterial cell stocks at –80 °C.

2.1 *In Silico Design*

1. Molecular biology software for DNA visualization (e.g., SnapGene or Ape).
2. Sequences of the plasmid backbones of choice (e.g., MultiBac donors).
3. Sequences of the desired DNA modification.

2.2 *Preparation of DNA Fragment of Choice*

1. PCR primers specific to your fragment of choice.
2. DNA polymerase (e.g., Phusion[®] High Fidelity DNA Polymerase, NEB), reaction buffer and dNTP solution.
3. Agarose gel electrophoresis equipment.
4. Method for visualizing DNA agarose gel (e.g., UV/Blue light box).
5. Gel extraction kit (e.g., Monarch DNA Gel Extraction Kit).
6. Method for quantifying DNA (e.g., Thermo Scientific NanoDrop 2000).

2.3 *Bacterial Cells Containing BAC and Plasmid pRed/ET*

1. *E. coli* cells DH10 β _BAC harboring the BAC of choice (e.g., from glycerol stock).
2. Lysogeny broth (LB) medium.
3. LB agar.
4. Petri dishes.
5. Antibiotic stock solutions appropriate for your BAC and DNA fragments.
6. IPTG and BluOGal (if BAC contains LacZ α cassette for blue/white screening).
7. 37 °C incubators (shaking and stationary).
8. Falcon tubes (15 and 50 mL), Eppendorf tubes (1.5 mL).
9. Spectrophotometer for OD measurements (not essential but preferred).
10. Cooled microfuge (4 °C).
11. Cold, sterile 10% glycerol.
12. pRed/ET plasmid (Gene Bridges GmbH).
13. 30 °C incubator (shaking and stationary).
14. Electroporator and cuvettes.

2.4 First Recombination Reaction

1. Sterile 10% L-arabinose.
2. 100–200 ng of the first fragment of DNA to be inserted.
3. Bacteria spreader.
4. Inoculation loops.
5. Primers for investigating correct clones.
6. PCR kit (e.g., Phusion[®] High Fidelity DNA Polymerase, NEB).

2.5 Second Recombination Reaction

1. *See* Subheading 2.4 **First recombination reaction**.
2. 100–200 ng of the second fragment of DNA to be inserted.

2.6 Removal of Selection Marker Res2

1. Qiagen Buffers P1, P2, P3 (or N3).
2. DH10 β cells.
3. Your chosen enzymes for removal of the selection marker (e.g., homing endonuclease and T4 DNA ligase, or a site-specific recombinase).
4. 0.22 μ m filter membranes.
5. Tweezers.

3 Methods

3.1 In Silico Design

1. Determine the boundaries of the segment of DNA from the BAC for modification and utilize the 50 bp up- and downstream as the homology arms, HomA and HomB.
2. Design a fragment of DNA to first replace the native BAC DNA beginning and ending with a blunt restriction enzyme site (e.g., StuI) and containing a gene encoding for a resistance marker (Res1, for example, gentamycin acetyltransferase) flanked by the 50 bp homology regions determined from the BAC (HomA and HomB): StuI_HomA_Res1_HomB_StuI (*see* **Notes 1** and **2**).
3. Decide upon a method for the removal of the second resistance marker (Res2). For example, this may consist of homing endonucleases (if not already present within the BAC) or site-specific recombinases (e.g., Cre recombinase or similar [**23**]).
4. Design the fragment of DNA to insert into the BAC. This should, once again, begin and end with blunt restriction sites (e.g., StuI) and the BAC homology regions (HomA and HomB), though this fragment should now contain the desired genetic material you wish to insert into the BAC and a second gene encoding for antibiotic resistance (e.g., β -lactamase), flanked by the recognition sites (RecogSite) of your chosen

method for its subsequent removal. Following the format: StuI_HomA_DesiredDNA_RecogSite_Res2_RecogSite_HomB_StuI.

5. Decide on your DNA assembly strategy (e.g., gene synthesis, PCR assembly or restriction/ligation).
6. Decide upon a donor plasmid backbone for the above fragments to be inserted into (e.g., pDS, pDK, or pDC [24]) and choose your DNA assembly strategy (e.g., gene synthesis, PCR assembly or restriction/ligation-based cloning) (*see Note 3*).
7. Design suitable primers for checking the modification of the BAC along the process (*see Note 4*).

3.2 Preparation of the DNA Fragment of Choice

1. Assemble the designed plasmids as per your desired method.
2. Digest several micrograms of the assembled plasmids with your chosen restriction enzyme, (e.g., StuI) to extract the fragments of interest (*see Note 5*).
3. Analyze the digestion using agarose gel electrophoresis to ensure the desired band sizes are observed, and to confirm that the digestion reactions are complete.
4. Purify the digested fragments using a commercial gel extraction kit and elute in the minimal volume defined by the manufacturer using sterile MQ H₂O (*see Note 6*).
5. Determine the concentration of the extracted DNAs spectrophotometrically.

3.3 Bacterial Cells Containing BAC and Plasmid pRed/ET

1. Streak out DH10 β _BAC cells from a glycerol stock that contain “*only*” the BAC of interest onto agar plates containing the appropriate antibiotics (*see Notes 7 and 8*).
2. Prepare a pre-culture for overnight growth containing 3 mL LB and the appropriate antibiotics, inoculate with a single colony of DH10 β _BAC and incubate shaking at 37 °C for 16–18 h.
3. Start a growth culture the next morning containing 14 mL LB supplemented with the appropriate antibiotics and add approximately 0.3–0.5 mL of the overnight pre-culture (to gain a desired starting OD₆₀₀ of approximately 0.08–0.1) and grow for approximately 2–3 h until OD₆₀₀ = approximately 0.6.
4. Prepare 1.4 mL cells for electroporation. Centrifuge 11,000 g, 1 min, 4 °C, remove the supernatant and gently resuspend the resulting pellet in 1 mL sterile 10% glycerol, pre-cooled to 4 °C. Repeat this step twice, finally resuspending in 50 μ L 10% glycerol.

5. Transform 100–200 ng pRed/ET into the prepared cells containing the BAC by electroporation following standard electroporation protocols (*see* **Note 9**).
6. Incubate the electroporation mixture in LB containing no antibiotics at 30 °C for 1–2 h, plate onto agar containing the appropriate antibiotics and incubate at 30 °C overnight (*see* **Notes 8 and 10**).

3.4 First Red/ET Recombination

1. Prepare a pre-culture for overnight growth containing 3 mL LB and the appropriate antibiotics, inoculate with a single colony of DH10 β _BAC_Red/ET and incubate shaking at 30 °C for 16–18 h.
2. The next morning, inoculate 14 mL LB containing the appropriate antibiotics with 0.5 mL pre-culture and grow at 30 °C until OD = 0.3 (*see* **Notes 10 and 11**). This should take approximately 2 h.
3. Transfer 1.4 mL of the bacterial cell culture into an Eppendorf tube, induce with 10% L-Arabinose (50 μ L), and incubate at 37 °C for 1 h.
4. Prepare the cells for electroporation as previously described (*see* Subheading 3.3, **step 4**).
5. Transform 100–200 ng of the first DNA fragment to be inserted (StuI_HomA_Res1_HomB_StuI) into the prepared cells by electroporation following standard electroporation protocols.
6. Incubate the electroporation mixture in LB containing no antibiotics at 37 °C for 2–3 h, plate 100 μ L onto agar containing the appropriate antibiotics and, using the same spreader, a second dilution plate. Incubate overnight at 37 °C (*see* **Note 12**).
7. Identify correct clones through colony growth on correct antibiotics. Additionally, confirm clones using PCR amplification of region of interest on the BAC to investigate the length of DNA present between the homology regions.

3.5 Second Recombination Reaction

1. Prepare electrocompetent DH10 β cells containing a confirmed correct clone of the newly modified BAC_HomA_Res1_HomB from Subheading 3.4, **step 7** and transform in pRed/ET as previously described (*see* Subheadings 3.3, **steps 2–6**).
2. The second Red/ET recombination reaction is then carried out repeating Subheadings 3.4, **steps 1–7** though instead, utilizing the second DNA fragment previously prepared: StuI_HomA_DesiredDNA_RecogSite_Res2_RecogSite_-HomB_StuI during Subheading 3.4, **step 5**.

3.6 Removal of Selection Marker Res2

1. Inoculate 3 mL LB containing the appropriate antibiotics with a single colony containing DH10 β _BAC_HomA_DesiredDNA_RecogSite_Res2_RecogSite_HomB and incubate at 37 °C for 16–18 h.
2. Centrifuge the overnight bacterial culture 3500 g, 10 min.
3. Resuspend the cell pellet in 300 μ L Buffer P1 (Qiagen) and transfer to a 1.5 mL Eppendorf.
4. Add 300 μ L Buffer P2 (Qiagen) and very gently, invert the tube 4–6 times until the mixture is homogeneous.
5. Add 300 μ L Buffer N3 (Qiagen) and very gently, invert the tube 4–6 times until the mixture is homogeneous.
6. Centrifuge at 17,000 g, 10 min then transfer to a clean Eppendorf (*see Note 13*).
7. Centrifuge at 17,000 g, 10 min.
8. Transfer 800 μ L of the mixture to a clean Eppendorf, slowly add 700 μ L isopropanol, mix by carefully inverting the tube, and incubate at 4 °C for 10 min.
9. Centrifuge at 17,000 g, 10 min, 4 °C.
10. Very carefully remove and discard the supernatant (*see Note 14*).
11. To wash the bacmid, slowly add 200 μ L 70% EtOH dropwise (*see Note 15*).
12. Centrifuge at 17,000 g, 5 min, 4 °C.
13. Very carefully remove and discard the supernatant.
14. Under sterile conditions, leave the Eppendorf tube open to allow the EtOH to evaporate for 10 min.
15. To resuspend the pellet, add 30 μ L sterile MQ H₂O and gently tap 10 times. (*see Notes 16 and 17*).
16. Incubate the BAC with the enzyme from your chosen method of removal of the resistance marker and if possible, subsequently heat inactivate the enzyme.
17. Dialyze the BAC on 0.22 μ m filter paper into MQ water for 45–60 min.
18. If a homing endonuclease was used in **step 16**, the BAC must be recircularized. In these instances, the dialyzed BAC was incubated with T4 DNA ligase at room temperature for 1 h. If another means of removal was used such as a site-specific recombinase, continue to **step 20**.
19. Repeat the dialysis of BAC as per **step 17**.
20. Transform the modified BAC into DH10 β cells by electroporation following standard electroporation protocols.

21. Incubate the electroporation mixture in 1 mL LB at 37 °C for 2–3 h, prior to plating onto agar containing the appropriate antibiotics (*see* **Note 8**).
22. Re-streak colonies onto agar plates with and without the antibiotic corresponding to the second resistance marker to identify correct clones that do not harbor Res2 (*see* **Note 18**).
23. Further confirm these clones with PCR amplification of the desired region of the modified BAC.

4 Notes

1. Utilizing blunt restriction sites to prepare the DNA fragment minimizes the risk of carrying forward mutations. *StuI* has been chosen as an example restriction enzyme in this case for the DNA fragment for integration, ensure that if you use this site, that it is not present elsewhere within the fragment. However, the DNA fragment can also be prepared by PCR amplification but it is advisable to verify the amplified fragment by sequencing analysis prior to use.
2. Ensure the choice of selection marker does not match the resistance of either the BAC, pRed/ET, or *E. coli* strain in use.
3. The fragments of DNA could, in theory, be inserted into any plasmid backbone, however, we suggest utilizing a plasmid that contains an alternative replication origin, (e.g., the donor plasmids from the ACEMBL suite [24]) to minimize the risk of plasmid retention and future contamination.
4. If possible, design primers with an annealing temperature of approximately 63 °C and away from any baculoviral homologous repeat (hr) regions. This reduces the chances of nonspecific binding.
5. It is possible to do this step without the restriction enzymes and to amplify by PCR. However, this increases the risk of the addition of mutations into the genetic sequences, which can have adverse effects later on. If PCR amplification is your method of choice, we strongly recommend retrieving sequencing data to confirm the correct sequence is present.
6. We recommend using sterile MQ H₂O for the elution at this stage as this fragment will be used during a subsequent electroporation transformation reaction, whereby the presence of salts (e.g., eluting using Qiagen EB buffer) has a detrimental effect.
7. If the DH10 β cells contain additional DNA plasmids (e.g., helper plasmid to produce Tn7 transposon [10]), these may hinder the homologous recombination reactions. Therefore, ensure that the cells utilized only contain the BAC to be

modified. To remove any additional plasmids, isolate the BAC (*see* Subheading 3.6, **steps 2–15**) and electroporate into a fresh aliquot of electrocompetent DH10 β cells, and subsequently plate onto agar that contains only the antibiotics present within the BAC. Re-streak several colonies onto both the combined antibiotics and singular antibiotics to confirm the loss of the additional plasmid. Due to the large size of the BAC, we would not recommend chemical transformation protocols.

8. If the BAC contains the LacZ α cassette, also add IPTG and BluOGal onto the agar plates to obtain positive blue colonies representing intact baculovirus genomes.
9. This step may also be performed by chemical transformation by replacing the 10% glycerol with 0.1 M NaCl; however, due to the large size of the pRed/ET plasmid we found electroporation to be more efficient.
10. It is most important to carry out this incubation at 30 °C to retain the pRed/ET plasmid. This plasmid contains a temperature-sensitive cassette and incubation at 37 °C will result in loss of pRed/ET.
11. The cellular replication will be slower at 30 °C; therefore, the volume of pre-culture added can be increased if previous attempts to reach OD = 0.3 took longer than 2 h.
12. If no or too few colonies are present at this stage, the rest of the cell culture can be plated onto an additional agar plate.
13. We recommend carrying out this step using a p1000 pipette and doing this in “one shot” in order to minimize the disturbance to the undesired pellet.
14. Be very careful not to disturb the DNA pellet, which should be small and mostly transparent. A big white pellet at this stage indicates that cell debris and/or DNA other than the BAC may have been brought forward.
15. Be careful not to dislodge the DNA pellet at the bottom of the Eppendorf. We recommend adding the EtOH dropwise to the opposite side of the tube where the pellet is expected.
16. Do not pipette up and down at this stage as this can cause the BAC to break.
17. We recommend taking a 10 μ L aliquot of the prepared BAC to analyze by agarose gel electrophoresis to confirm the presence of bacmid within the isolated sample.
18. We recommend re-streaking in excess of 20 colonies to increase chances of identifying a clone that is negative for the resistance marker that was removed. The efficiency of this reaction will greatly depend on your chosen method.

Acknowledgments

We thank all members of the Berger laboratory for their contributions. We are grateful to Robert Roth (AstraZeneca) for helpful discussions. The authors thank Francis Stewart for the pRed/ET plasmid. B.G. is supported by the Biotechnology and Biological Research Council (BBSRC) through a scholarship from the South West Doctoral Training Programme, SWDTP. I.B. is supported by a European Research Council ERC Advanced Grant (DNA-DOCK) and is recipient of a Wellcome Trust Senior Investigator Award.

Competing Financial Interest Statement: *The authors declare competing financial interest. I.B. is inventor on patents protecting MultiBac. I.B. is also shareholder of biotech companies commercializing MultiBac applications.*

References

1. Gupta K, Tölzer C, Sari-Ak D et al (2019) MultiBac: Baculovirus-mediated multigene DNA cargo delivery in insect and mammalian cells. *Viruses* 11:1–14
2. Summers MD (2006) Milestones leading to the genetic engineering of baculoviruses as expression vector systems and viral pesticides. *Adv Virus Res* 68:3–73
3. van Oers MM, Pijlman GP, Vlak JM (2015) Thirty years of baculovirus-insect cell protein expression: from dark horse to mainstream technology. *J Gen Virol* 96:6–23
4. Assenberg R, Wan PT, Geisse S, Mayr LM (2013) Advances in recombinant protein expression for use in pharmaceutical research. *Curr Opin Struct Biol* 23:393–402. <https://doi.org/10.1016/J.SBI.2013.03.008>
5. Fernandes F, Teixeira AP, Carinhas N et al (2013) Insect cells as a production platform of complex virus-like particles. *Expert Rev Vaccines* 12:225–236
6. Airene KJ, Hu Y-C, Kost TA et al (2013) Baculovirus: an insect-derived vector for diverse gene transfer applications. *Mol Ther* 21:739–749
7. Kost TA, Condreay JP (2002) Recombinant baculoviruses as mammalian cell gene-delivery vectors. *Trends Biotechnol* 20:173–180
8. Hartig PC, Cardon MC (1992) Rapid efficient production of baculovirus expression vectors. *J Virol Methods* 38:61–70
9. Hitchman RB, Possee RD, King LA (2012) High-throughput baculovirus expression in insect cells. In: Lorence A (ed) *Methods in molecular biology*, 3rd edn. Humana Press, Totowa, NJ, pp 609–627
10. Luckow V, Lee S, Barry G, Olins P (1993) Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in *Escherichia coli*. *J Virol* 67:4566–4579
11. Ayres MD, Howard SC, Kuzio J et al (1994) The complete DNA sequence of Autographa californica nuclear polyhedrosis virus. *Virology* 202:586–605
12. Rohrmann G (2013) *Baculovirus Molecular Biology*. National Center for Biotechnology Information (US), Bethesda, MD
13. Slack JM, Kuzio J, Faulkner P (1995) Characterization of v-cath, a cathepsin L-like proteinase expressed by the baculovirus Autographa californica multiple nuclear polyhedrosis virus. *J Gen Virol* 76:1091–1098
14. Hom LG, Volkman LE (2000) Autographa californica M Nucleopolyhedrovirus chiA is required for processing of V-CATH. *Virology* 277:178–183
15. Hawtin RE, Zarkowska T, Arnold K et al (1997) Liquefaction of Autographa californica nucleopolyhedrovirus-infected insects is dependent on the integrity of virus-encoded chitinase and cathepsin genes. *Virology* 238:243–253
16. Kaba SA, Salcedo AM, Wafula PO et al (2004) Development of a chitinase and v-cathepsin negative bacmid for improved integrity of

- secreted recombinant proteins. *J Virol Methods* 122:113–118
17. Berger I, Fitzgerald DJ, Richmond TJ (2004) Baculovirus expression system for heterologous multiprotein complexes. *Nat Biotechnol* 22:1583–1587
 18. Vijayachandran LS, Thimiri Govinda Raj DB, Edelweiss E et al (2013) Gene gymnastics synthetic biology for baculovirus expression vector system engineering. *Bioengineered* 4:279–287
 19. Pelosse M, Crocker H, Gorda et al (2017) MultiBac: from protein complex structures to synthetic viral nanosystems. *BMC Biol* 15:99
 20. Zhang Y, Buchholz F, Muyrers JPP, Stewart FA (1998) A new logic for DNA engineering using recombination in *Escherichia coli*. *Nat Genet* 20:123–128
 21. Muyrers J, Zhang Y, Testa G, Stewart FA (1999) Rapid modification of bacterial artificial chromosomes by ET- recombination. *Nucleic Acids Res* 27:1555–1557
 22. Tischer BK, Von Einem J, Kaufer B, Osterrieder N (2006) Two-step red-mediated recombination for versatile high-efficiency markerless DNA manipulation in *Escherichia coli*. *Bio-Techniques* 40:191–197
 23. Meinke G, Bohm A, Hauber J et al (2016) Cre recombinase and other tyrosine recombinases. *Am Chem Soc* 116:12785–12820
 24. Nie Y, Chaillet M, Becke C et al (2016) ACEMBL tool-kits for high-throughput multi-gene delivery and expression in prokaryotic and eukaryotic hosts. In: *Advanced technologies for protein complex production and characterization*, 896th edn. Springer, Cham, pp 27–42



Gene Tagging with the CRISPR-Cas9 System to Facilitate Macromolecular Complex Purification

Sylvain Geny, Simon Pichard, Arnaud Poterszman,
and Jean-Paul Concordet

Abstract

The need to generate modified cell lines that express tagged proteins of interest has become increasingly important. Here, we describe a detailed protocol for facile CRISPR/Cas9-mediated gene tagging and isolation of modified cells. In this protocol, we combine two previously published strategies that promote CRISPR/Cas9-mediated gene tagging: using chemically modified single-stranded oligonucleotides as donor templates and a co-selection strategy targeting the *ATP1A1* gene at the same time as the gene of interest. Altogether, the protocol proposed here is both easier and saves time compared to other approaches for generating cells that express tagged proteins of interest, which is crucial to purify native complex from human cells.

Key words CRISPR/Cas9, Co-selection, Complex purification, Single-stranded oligonucleotide donor

1 Introduction

Using cells that express a genetically tagged subunit is a simple and efficient way to undertake affinity purification of a macromolecular complex of interest [1]. Antibodies to specific subunits could also be used but their isolation is often challenging, and they are seldom available. This obstacle can be overcome by expressing a fusion of the protein of interest with a peptide tag for which high affinity reagents are already available and well characterized. The method presented here consists in using optimized gene editing with the CRISPR-Cas9 system to introduce the coding sequence for a small peptide tag at the 5' or 3' end of the chosen subunit in order to facilitate affinity purification. Importantly, in contrast to transfection experiments, where the tagged protein is expressed from an exogenous promoter, the tagged protein is expressed from the endogenous locus and subject to the full extent of physiological control of gene expression. This approach therefore minimizes the

risk of complex perturbation due to inappropriate expression levels associated with traditional transfection methods.

The CRISPR-Cas9 system has revolutionized many fields of life sciences by making it possible to modify the genome sequence with unprecedented efficiency in a great number of biological systems [2, 3]. Using a specific guide RNA, the Cas9 nuclease can be directed to generate a double-strand break (DSB) at virtually any chosen genomic site. To ensure their survival, cells will repair the damaged DNA. If DSB repair proceeds by end-joining repair pathways, small insertions or deletions are introduced at the site of the DSB while if template donor DNA is introduced together with the CRISPR-Cas9 nuclease, homology-directed repair (HDR) can take place and precise genome editing is achieved [4]. Donor DNA that can be used is either double-stranded plasmid or single-stranded oligonucleotides (ssODN). Precise genome editing with the CRISPR-Cas9 system is therefore a powerful method to introduce tag-coding sequences into genes of interest and greatly facilitates their analysis. However, HDR is generally less efficient than end-joining and HDR-based gene editing needs to be carefully optimized to achieve efficient insertion of tag-coding sequences.

The precise gene editing method presented here is optimized at two levels, based on previously published studies. First, chemically modified ssODNs are used to increase the efficiency of precise gene editing [5] and second, a co-selection strategy is applied to increase the proportion of cells with the gene modification of interest.

The co-selection strategy was devised by the Doyon laboratory [6]. It consists of co-targeting the *ATP1A1* gene with a specific guide RNA and ssODN designed to introduce an *ATP1A1* mutation conferring resistance to ouabain, a toxic glycoside. After ouabain selection, the proportion of cells with gene editing at the locus of interest is significantly increased. The principle and efficiency of co-selection were initially reported in gene editing experiments with zinc finger nucleases [7]. The cellular mechanism for co-selection is not characterized, but it is thought that successful HDR at the selection gene locus likely corresponds to favorable cellular conditions for simultaneous HDR at the target locus, for example, during S/G2 cell cycle phases.

The co-selection of ouabain-resistant cells devised by the Doyon lab is more convenient than previous co-selection procedures [8] because it does not require the introduction of exogenous selection cassettes and can be very efficient, resulting in up to ten-fold enrichment. Importantly, co-selection generally makes it easier to isolate cells with modification of multiple alleles of the targeted gene. When tagging the subunit of a macromolecular complex, modification of multiple alleles increases the proportion of the complex that can be purified from edited cells. Furthermore, if all alleles are modified, it is easier to confirm that protein tagging does not perturb activity and regulation of the macromolecular

complex of interest. Finally, the protocol proposed here is very efficient which is critical to achieve gene tagging with ssODNs that are too small to include selection cassettes in addition to the tag-coding sequence. The lymphoblastoid cell line, K562, has been chosen for gene editing since the cells can be grown in to large-scale volumes in suspension, enabling the production of sufficient amounts of complex for purification.

The ATP1A1 co-selection strategy can be applied to the plasmid or ssODN donors [6]. ssODNs are particularly appealing to achieve efficient insertion of tag-coding sequences because they alleviate the need to construct donor plasmids. Although commercially available ssODNs have a length restriction under 200 nt, they can generally be used because a majority of purification tags require only a short-coding sequence. The ATP1A1 co-selection approach using ssODNs makes it possible to generate cells with several different purification tags in parallel to determine the optimal tag for purification of the macromolecular complex of interest. This cannot necessarily be predicted in advance [9, 10].

Here, we demonstrate the feasibility of the method by tagging the XPB subunit of the TFIIH complex that plays a major role in DNA nucleotide excision repair and transcription [11]. We show that despite low DNA cleavage efficiency of the XPB gene with the CRISPR/Cas9 system in our experimental conditions, our approach is robust enough to isolate cells with insertion of tag-coding sequences by screening of only a small number of clones. We provide a detailed protocol applicable to any gene of interest to purify macromolecular complexes that cannot be isolated following previously described purification methods.

2 Materials

Prepare all solutions using ultrapure water and analytical grade reagents. Perform all cell culture work under sterile conditions. Diligently follow all waste disposal regulations when disposing waste materials.

Procedures described here need access to standard equipment for molecular biology, cell culture, and protein purification and require basic knowledge in these fields is required. Specific equipment and material is detailed below.

2.1 Oligonucleotides and Plasmids

1. Oligonucleotides for construction of sgRNA expression plasmids are listed in Table 1.
2. ssODN donor oligonucleotide for ATPA1 and XPB (Table 2).
3. Forward and reverse primers for genotyping (Table 3).
4. Expression plasmid for sgRNA: MLM3636 (Addgene cat. no. 43860).

Table 1

Target sequences for sgRNAs. The 20 nt-long target sequence (N_{20}) of each guide RNA and adjacent PAM are indicated. Oligonucleotides for cloning into the BsmB1 site MLM3636 should contain the appropriate 5' and 3' extensions: 5'ACACC (G) N_{20} G 3' for the sense oligonucleotide and 5'AAAA n_{20} C3' for the antisense oligonucleotide extension where n_{20} is the reverse complement of N_{20} . Online applications such as CRISPOR design the sequences to be ordered and examined if the target sequence starts with a G (required for efficient U6 transcription). If this is not the case, an extra G is added 5' to the guide sequence in order to ensure efficient transcription from the U6 promoter

Guide RNA Plasmid	Target sequence	PAM
sgRNA-ATP1A1	GAGTTCTGTAATTCAGCATA TGG	TGG
sgRNA-XPB1	TAGGAAATGATGCTTAGGCA GGG	GGG
sgRNA-XPB2	TTAGGAAATGATGCTTAGGC AGG	AGG
sgRNA-XPB3	CGCTTTAGGAAATGATGCTT AGG	AGG

5. Expression plasmid for SpCas9: JDS246 (Addgene cat. no. 43861).
6. Plasmid sgG3-ATP1A1-Cas9 (Addgene cat. no. 86611).

2.2 Cell Culture

1. Tissue culture flasks (25, 75, and 175 cm²), reagent glass bottles of various sizes (250 mL ($h = 105$ mm, $d = 95$ mm, opening = 32 mm), 1000 mL ($h = 222$ mm, $d = 101$ mm, opening = 32 mm), and 5000 ml ($h = 314$ mm, $d = 182$ mm, opening = 68 mm), Duran® GLS 80® laboratory bottles) and gas permeable adhesive seal. Bottle caps are removed replaced with an aluminum foil and sterilized by autoclaving (134 °C, 15 psi, 20 min).
2. RPMI 1640 with L-Glutamine.
3. Fetal bovine serum (heat inactivated) (FBS).
4. Penicillin-Streptomycin 10,000 U/mL.
5. Human chronic myelogenous leukemia K562 cells (ATCC, cat. no. CCL-243).
6. Phosphate-buffered saline.
7. Cell counter.
8. Cell counting chambers.
9. 0.4% Trypan blue dye solution in PBS.
10. Humidity, CO₂, and temperature controlled orbital shaker fitted for 50 mL to 5 L glass bottles with 50 mm orbital and shaking speed of up to 150 rpm (InforsMultitron™).
11. Centrifuge with adaptors for 15 mL, 50 mL, 250 mL, and 1 L. For the 1 L tubes, use preferably a fixed angle rotor.

Table 2

Sequences of the single-stranded donor oligonucleotides. Homology arms are indicated in lower case. Epitope tag sequences 3Flag, His8, HiBiT, and cleavage site PreScission are in uppercase and underlined. Many different tags have been characterized and can be chosen for specific applications [15]. For affinity purification, peptide tags are frequently used and small tags such as 3 × Flag and polyhistidine are generally preferred because they are considered less likely to affect protein function. Linker regions and stop codons are in uppercase. Phosphorothioate linkages are marked with an asterisk. The sequence of the ssODN mutATP1A1 donor is from [6]

Donor ON	Sequence
mutATP1A1:	C*A*ATGTTACTGTGGATTGGAGCGATTCTTTGTTTCTTGGCTTATAGCATCAGAGCTGC TACA GAAGAGGGAACCTCAAAACGATGACGTGAGTTCGTGTAATTCAGCATATCGATTGTGTAGTACAC ATCAGATATC*T*T
XPB-Hibit	c*g* <u>cccagcaaacatgtacaccgcgtctcttcaagcgttttaggaaaGGTTCCGTGAGCGGCTGGCGGGCTGTTCAA</u> GAAGATTAGCTGAcagggtacttcgttcaagacggcgcttggcacccctgttggga*a*a
ON1 XPB-Flag3X-His8	c*g* <u>cccagcaaacatgtacaccgcgtctcttcaagcgttttaggaaaGGTTCCGACTACAAAGACCATGACGGTGATTA</u> TAAAGATCATGACATCGATTACAAGGATGACGATGACAAAGGCGAGCGGCCCATCATCACCCAC CATCACCCACCAJTAGcagggtacttcgttcaagacggcgcttggcacccctgttggga*a*a
ON2 XPB-PreScission-His10-CaptureSelect	c*g* <u>cccagcaaacatgtacaccgcgtctcttcaagcgttttaggaaAAGTTCCCTTGGAAGTTCTGTTCACAGGGGCCCC</u> ctgggatccCATCATCACCCACCATCACCAACCATCATCACGAACCTGAAGCCTAGCTAGcaggga Cttcgttcaagacggcgcttggcacccctgttggga*a*a
ON3 XPB-PreScission-Flag3X	c*g* <u>cccagcaaacatgtacaccgcgtctcttcaagcgttttaggaaaGGTTCCCTGGAAGTTCTGTTCACAGGGGCCCC</u> CTGGGATCCGACTACAAAGACCATGACGGTGATTATAAAAGATCATGACATCGATTACAA GGATGACGATGACAAAGTAGcagggtacttcgttcaagacggcgcttggcacccctgttggga*a*a

Table 3
Sequences of the primers used for genotyping the clones

XPBfw	AGACAGTAAGCGATCTGTAAACA
XPBbv	ACCCCACTCCCCAAAAAGTT

**2.3 Nucleofection
AMAXA**

1. Tissue culture plate, 6 wells.
2. AMAXA cuvettes (Lonza).
3. AMAXA nucleofector machine (Lonza).
4. AMAXA pipettes (Lonza).
5. Solution V AMAXA (Lonza).

**2.4 Nano-Glo HiBiT
Detection System**

1. ViewPlate 96-wells, white.
2. Nano-Glo HiBiT Lytic detection system (Promega).
3. Microplate Reader Wallac Victor 1420 (Perkin Elmer).

**2.5 Ouabain
Selection, Pool
and Clone Analysis**

1. Tissue culture plate, 6 wells.
2. Tissue culture plate, 96 wells.
3. Ouabain octahydrate.
4. DNA lysis buffer (Viagen).
5. Phusion polymerase (New England Biolabs).
6. dNTPs set.
7. Agarose powder.
8. Ethidium bromide.
9. Tris-Borate EDTA UltraPure 10X.
10. Bromophenol blue.
11. 1 Kb Plus DNA Ladder (Invitrogen).

**2.6 Biochemical
Characterization
and Purification**

1. Bioruptor (Diagenode) or any small-scale sonicator.
2. PBS containing 30% w/v glycerol.
3. RIPA buffer: 20 mM Tris-HCl or HEPES pH 7.5 120 mM KCl 1% NP-40 0,1% SDS, 1 mM EDTA, 0.5% Na-Deoxycholate, supplemented with protein inhibitor cocktail (Roche™) and 0.5 mM 1,4-dithreothiol (DTT) (Sigma Aldrich).
4. Lysis buffer: 20 mM Tris or HEPES, 250 mM KCl, 20% glycerol, NP-40 0.05% supplemented with protein inhibitor cocktail (Roche™) and 0.5 mM DDT.
5. DNase and RNase.
6. M2 agarose beads.
7. FLAG peptide (dykddddk) solution (10 mg/ml, pH adjusted to 7.5) and/or purified PreScission Protease (1 mg/mL).

8. Temperature controlled thermomixer.
9. TBST: 20 mM Tris/HCl pH 7.5, 150 mM NaCl, 0.1% Tween-20.
10. Blocking buffer: 3% w/v dry skimmed milk or BSA solution in TBST.
11. Laemmli buffer 4X: 60 mM Tris-HCl pH 6.8, 10% glycerol, 2% SDS, 0.0005% Bromophenol Blue, 355 mM β -mercaptoethanol.

3 Methods

We detail the different steps of our gene editing protocol (Fig. 1.) The design of both sgRNAs targeting the gene of interest and the ssODN for the chosen peptide tag are detailed first (Subheadings 3.1 and 3.2). Then, we describe how to take advantage of the newly developed Nano-Glo HiBiT Lytic Detection System to select the most efficient sgRNA for editing your target gene (Subheading 3.3). Next, we provide the protocol for transfection of gene editing reagents into cells, ouabain selection (Subheading 3.4), and isolation of correctly modified cells (Subheading 3.5), carrying the tagged gene of interest. Finally, we report validation of the modified cell line by western blot analysis (Subheading 3.6) and large-scale suspension cultures to purify the target complex (Subheading 3.7).

3.1 Design of sgRNAs for Targeting the Gene of Interest and Cloning into Guide RNA Expression Plasmid

Protein are usually tagged either at the N- or C-terminal end [13]. There is no easy method to predict how the tag will modify the protein structure and to conclude whether the protein and the tag will be accessible for their functions. Here, we show the step-by-step approach to introduce a purification tag at the C-terminal end of the protein XPB but a similar approach can be applied to tag the N-terminal end. To tag a protein at C-terminal end, the cut site introduced with the CRISPR/Cas9 system needs to be in the vicinity of the stop codon of the gene of interest.

1. Screen the genomic regions of interest using online CRISPR design tool CRISPOR (<http://crispor.tefor.net/>) to identify and rank the guide sequences. We recommend testing at least three sgRNAs with cut sites less than 20 bp away from the insertion site to find an optimal sgRNA.
2. Order the sense and antisense oligonucleotides for cloning of sgRNA sequences into plasmid MLM3636 from your preferred supplier. The oligonucleotide sequences needed are given by the CRISPOR website by clicking on the “Cloning/PCR primers” hyperlink for the selected guide RNA in the CRISPOR results web page. Resuspend the oligonucleotides for each

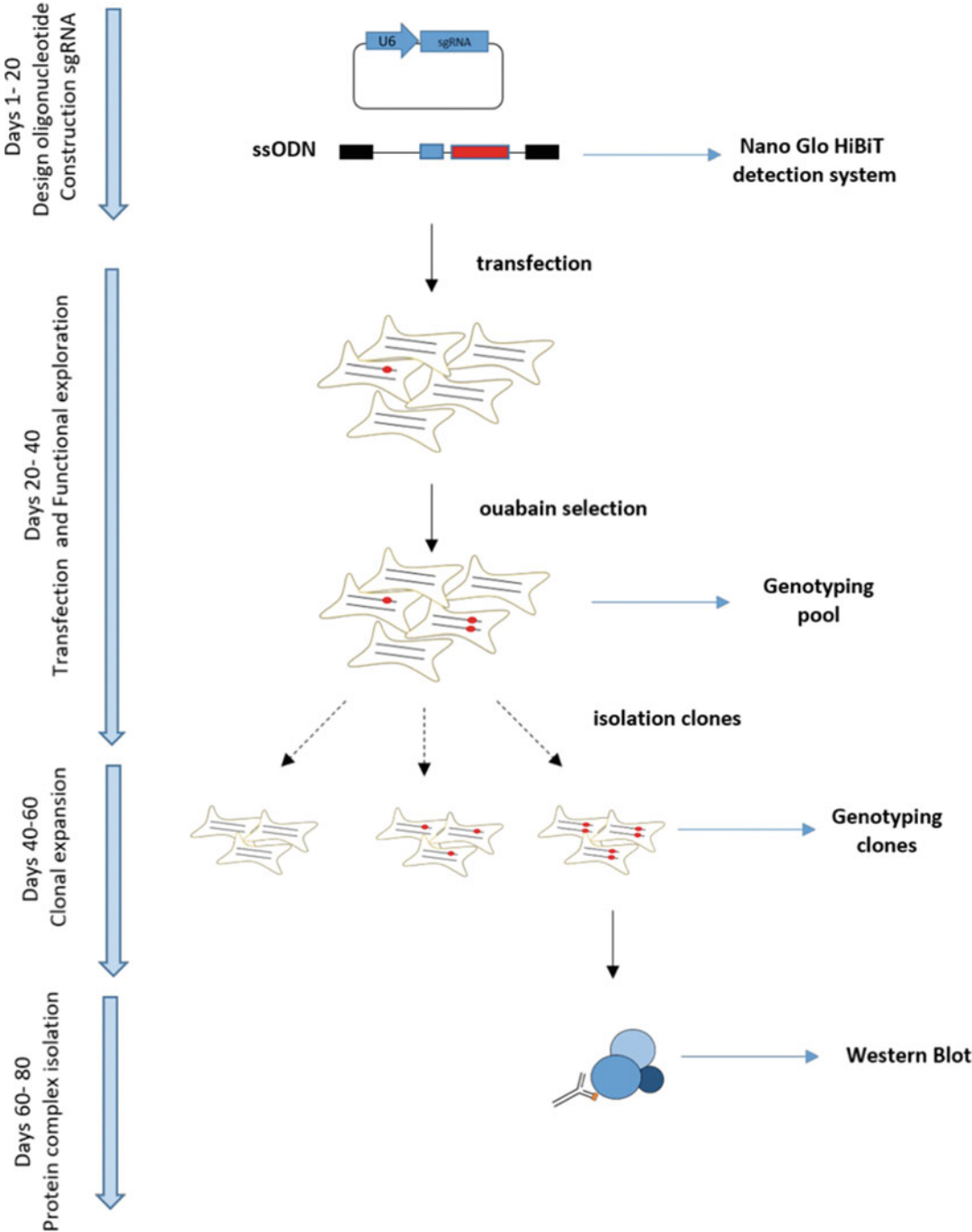


Fig. 1 Timeline of a gene tagging experiment for protein complex purification. The main steps for a gene tagging experiment for protein complex purification are depicted. sgRNA guide sequences are cloned into an expression plasmid. Single-stranded donor oligonucleotides and plasmids expressing sgRNA and SpCas9 are transfected into cells. Treatment with ouabain is performed to enrich for gene-tagged cells among the total cell population. Finally, positive cells are clonally expanded to derive isogenic cell lines that express the tagged subunit protein to isolate the full macromolecular complex

sgRNA at a concentration of 100 μ M. Prepare the following mixture for annealing the sgRNA oligos: 1 μ L of each oligonucleotides (100 μ M), 2 μ L buffer NEB2x10, and 16 μ L water.

3. Anneal the oligonucleotides in a thermocycler by using the following parameters: 37 °C for 15 min; 95 °C for 5 min; ramp down to 25 °C at 5 °C min⁻¹.
4. Digest 10 μ g plasmid MLM3636 with restriction enzyme BsmBI. Purify the linear plasmid using PCR cleanup kit. This step is sufficient to eliminate the short linker sequence between the two BsmBI sites of MLM3636.
5. Set up a ligation reaction for each sgRNA: 1 μ L of pre-digested plasmid vector sgRNA (50 ng/ μ L), 0.5 μ L of your annealing solution of sgRNA oligonucleotides, 1 μ L of T4 DNA ligase, 1.5 μ L of T4 DNA ligase buffer, 11 μ L of water.

We recommend also setting up a no-insert, vector only negative control for ligation to observe the background amount of undesired ligation. Incubate the ligation reactions at room temperature for 1 h.

6. Heat inactivate T4 DNA ligase with incubation at 65 °C for 10 min. Treat the ligation reaction for 30 min with BsmBI restriction enzyme to reduce background (arising from residual amounts of undigested plasmid or self-ligation of plasmid that was only digested once with BsmBI).
7. Transform the ligation product into a competent *E. coli* strain, according to the protocol supplied with the cells. We recommend the DH5 α or XL10 strains for quick transformation. Briefly, add 3 μ L of the ligation product into 30 μ L of ice-cold chemically competent cells, incubate the mixture on ice for 10 min, heat-shock it at 42 °C for 30 s and return it immediately to ice for 2 min. Add 500 μ L of SOC medium and incubate 1 h under shaking at 37 °C. Plate the mixture onto an LB plate containing 100 μ g/ml ampicillin. Incubate it overnight at 37 °C. Pick up a colony to grow overnight at 37 °C in an appropriate volume of LB medium with antibiotics.
8. Extract plasmid DNA using a Midi or Maxi Prep kit depending on the amount of plasmid required. Confirm successful sgRNA cloning by DNA sequencing.

3.2 Design of the Single-Stranded Oligonucleotide (ssODN) Donor

The oligonucleotide donors contain the coding sequence for one or several purification tags flanked by two homology regions as depicted in Fig. 2. Total length of the donor oligonucleotides should remain less than 200 bases to ensure efficient synthesis. For optimal HDR efficiency, we recommend designing homology regions of at least 45 bases on either side of the insert. If necessary, silent mutations in the guide RNA or PAM sequence should be added to prevent cleavage by Cas9 after oligonucleotide sequence

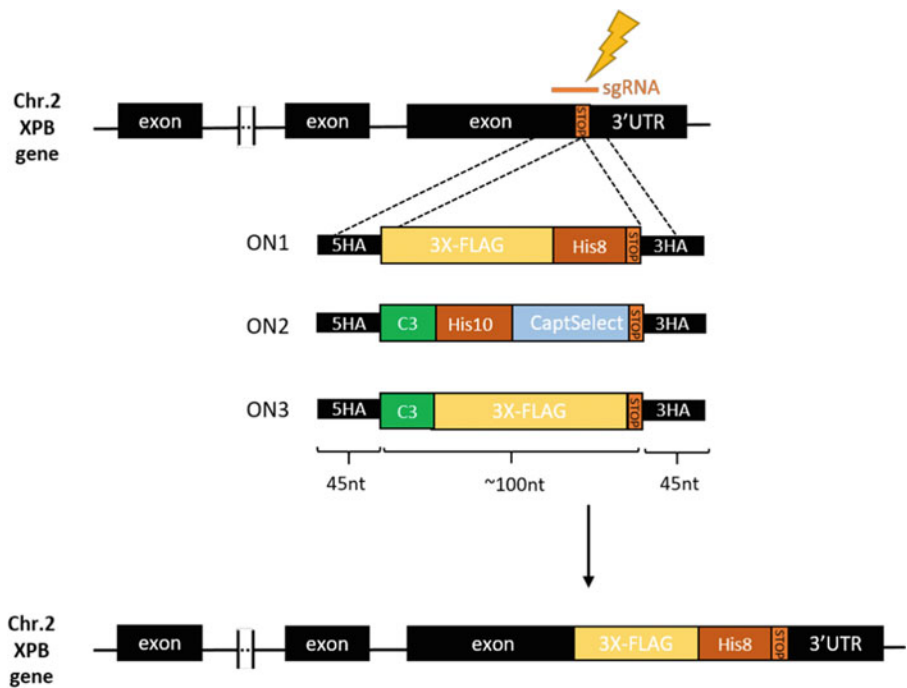


Fig. 2 Schematic of the tagging strategy for the XPB locus. CRISPR/Cas9-induced DSB takes place in vicinity of the stop codon of the XPB gene with the selected guide RNA (sgRNA). The various single-stranded donor oligonucleotides tested (ON1,ON2,ON3) contain tag sequences (3X-FLAG/His8, His10/CaptSelect, and 3X-FLAG, respectively) flanked by two homology arms (5HA, 3HA) identical to the 5'- and 3'-genomic sequences adjacent to the cleavage site, respectively. As an example, the XPB locus is shown after integration of ON1 during repair of the DSB introduced by the complex of Cas9 and the sgRNA

integration. In that case, the homology arm should be chosen to be at least 45 bases long starting from the corresponding mutation. Two phosphorothioate linkages are added at both 5' and 3' ends between the three last nucleotides to protect donor oligonucleotides from degradation by cellular exonucleases and increase genome-editing efficiency [5]. No specific additional purification step, such as PAGE purification, is required after synthesis. If possible, we also recommend testing both sense and antisense oligonucleotides for optimal efficiency.

3.3 Optimal sgRNA Selection with the Nano-Glo HiBiT Lytic Detection System

We detail here a simple and highly sensitive protocol to find the optimal guide RNA for editing the gene of interest. A standard method for testing guide RNA efficiency is to measure the proportion of mutant sequences induced at the target site with the T7 endonuclease assay. However, this method is not very sensitive and does not reliably detect mutation rates below 5%. As we could not detect robust CRISPR/Cas9 induced-cleavage assay at the XPB locus using the T7 endonuclease assay, we used a more sensitive and quicker Nano-Glo HiBiT detection system to determine the

optimal sgRNA guide. The guide RNAs are used to stimulate integration of ssODN coding for a HiBit tag-coding sequence. The Nano-Glo HiBiT Lytic Detection System is highly convenient to detect HiBiT-tagged proteins in cell lysates through luciferase detection [14]. HiBiT is an 11-amino-acid peptide tag that can be fused to the N- or C-terminal end of a protein by genome editing with the CRISPR/Cas9 system. The Nano-Glo HiBiT detection system relies on a NanoLuc enzyme, which exhibits a 150-fold increase in luminescence compared to the traditional luciferases [15]. This approach increases the sensitivity of the method and allows robust detection of HiBit-tagged proteins, even when expressed at low levels such as transcription factors. In this protocol, we used distinct oligonucleotides containing either the purification tag or the HiBiT tag. If possible, it is very convenient to introduce HiBiT and purification tags in tandem. This allows integration efficiency to be tested easily and enables the isolation of modified cells from those exhibiting the highest integration efficiency.

1. Thaw K562 cells using your favorite protocol. Cells are grown in RPMI 1640 with L-Glutamine supplemented with FBS (10%) and PenStrep (1%).
2. Split K562 cells 1/10 every 3–4 days by transferring 1 mL of suspension in a new tissue culture flask containing 9 mL of pre-warmed fresh media. We limit the number of passages to 30 to minimize genetic drift.
3. One day before nucleofection, determine the cell concentration, control that the viability is better than 95% and split cells in order to reach a cell count between 500,000–1,000,000 cells/mL.
4. Prepare DNA master mixes in Eppendorf tubes containing: 6 µg of the donor oligonucleotide HiBiT, 2 µg of guide sgRNA plasmid (sgXPB), and 2 µg of plasmid expressing the SpCas9 protein (JDS246). The total volume should be lower than 10 µL to avoid lower nucleofection efficiency.

In a control well, we recommend transfection of 6 µg of the donor oligonucleotide HiBiT on its own to measure the luminescent background signal in the experiment.

5. Count cells and centrifuge the required number of cells (1×10^6 cells per sample) at 90 g at room temperature for 10 min. Discard supernatant completely so that no residual medium covers the cell pellet. Resuspend cells in Nucleofector[®] solution V at 10^6 cells/100 µL.
6. Mix the DNA master mixes with 100 µL of cell suspension and transfer to a cuvette. Process the samples quickly to avoid storing the cells longer than 15 min in Nucleofector[®] Solution V.

7. Insert the cuvette into the Nucleofector[®], select the cell type-specific program X-001, and press the start button. Using the provided pipette, immediately remove sample from the cuvette and transfer into the 6-well plate containing 2 mL of RPMI +1% PenStrep.
8. Incubate the 6-well plate at 37 °C, 5% CO₂ for 3 days then spin down and wash the cells with PBS. Resuspend the cells in 100μL PBS and transfer to an opaque tissue culture plate to minimize absorption of the emitted light and cross-talk between wells.
9. Prepare a master mix of Nano-Glo HiBiT Lytic Detection System depending on the number of samples in your assay. Dilute the LgBiT Protein 1:100 and the Nano-Glo HiBiT Lytic Substrate 1:50 into an appropriate volume of Nano-Glo HiBiT Lytic Buffer at room temperature. Add 50μL of Nano-Glo HiBiT Lytic Reagent in each well and mix. Wait at least 15 min for equilibration of LgBiT and HiBiT in the lysate.
10. Measure luminescence using settings specific to your instrument and consider the sgRNA with the highest luminescent signal as the optimal guide RNA plasmid for the following steps in the protocol.

The results of the Nano-Glo HiBiT Lytic Detection System is shown in Fig. 3 (*See Note 1*).

3.4 Transfection of Gene Editing Reagents into K562 Cells and Ouabain Selection

After selecting the optimal sgRNA guide, we proceed to genome editing with CRISPR/Cas9 to generate cell lines that express an endogenously tagged protein of interest. To ensure efficient integration in the cell genome, we combine chemically modified single-stranded oligonucleotides as homology donors and a co-selection strategy for enrichment in homology-driven repair during

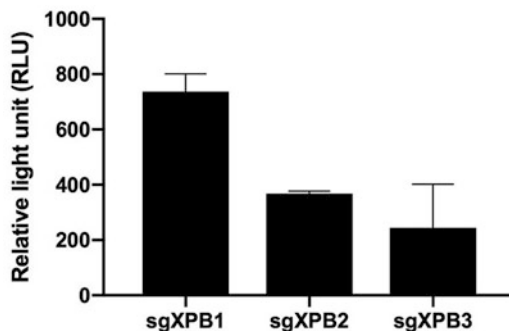


Fig. 3 Comparison of integration efficiency between the various sgRNAs targeting XPB. K652 cells were transfected with HiBiT ssODN donor, indicated guide RNA plasmid and JDS246 plasmid expressing the protein SpCas9. After 3 days, the cells were lysed using Nano-Glo HiBiT lytic detection system to measure luminescence. Higher luminescent signal correlates with higher integration of the donor HiBiT in cells

CRISPR/Cas9-mediated genome editing. This strategy generates a double integration at two genomic loci. The first integration corresponds to the tag insertion and is located downstream or upstream of the gene of interest. The second integration occurs in the ATP1A1 gene coding for the sodium-potassium pump. Introducing a specific known mutation via ssODN in the gene ATP1A1 confers to the cell resistance to ouabain, an ATP1A1 inhibitor of the enzyme responsible for sodium/potassium pump. The cells that have integrated ssODN at the ATP1A1 gene are more likely to integrate another ssODN at the locus of interest.

The selection with ouabain is performed according to the protocol described by Agudelo et al. [6] as detailed below.

1. Prepare stock solution of ouabain octahydrate (68.6 mM), by dissolving 50 mg of in 1 mL of water by heating the solution at 90 °C for 5–10 min and vortexing until complete dissolved.
2. Prepare by serial dilution 1/10, 1/100, and 1/1000 solutions. All dilution steps should be carried out in warm water to ensure the homogeneity of the solution. Add 7.3 µL of solution 1/1000 (68.6 µM) for each milliliter of media to reach a concentration of 0.5 µM.
3. In each transfection, add 6 µg of the donor oligonucleotide ON1/ON2/ON3, 2 µg of guide RNA plasmid (gRNA) vector, 4 µg of the donor oligonucleotide mutATP1A1 and 2 µg of plasmid sgG3-ATP1A1-Cas9.
4. After 72 h following nucleofection, split cells from each transfection into two 6-well plates. In both plates, add 1 mL of the cells for each condition and top up with RPMI 1640 with L-Glutamine supplemented with FBS (10%) and PenStrep (1%). In the first plate, add 15 µL of ouabain solution 1/1000 in each well. The second plate serves as a control to monitor the death of cells in the first plate (*see Note 2*).
5. Extract genomic DNA from the transfected cells using Quick-Extract DNA extraction solution according to the manufacturer's recommendations. Dilute genomic DNA at a final concentration around 50 ng/µL.
6. Amplify the locus of interest by PCR as described in Table 4. (*see Note 3*).
7. Run 2 µL of PCR products on 1% agarose gel TBE 0.5x at 100 volts for 90 min (*see Note 4*).
8. Visualize gel using a UV transilluminator.

After selection, we assess the integration of our donor oligonucleotide in the pool of transfected cells by a PCR where primers bind in the genome regions flanking the cut site.

Tag sequence insertion increases the PCR fragment size (between 75 and 100 bp for the various combination of tags used

Table 4
Composition of the PCR mix used to amplify the genomic site XPB. Use the following PCR program: 98 °C/30s 98°C/10 s 65 °C/20 s 72 °C/20 s 30 cycles 72 °C/5 min 4 °C/Hold

Genomic DNA(50 ng/μL)	1μL
dNTPs (25 mM)	0.4μL
Phusion DNA polymerase (2000 U/mL)	0.5μL
Buffer PCR 5X	10μL
Forward primer (10μM)	2.5μL
Reverse primer (10μM)	2.5μL
H ₂ O	33.1μL

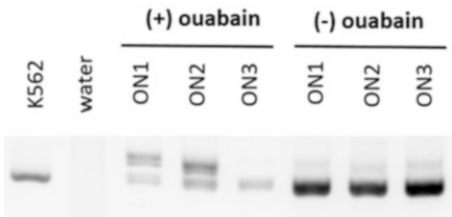


Fig. 4 PCR Analysis of the mixed cell pools. Cell pools were collected after transfection with various ssODNs (ON1, –2, or –3 as indicated), with or without ouabian selection. Targeted genomic integration of the tag-coding sequence gives rise to a PCR fragment with bigger size (upper band) compared to the wild-type PCR fragment (lower band). Ouabain selection significantly increased the proportion of cells with integration of tag-coding sequences for ON1 and ON2

here) and can be readily compared between different experimental conditions (e.g., comparing sense and antisense donor oligonucleotides). An example of the analysis is shown in Fig. 4. If the proportion of PCR products with tag insertion is low (corresponding to less than 5% of the PCR products), the proportion of cells carrying the modified allele is probably too low to consider isolating a clone of modified cells by single-cell cloning. The protocol may best be repeated using another donor oligonucleotide design or after further optimization of transfection conditions.

**3.5 Isolation
of Gene-Edited Cells by
Single-Cell Cloning**

Different approaches can be used to isolate clonal populations of gene-edited cells. Cell types can vary greatly in their responses to FACS or serial dilution. For K562 cells, we used successfully both methods.

1. Sort single cells into 96-well plate with your favorite method (see Note 5).

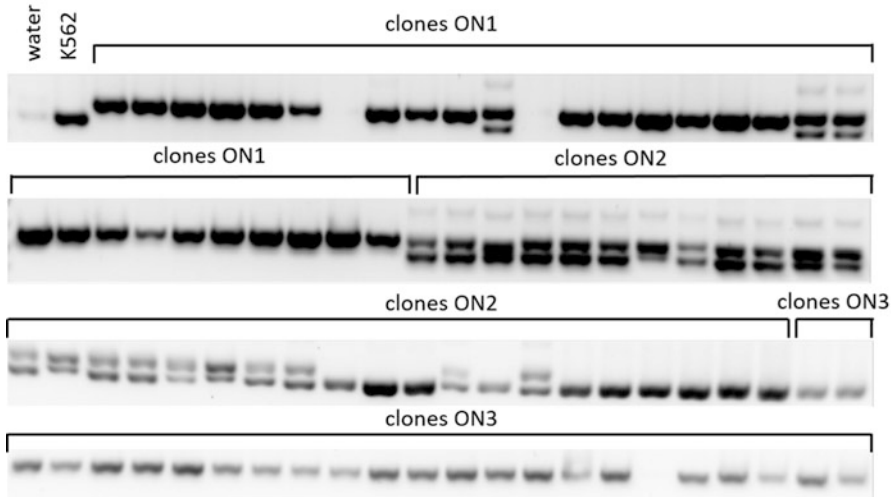


Fig. 5 Genotyping of the clones. Integration of ssODN donor at the XPB locus gives rise to a bigger PCR product (upper band) compared to wild-type PCR product (lower band). The majority of clones generated with ssODN ON1 are homozygous for integration whereas the majority are heterozygous with ssODN ON2. None of the clones generated with ON3 integrated the ssODN

2. Two weeks after cell sorting, visually screen plates using a microscope to identify the wells containing cells colonies. Split cells from each positive well into two 96-well plates (P1 and P2).
3. The day after, spin down and wash with PBS the cells from the plate P1. Extract genomic DNA directly in the wells using 25 μ L of commercially available DNA lysis buffer.
4. Proceed to PCR amplification for each clone following a similar protocol to the PCR analysis of the pool. Run 2 μ L of the PCR on 1% agarose TBE 1x gel at 100 volts for 1 h. Visualize gel using a UV transilluminator. In this genotyping, both knock-in and non-knock-in alleles could be amplified, showing whether both, only one or none of the alleles were modified to generate either homozygote or heterozygote clones. An example of gel is shown (Fig. 5).
5. Sequence the PCR fragment of positive clones to ensure correct sequence tag insertion. For homozygous clones, Sanger sequencing chromatograms display only one sequence. For heterozygous clones, Sanger sequencing chromatograms will contain mixtures of sequences that can be de-convoluted separately using online tool CRISP-ID (<http://crispid.gbiomed.kuleuven.be/>).
6. Expand positive clones from plate P2 and prepare frozen stocks, e.g., 10 batches of 10×10^6 frozen cells.

3.6 Validation of Gene Tagging by Western Blot

Having isolated a clonal population of modified cells, expression of the tagged gene is validated at the protein level. We generally use approximately 2.5×10^7 cells and validate the presence of the tagged protein either directly in a whole cell extract or after immunopurification.

3.6.1 Preparation of Cell Pellets

1. Seed several T175 flasks containing 25 mL RPMI 1640 supplemented with 10% FBS, L-Glutamine, and antibiotics.
2. Harvest cells by centrifugation at 1200 g for 10 min at 4 °C, wash twice with ice-cold PBS containing 30% w/v glycerol, snap freeze in liquid nitrogen, and stored at –80 °C. We usually prepare batches of 1.5×10^6 and 2.5×10^6 cells.

3.6.2 Analysis of Tagged Proteins in the Whole Cell Extract

1. Resuspend 1.5×10^6 cells in 150 µL of RIPA buffer and incubate for 10 min with periodic pipetting.
2. Sonicate 2 times 30 s on ice. We use a Bioruptor™ sonication system (amplitude 30 and 0.5 s pulse on ice) (Optional).
3. Centrifuge 15 min at 14000 g at 4 °C, collect the supernatant, and estimate the total protein concentration using a Bradford assay. A concentration of 3–4 mg/mL is expected.
4. Heat 20 µg of total protein from the soluble extract prepared in Laemmli buffer at 95 °C for 5 min and centrifuge samples at 10,000 g for 30 s to bring down the condensate and remove insoluble debris.
5. Load the centrifuged sample on a polyacrylamide gel, electrophorese, and transfer proteins from the gel matrix to a solid-support membrane using your favorite device (*see Note 6*).
6. Block the membrane for 1 h at room temperature or overnight at 4 °C, incubate the membrane with an appropriate dilution of a primary antibody directed against the affinity-tag or the tagged subunit in the same buffer for 1 h at room temperature or overnight at 4 °C.
7. Wash the membrane three times for 5 min in TBST, incubate with the recommended dilution of conjugated secondary antibody in TBST at room temperature for 1 h.
8. Develop the western blot.

3.6.3 Analysis of the Target Complex after Immunopurification

1. Resuspend 2.5×10^6 cells in 1 mL of lysis buffer (*see Note 7*) and sonicate 2 times 30 s on ice with a Bioruptor (amplitude 30 and 0.5 s pulse on ice).
2. Centrifuge 15 min at 14,000 g, transfer the supernatant into an Eppendorf tube containing 20 µL of anti-FLAG affinity resin, and incubate for 3–4 h with gentle agitation. All manipulations should be performed on ice or at 4 °C.

3. Centrifuge at 1000 g for 2 min to sediment the resin and discard the supernatant. Resuspend the resin in 1 mL of lysis buffer and pellet the beads again. Repeat this washing step twice and carefully remove the supernatant using a pipette with a narrow-end pipette tip.
4. For analysis under denaturing conditions, elute bound proteins by addition of: 60 μ L of Laemmli buffer. For analysis in native condition, add 60 μ L of lysis buffer containing 0.5 mg/mL FLAG peptide or 5 μ g/mL of PreScission protease. After incubation for 12 h at 4 °C under gentle agitation, sediment the resin by centrifugation, and mix 45 μ L of the eluate with 15 μ L of 4X Laemmli buffer.
5. Heat the Laemmli-containing eluate at 95 °C for 5 min, and after centrifugation at 10000 g for 30 s load the sample on a polyacrylamide gel and perform a western blot analysis as detailed in Subheading 3.7.2 (*see Note 8*).
6. The result of an experiment in which the gene encoding the XPB subunit of the TFIIH transcription/DNA repair was edited to add a C-terminal PreScission protease cleavage site (3C) followed by a 3X-FLAG peptide is shown (Fig. 6).

3.7 Suspension Cultures and Purification

After validation of the modified cell line, one clone is selected for scale-up and purification. We generally use batches of 10^9 cells for the first purification trials and adapt subsequently. In this section, we detail large-scale suspension cultures and immunopurification.

3.7.1 Large-Scale Suspension Cultures

1. Thaw a vial of a positive clone using your favorite protocol. As above, cells are grown in RPMI 1640 with L-Glutamine supplemented with 10% fetal bovine serum (FBS) and 1% PenStrep.
2. Expand cells in a tissue culture flask until a total of 15×10^6 cells with a minimum viability of 95% is reached.
3. Transfer cells in a sterile 250 ml glass bottle and dilute with fresh medium to a cell density $\sim 0.2\text{--}0.4 \times 10^6$ cells/mL in a total volume of 40–60 mL, cover the opening with gas permeable adhesive seal and incubate at 37 °C, 5% CO₂, 75% humidity under agitation (90 rpm).
4. After 24–48 h, when a cell density of $\sim 0.7\text{--}1.0 \times 10^6$ cells/mL is reached transfer cells into a 500 ml glass bottle and dilute to a cell density $\sim 0.2\text{--}0.4 \times 10^6$ cells/mL in a total volume of 80–120 mL and incubate as above.
5. Repeat **step 4** to seed successively a 1000 ml bottle with a working volume of 150–250 ml and four 5000 ml bottles with working volumes of 750–1250 mL.

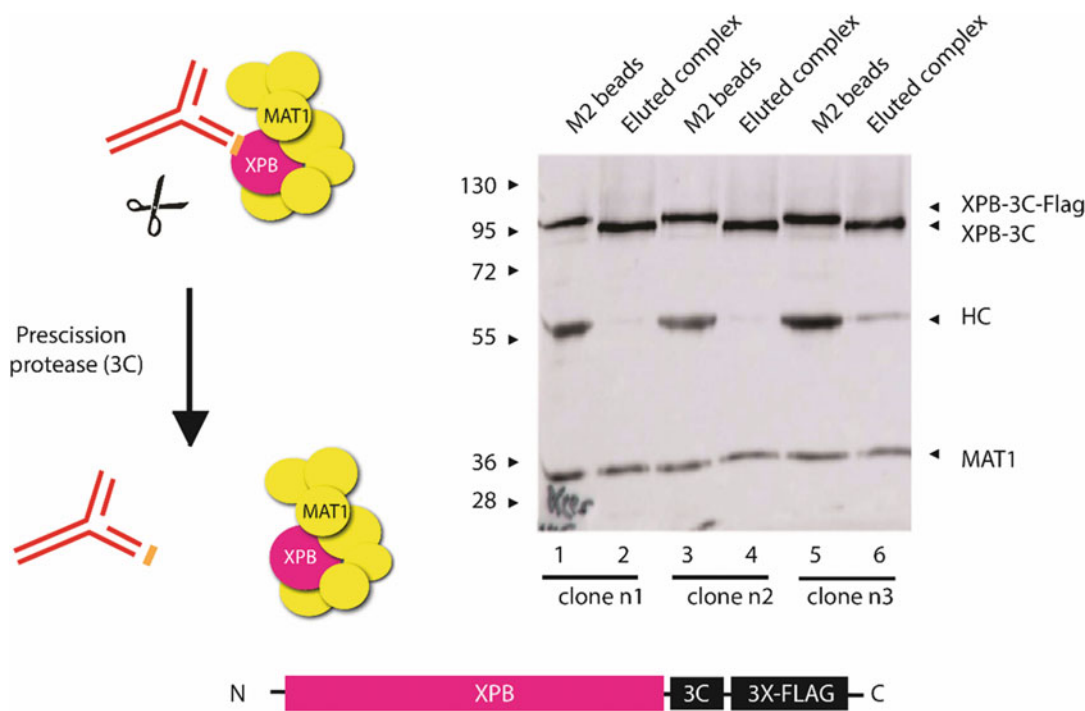


Fig. 6 Validation of engineered cell lines by western blot analysis. The gene encoding the XPB subunit of the TFIIF transcription/DNA repair was edited to fuse its C-terminus to a PreScission protease cleavage site (3C) followed by a 3X-FLAG peptide. For validation of selected clones, the corresponding soluble extracts were incubated with anti-FLAG affinity resin, and after extensive washing, bound proteins were eluted in denaturing conditions with Laemmli buffer (lanes 1, 3, and 5). Alternatively, immobilized complexes can be eluted in native condition by addition of FLAG competitor peptide (not shown) or PreScission protease to cleave the tag (lanes 2, 4, and 6). Immuno-purified proteins were resolved using a 12% SDS-polyacrylamide gel and subjected to western blot analysis. Antibodies directed against XPB and MAT1, another TFIIF subunit were used for protein detection

6. When the cell density of $\sim 0.7\text{--}1.0 \times 10^6$ cells/mL is reached in the 5000 mL bottles add fresh medium and incubate further.
7. Harvest cells by centrifugation at 1000 g for 10 min at 4 °C, wash twice with ice-cold PBS containing 30% w/v glycerol, snap freeze in liquid nitrogen and store at -80°C . We usually prepare batches of 10^9 cells.

A typical amplification workflow is shown in Fig. 7a. As modified cell lines can be unstable and prone genetic drift, we carefully monitor the number of passages and always start with a freshly thawed vial.

3.7.2 Immunopurification

1. Resuspend one batch of 10^9 cells in 20 mL of lysis buffer (see Note 7) and sonicate 4 times 30 s on ice with a Bioruptor (amplitude 30 and 0.5 s pulse on ice).

2. Add DNase and RNase to obtain a final concentration of 12.5 µg/mL for both nucleases and incubate for 20 min at room temperature under gentle agitation.
3. Centrifuge 60 min at 50,000 g.
4. Load the extract on a 5 ml heparin column and collect the fractions containing the complex of interest.
5. Incubate the extract or the fractions containing the complex with 50 µl of anti-FLAG affinity resin during 3–4 h under gentle agitation in a 15 ml Falcon tube. All manipulations should be performed on ice or at 4 °C.
6. Centrifuge at 1000 g for 2 min to sediment the resin and discard the supernatant. Resuspend the resin in 1 mL of lysis buffer, transfer into a 1.5 mL Eppendorf tube and pellet the beads again. Repeat this washing step twice and carefully remove the supernatant using a pipette with a narrow-end pipette tip.
7. Add 50 µL of lysis buffer containing 0.5 mg/mL FLAG peptide and incubate for 12 h at 4 °C under periodic agitation (30s every 5 min).
8. Collect the supernatant and repeat **step 7** with an incubation of 2 h and pool the two eluates.
9. Mix 20 µL of the eluate with 5 µL of 4X Laemmli buffer and analyze on an SDS-PAGE (12.5% acrylamide). The corresponding gel is shown in Fig. 7b.
10. Aliquot and snap freeze the purified sample.

4 Notes

1. The relative luciferase values detected here with Nano-Glo HiBiT Lytic Detection System were lower than in the other experiments performed at various genomic loci and in other cell lines. This is consistent with the results obtained in T7 endonuclease assay that indicate a very low activity of sgRNAs targeting the stop codon of the XPB gene.
2. Within a few days, only the cells that have gained ouabain resistance will proliferate. Nonetheless, subsequent splitting of the cells is essential to avoid interference in molecular analyses from the genomic DNA of dead cells. After 2 weeks of ouabain treatment, the cell population is considered stable.
3. Suitable PCR primers to the target region can be designed with CRISPOR by clicking on the “Cloning/PCR primers” hyper-link for the selected guide RNA on the CRISPOR results web page or using the online tool Primer3 (<http://www.bioinfo.ut>).

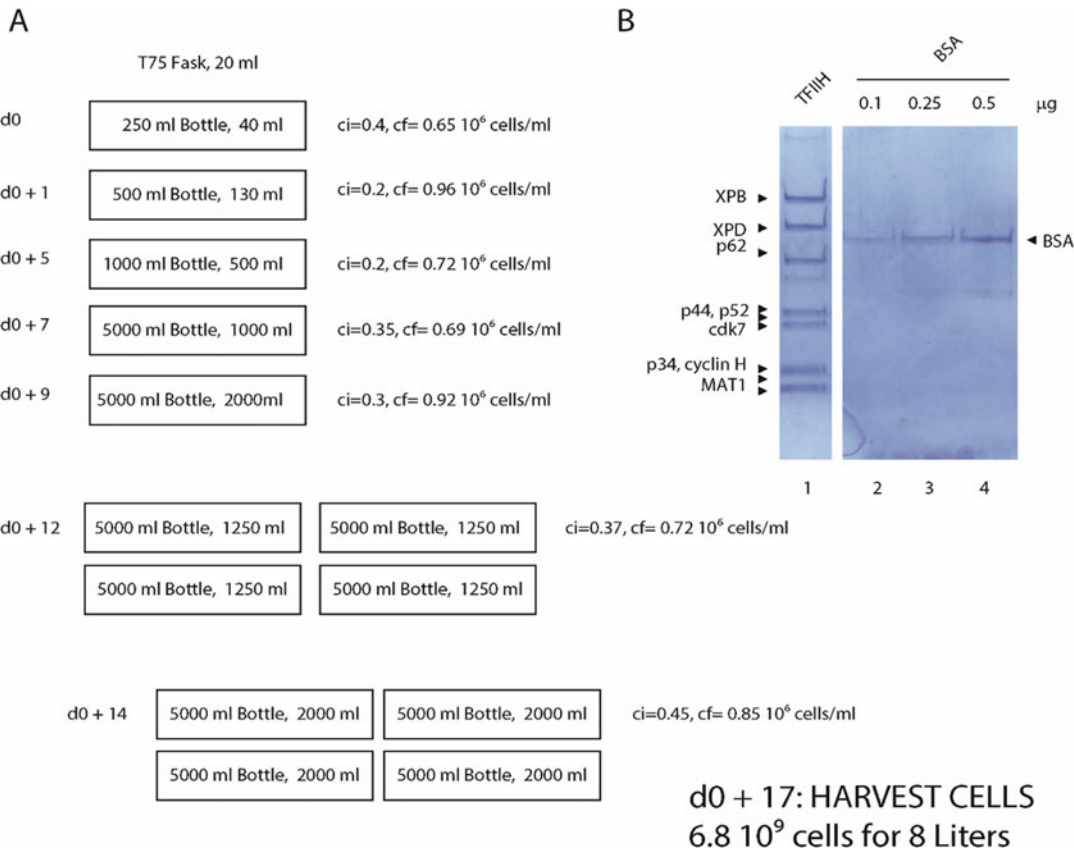


Fig. 7 Large-scale cultures and immunopurification. (a) Freshly thawed cells of a representative modified K562 cell line were expanded in a T75 tissue culture flask and then successively transferred into 250, 500, 1000, and 5000 mL glass bottles. Initial and final cell densities (ci and cf., respectively) are indicated. (b) Purified TFIIH resolved on a Coomassie-stained 12.5% SDS-polyacrylamide gel

- [cc/primer3/](#)). PCR primers should be designed that flank the DSB site within a total length of between 300 and 800 bp.
4. Make sure to run the electrophoresis for a sufficient length of time because the separation between the PCR fragments is challenging due to the short length of the purification tag sequence relative to the amplified DNA region.
 5. Some adherent cell lines do not tolerate single-cell dilution in 96-well plates and may need using conditioned medium or can alternatively be dilution plated in 10 cm cell culture dishes.
 6. For most proteins, we run a 12.5% a 1 mm thick polyacrylamide gel for 1 h 30 (35 mA) with a TGS running buffer (25 mM Tris-HCl pH 8.3, 190 mM glycine, 0.1% SDS). Do not omit to include a sample prepared from a non-modified cell line as negative control.

7. The composition of the RIPA and lysis buffers can be adjusted by using a few detergents (ionic and nonionic detergent) and different salt concentrations (low, medium, and high salt). High ionic strength enhances solubility of many proteins but can lead to complex dissociation. For intracellular proteins, care should be taken to maintain a reducing environment.
8. As analyses were performed in native conditions, antibodies directed against the affinity-tag or the tagged protein as well as against another subunit of the complex can be used.

Acknowledgments

This work was supported by the Centre National pour la Recherche Scientifique (CNRS), the Institut National de la Santé et de la Recherche Médicale (INSERM), Université de Strasbourg (UdS), Association pour la Recherche sur le Cancer (ARC), the Ligue nationale contre le cancer, Institut National du Cancer (INCa; INCA 9378), Agence National pour la Recherche (ANR-12-BSV8-0015-01 and ANR-10-LABX-0030-INRT under the program Investissements d'Avenir ANR-10-IDEX-0002-02), Instruct-ERIC (R&D Project Funding) and by Instruct-ULTRA (Coordination and Support Action Number ID 731005) funded by the EU H2020 framework to further develop the services of Instruct-ERIC.

References

1. Dalvai MJ, Loehr K, Jacquet CC et al (2015) Scalable genome-editing-based approach for mapping multiprotein complexes in human cells. *Cell Rep* 13:621–633
2. Doudna JA, Charpentier E (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346:1258096
3. Jinek MK, Chylinski I, Fonfara M et al (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337:816–821
4. Her J, Bunting SF (2018) How cells ensure correct repair of DNA double-strand breaks. *J Biol Chem* 293:10502–10511
5. Renaud JB, Boix C, Charpentier M et al (2016) Improved genome editing efficiency and flexibility using modified oligonucleotides with TALEN and CRISPR-Cas9 nucleases. *Cell Rep* 14:2263–2272
6. Agudelo D, Düringer A, Bozoyan CC et al (2017) Marker-free co-selection for CRISPR-driven genome editing in human cells. *Nat Methods* 14:615–620
7. Santiago YE, Chan PQ, Liu S et al (2008) Targeted gene knockout in mammalian cells by using engineered zinc-finger nucleases. *Proc Natl Acad Sci U S A* 105:5809–5814
8. Kim H, Kim JS (2014) A guide to genome engineering with programmable nucleases. *Nat Rev Genet* 15:321–334
9. Giraud GR, Stadhouders A, Conidi D et al (2014) NLS-tagging: an alternative strategy to tag nuclear proteins. *Nucleic Acids Res* 42:e163
10. Oesterle S, Roberts TM, Widmer LA et al (2017) Sequence-based prediction of permissive stretches for internal protein tagging and knockdown. *BMC Biol* 15:100

11. Kolesnikova O, Radu L, Poterszman A (2019) TFIIH: a multi-subunit complex at the crossroads of transcription and DNA repair. *Adv Protein Chem Struct Biol* 115:21–67
12. Kimple ME, Brill AL, Pasker RL (2013) Overview of affinity tags for protein purification. *Curr Protoc Protein Sci* 73:9.9.1–9.9.23
13. Schwinn MK, Machleidt T, Zimmerman K et al (2018) CRISPR-mediated tagging of endogenous proteins with a luminescent peptide. *ACS Chem Biol* 13:467–474
14. England CG, Ehlerding EB, Cai W (2016) NanoLuc: a small luciferase is brightening up the field of bioluminescence. *Bioconjug Chem* 27:1175–1187
15. Wood DW (2014) *Curr Opin Struct Biol.* Jun;26:54–61. <https://doi.org/10.1016/j.sbi.2014.04.006>. Epub 2014 Jun 12. PMID: 24859434



Chapter 9

Synthesis of Fluorescently Labeled Antibodies Using Non-Canonical Amino Acids in Eukaryotic Cell-Free Systems

Marlitt Stech, Nathanaël Rakotoarinoro, Tamara Teichmann,
Anne Zemella, Lena Thoring, and Stefan Kubick

Abstract

Cell-free protein synthesis (CFPS) enables the development of antibody conjugates, such as fluorophore conjugates and antibody-drug conjugates (ADCs), in a rapid and straightforward manner. In the first part, we describe the cell-free synthesis of antibodies containing fluorescent non-canonical amino acids (ncaa) by using pre-charged tRNA. In the second part, we describe the cell-free synthesis of antibodies containing ncaa by using an orthogonal system, followed by the site-specific conjugation of the fluorescent dye DyLight 650-phosphine. The expression of the antibodies containing ncaa was analyzed by SDS-PAGE, followed by autoradiography and the labeling by in-gel fluorescence. Two different fluorescently labeled antibodies could be generated.

Key words Cell-free protein synthesis, Antibody, Antibody conjugates, IgG1, Non-canonical amino acid, Conjugation

1 Introduction

ADCs represent one of the most promising strategies in the pharmaceutical industry to treat solid and hematological cancers. Up to date, nine of them are already approved [1]. The concept of ADCs combines the effect of highly specific tumor-targeting immunotherapy with the concept of chemotherapy relying on highly cytotoxic drugs. ADCs are considered as a strategy to provide a better cytotoxicity to immunotherapy and a better specificity to chemotherapy. ADCs are composed of a monoclonal antibody conjugated to a cytotoxic drug by a linker. In the early stages of the development of ADCs, non-toxic fluorescent agents can be used to optimize the conjugation efficiency, to analyze the internalization efficiency, and to locate and quantify the antibody *in vitro* and *ex vivo*, prior to the conjugation to the cytotoxic agent.

While full-length Immunoglobulin G (IgG) is mainly produced in the mammalian system [2], such as Chinese Hamster Ovary

(CHO) cell lines, the drug is produced either by hemi-synthesis or by chemical synthesis. One strategy to conjugate the latter to the former is the amber suppression technology [3–6]. This technology is based on the introduction of an amber stop codon into the gene sequence at a desired position. By adding an engineered tRNA and synthetase the ncaa is incorporated in the antibody sequence during its synthesis exactly at the position of the amber stop codon. Following protein synthesis, the drug containing the corresponding reactive group can be conjugated to the antibody by the reactive group of the ncaa. The basis of the amber suppression technology is an orthogonal tRNA/synthetase pair. The orthogonal synthetase is engineered to amino-acylate specifically the ncaa at the 3'-end of the orthogonal tRNA. The latter is engineered to recognize the amber stop codon, thus allowing for the incorporation of the ncaa in a site-specific manner. Most importantly, the orthogonal tRNA/synthetase pair should not show cross-reactivities between endogenous amino acids, tRNAs, and synthetases.

Orthogonal systems and the resulting products can be developed by using a CFPS system [5]. The CFPS system based on CHO lysate contains microsomes derived from the endoplasmic reticulum of CHO cells as previously described [7]. By translocating *de novo* synthesized proteins into the lumen of these microsomes by using a melittin signal peptide, post-translational modifications such as disulfide bridge formation and glycosylation can be performed [7]. Using this system, full-length post-translationally modified antibodies can be produced within hours, making the screening of orthogonal tRNAs, or synthetases, ncaa, antibody candidates, and ADCs rapid and straightforward [5].

In this chapter, we describe the proof-of-concept for the cell-free synthesis of antibodies containing ncaa. To allow for the site-specific introduction of ncaa, the chosen model antibody contained an amber stop codon at amino acid position 134 in the CH1 domain of antibody heavy chain (HC). In the first part, we expressed antibodies containing a Bodipy-TMR-lysine, pre-charged on a tRNA. Autoradiography and in-gel fluorescence analysis showed the expression of the fluorescently labeled antibody of interest. In the second part, we expressed antibodies containing p-azido-L-phenylalanine (AzF) by using an orthogonal system, composed of an engineered *E. coli* tRNA [8] amino-acylated by an engineered *E. coli* tyrosine synthetase [9]. Autoradiography confirmed the synthesis of suppression and full-length product. Subsequently, the fluorophore DyLight 650-phosphine was conjugated to the antibody by Staudinger ligation. The successful conjugation was shown by in-gel fluorescence.

2 Materials

Prepare all buffers and solutions using ultrapure water.

2.1 CFPS Using CHO Lysate

1. Ice pan.
2. 1.5 mL reaction tubes.
3. 10× translation mix: 300 mM HEPES-KOH (pH 7.6), 2250 mM KOAc, 2.5 mM spermidine, 1 mM of each canonical amino acid, and 39 mM Mg(OAc)₂. Store at −80 °C.
4. CHO lysate prepared as described previously [10, 11] (*see Note 1*). Shock-freeze in liquid nitrogen after every usage and store at −80 °C.
5. T7 RNA polymerase.
6. 5× energy mix: 100 mM creatine phosphate, 1.5 mM GTP, 1.5 mM CTP, 1.5 mM UTP, 8.75 mM ATP, and 0.5 mM m⁷G (ppp)G cap analog. Store at −80 °C.
7. ¹⁴C-leucine (200 dpm/pmol, 100 dpm/pmol). Store at −20 °C.
8. Plasmid encoding antibody HC, plasmid encoding antibody light chain (LC) (Fig. 1) (*see Note 2*). Store at −20 °C.
9. Ultrapure water.
10. Thermomixer.
11. Phosphate-buffered saline (PBS) containing 0.2% *n*-Dodecyl β-D-maltoside (DDM). Store at 4 °C.

2.2 Preparation of Orthogonal Synthetase

1. Ice pan.
2. 1.5 mL reaction tubes.
3. DNA template encoding modified tyrosyl-tRNA-synthetase (eAzFRS, including the mutations Thr37, Ser182, Ala183, and Arg265 [9] and a C-terminal Strep-Tag) from *E. coli*, cloned into a vector containing a T7 promotor. We used the vector pQE2 vector (pQE2-eAzFRS-SII) (*see Note 3*).
4. *E. coli* expression system (RTS 500 *E. coli* HY Kit, Biotech rabbit).
5. Thermomixer with RTS 500 thermomixer adapter.
6. 100 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG).
7. Gravity flow Strep-Tactin[®] superflow mini-column (0.2 mL).
8. Strep-Tactin[®] Purification Buffer Set: 10× Washing Buffer (1 M Tris-Cl, pH 8.0, 1.5 M NaCl, 10 mM EDTA), 10× Elution Buffer (1 M Tris-Cl, pH 8.0, 1.5 M NaCl, 10 mM EDTA, 25 mM Desthiobiotin) and 10× Regeneration Buffer

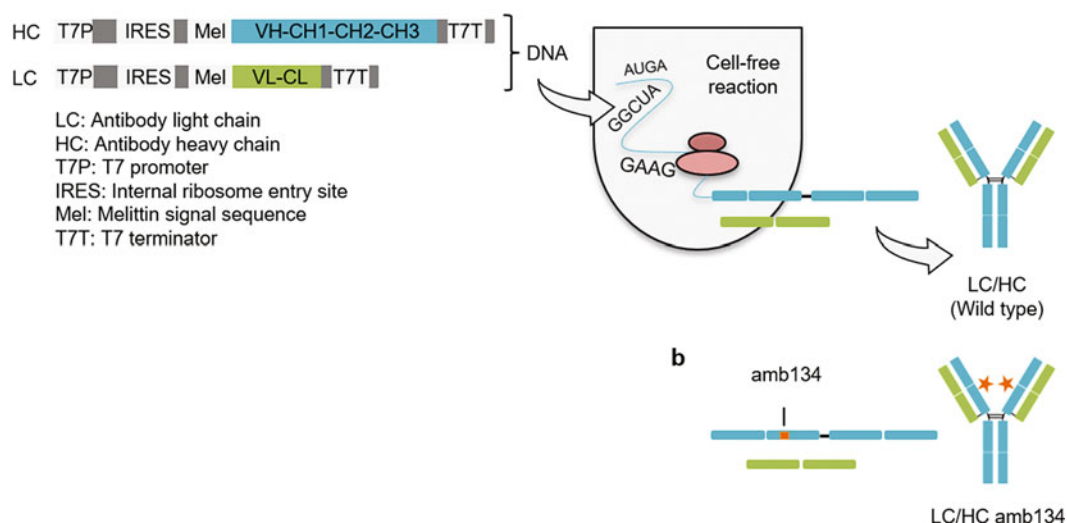
a

Fig. 1 Schematic presentation of the template design of antibody heavy (HC) and light chain (LC) and their rapid cell-free synthesis and assembly to functional IgG. **(a)** Template design without amber (amb) stop codon. **(b)** Template design with amber stop codon to allow for the site-specific incorporation of a ncaa into cell-free synthesized antibody HC. The amb stop codon TAG was positioned in the CH1 domain (replacing S134, EU numbering) of HC. Orange asterisks indicate fluorescent dye conjugated to the incorporated ncaa.

(1 M Tris-Cl, 1.5 M NaCl, 10 mM EDTA, 10 mM HABA (hydroxyl-azophenyl-benzoic acid)).

9. Zeba™ Spin Desalting Columns (7 K MWCO, 0.5 mL).
10. Amicon® Ultra Centrifugal Filters (10 K device, 0.5 mL).
11. Synthetase storage buffer: 50 mM HEPES pH 7.6, 10 mM KOAc, 1 mM MgCl₂, 4 mM DTT.
12. NanoDrop 2000c.

2.3 Preparation of Orthogonal tRNA

2.3.1 PCR Amplification of the tRNA Gene

1. Vector containing the nucleotide sequence of tRNA^{Tyr}CUA (SupF Gene).
2. tRNA^{Tyr}CUA-specific forward primer (5' CgA gCT CgC CCA CCg gAA TTC 3') and 2'-OMe reverse primer (5' Tgg Tgg Tgg ggg AAg gAT TCg 3').
3. 0.2 mL PCR tubes.
4. PCR cyclor.
5. Taq DNA polymerase.
6. Taq buffer.
7. dNTPs.
8. 25 mM MgCl₂.
9. Agarose gel electrophoresis chamber.

10. Agarose.
11. Rotiphorese 10 × TBE buffer.
12. DNA stain.
13. DNA ladder.
14. PCR Purification Kit.

**2.3.2 Transcription,
Isolation, and Folding
of tRNA**

1. 5× transcription buffer: 400 mM HEPES-KOH, 0.5 mM spermidine, 50 mM DTE and 75 mM MgCl₂.
2. 5 × NTP mix containing 18.75 mM ATP, 18.75 mM CTP, 18.75 mM UTP, and 7.5 mM GTP.
3. T7 RNA Polymerase.
4. DNaseI (1 U per µg plasmid DNA).
5. 10× MOPS buffer: 200 mM MOPS, 50 mM NaOAc, 10 mM EDTA, pH 8.0.
6. MOPS sample buffer: 8% (v/v) formaldehyde, 12 mL formamide, 2.4 mL 10× MOPS buffer, 0.05% (v/v) bromophenol blue to a total volume of 24 mL.
7. TRIzol reagent.
8. High Performance Liquid Chromatography (HPLC) grade Chloroform.
9. HPLC grade isopropanol.
10. 75% ethanol.
11. Cooled centrifuge.
12. NanoDrop 2000c.
13. PCR cycler.

**2.4 Site-Specific
Incorporation
of Non-canonical
Amino Acids**

1. 100 µM Bodipy-TMR-lysine-tRNACUA (Biotech rabbit). Store at −80 °C.
2. 100 µM eAzFRS. Store at −80 °C.
3. 100 µM tRNATyrCUA. Store at −80 °C.
4. 100 mM AzF (Bachem AG; Bubendorf, Schweiz). Store at −80 °C.
5. Phosphate-buffered saline (PBS). Store at 4 °C.
6. PBS containing 0.2% DDM. Store at 4 °C.

**2.5 Qualitative
Protein Analysis**

1. SDS-PAGE sample buffer: 1× LDS buffer containing 106 mM Tris HCl, 141 mM Tris base, 2% LDS, 10% glycerol, 0.51 mM EDTA, 0.22 mM SERVA Blue G, 0.175 mM Phenol Red, pH 8.5.
2. 3–8% Tris acetate gels.

3. Fluorescently labeled protein ladder for SDS-PAGE.
4. SDS-PAGE running buffer: 50 mM MES, 50 mM Tris Base, 0.1% SDS, 1 mM EDTA, pH 7.3.
5. SDS-PAGE gel tank system.
6. Radioactive ink.
7. Acetone.
8. Water bath.
9. Fluorescence/phosphorimager.
10. Gel dryer.
11. Phosphor screens.

3 Methods

3.1 Cell-free Synthesis and Fluorescence Labeling of IgG Using Pre-Charged tRNA

3.1.1 Batch-Based CFPS

Batch-based cell-free reactions are carried out as coupled transcription-translation reaction, in which transcription and translation take place simultaneously in the same reaction compartment (“one-pot”).

1. Thaw all components of the cell-free reaction on ice. Mix all components thoroughly before usage. Protect pre-charged tRNA Bodipy-TMR-lysine-tRNACUA from light (*see Note 4*).
2. Pipet the following components on ice using a 1.5 mL reaction vessel: 5 μ L 10 \times translation mix (f.c. 1 \times), 20 μ L CHO lysate (f.c. 40%), 1 μ L T7 RNA polymerase (f.c. 1 U/ μ L), and 10 μ L 5 \times energy mix (f.c. 1 \times). Mix thoroughly after addition of each component (*see Note 5*).
3. Add 2.5 μ L of 200 dpm/pmol 14 C-leucine (specific radioactivity of 66.67 dpm/pmol f.c.) for subsequent qualitative analysis by autoradiography.
4. For fluorescence labeling, supplement the cell-free reaction with 2 μ M Bodipy-TMR-lysine-tRNACUA. Protect the translation mixture from light during pipetting and incubation.
5. Add HC and LC encoding plasmid at a final concentration of 60 nM each (*see Note 6*).
6. Adjust the final volume of the reaction mix with ultrapure water to 50 μ L (*see Note 7*).
7. Mix all components thoroughly and incubate the reaction at 27 °C and 500 rpm for 3 h in a thermomixer (*see Note 8*). After completing the cell-free reaction place reaction vessels on ice and proceed with the procedure described in Subheadings 3.2 and 3.4 (Fig. 2).

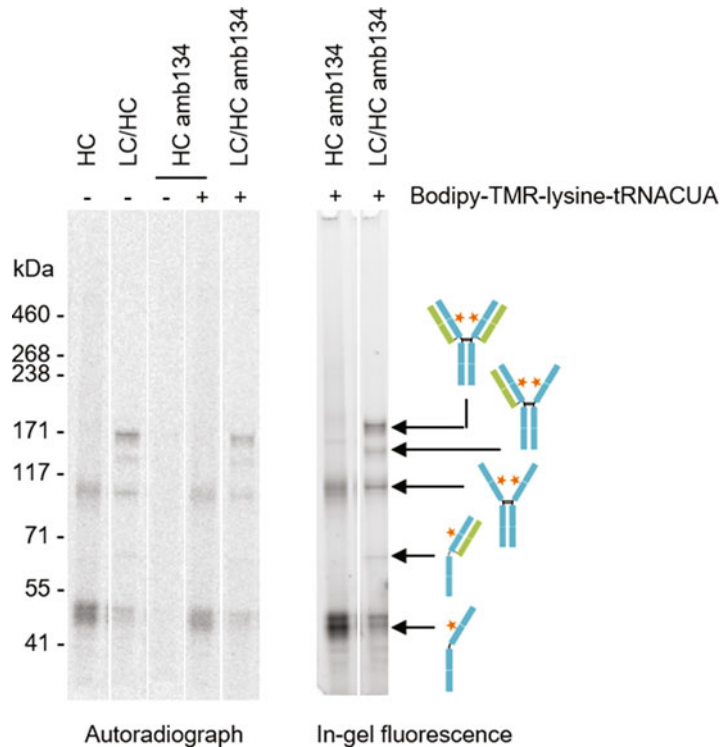


Fig. 2 Batch-based cell-free synthesis and site-specific fluorescence labeling of antibodies with Bodipy-TMR-lysine-tRNACUA by amber (amb) suppression. Qualitative analysis of cell-free synthesized antibody heavy chain (HC) and light chain (LC) was performed by SDS-Page followed by autoradiography (left side) and in-gel fluorescence (right side) (analysis of MF). Cell-free synthesis was performed in the presence of ^{14}C -leucine and in the presence (+) or absence (–) of Bodipy-TMR-lysine-tRNACUA. The amb stop codon TAG was positioned in the CH1 domain (replacing S134, EU numbering) of HC. Orange asterisks indicate fluorescent dye conjugated to the incorporated ncaa. Unassembled LC (25.4 kDa) and termination product of HC amb134 (16.4 kDa) cannot be visualized in the autoradiograph because of its low molecular weight

3.2 Fractionation of Translation Mixtures

1. Centrifuge translation mixtures at $16,000\times g$ for 10 min at $4\text{ }^{\circ}\text{C}$. Take off the supernatant (SN1) and discard.
2. Carefully resuspend the microsomal pellet in PBS supplemented with 0.2% DDM by pipetting up and down for several times. Incubate the solution for 45 min at room temperature (RT) under intense agitation (*see* **Note 9**).
3. Centrifuge the solution at $16,000\times g$ for 10 min at $4\text{ }^{\circ}\text{C}$ to separate soluble antibodies, located in the supernatant (supernatant 2, SN2), from the microsomes. Take off the supernatant (containing soluble antibodies) and place it on ice. Analyze cell-free synthesized antibodies as described in Subheading 3.4.

3.3 Cell-free Synthesis and Fluorescence Labeling of IgG Using Orthogonal System

3.3.1 Preparation of Orthogonal Synthetase

1. Enhanced tRNA synthetase is produced using the *E. coli*-based CECF-system “RTS500 ProteoMaster *E. coli* HY Kit.”
2. First, reconstitute the buffers and the *E. coli* lysate in reconstitution buffer according to the manufacturer’s instructions. Pipet everything on ice and mix the buffers and solutions thoroughly.
3. Prepare the reaction mixture as follows: Pipet 525 μL *E. coli* lysate, 225 μL reaction mix, 270 μL amino acid mix without methionine, 30 μL methionine, 11 μL IPTG and 39 μL template pQE2-eAzFRS-SII containing 110 μg plasmid DNA, and mix the solution thoroughly.
4. For the feeding mixture, pipet 7990 μL feeding mix, 110 μL IPTG, 2650 μL amino acid mix without methionine, and 300 μL methionine. Mix the solution thoroughly.
5. Fill the reaction chamber (red lid) with the complete volume of reaction mix and the feeding chamber (colorless lid) with the complete volume of feeding mix. Insert the chamber into the RTS 500 adapter in a thermomixer. Incubate the reaction at 30 °C for 24 h at 1000 rpm. Harvest the reaction mix from the reaction chamber and place the reaction mixture on ice.
6. In order to separate soluble from insoluble protein, centrifuge the translation mix at $16,000 \times g$ for 10 min at 4 °C. Harvest the supernatant containing soluble eAzFRS by pipetting.
7. Subsequently, eAzFRS is purified via its C-terminal Strep-Tag. Purification is performed using Strep-Tactin Gravity Flow Columns (200 μL).
8. Equilibrate each column with $2 \times 800 \mu\text{L}$ washing buffer and add 500 μL of the supernatant from this step to each column.
9. Once the supernatant has completely entered the column, wash each column $5 \times$ with 200 μL washing buffer (*see Note 10*).
10. To elute the synthetase, add $6 \times 100 \mu\text{L}$ elution buffer to each column. Collect each flow-through and analyze separately. Afterwards, pool all fractions containing the target protein.
11. To regenerate the column, add $3 \times 1 \text{ mL}$ regeneration buffer to each column, followed by addition of $2 \times 800 \mu\text{L}$ of $1 \times$ washing buffer. For storage add 2 mL washing buffer and place the columns at 4 °C.
12. To exchange the elution buffer to synthetase storage buffer, apply pooled elution fractions to Zeba™ Spin Desalting Columns. First, remove the storage solution of the Zeba™ Spin Desalting Column by centrifugation of the column at $1500 \times g$ for 1 min. Subsequently, add 300 μL synthetase storage buffer to the resin bed and centrifuge the column at $1500 \times g$ for 1 min. After repeating this step $2 \times$ place the

column into a new collection tube and apply 100 μL of the pooled synthetase solution to each column. Collect target proteins by a final centrifugation step at $2000 \times g$ for 2 min.

13. Concentrate target protein by using Amicon Centrifugal Filter Devices 0.5 mL. Adjust the volume of the synthetase from **step 12** to 500 μL with storage buffer and add this solution to the concentrator. Centrifuge at $14,000 \times g$ for 10 min and 4°C . Collect the flow-through. The concentration of the flow-through can be determined by photometric measurement using NanoDrop based on the calculated molecular mass of the synthetase (48.6 kDa) and the extinction coefficient (54.3) (*see Note 11*). Store synthetase at -80°C after shock freezing in liquid nitrogen.

3.3.2 Preparation of Orthogonal tRNA

1. First, suitable DNA templates of the tRNA gene need to be generated. For this, a reverse primer containing a 2'-OMe group has to be used in order to prevent unspecific addition of nucleotides to the 3' end by the T7 RNA polymerase.
2. The PCR reaction is composed of the following components: $1 \times$ Taq Buffer, 0.2 mM dNTP mix, 0.5 μM forward primer, 0.5 μM reverse primer, 2.5 mM MgCl_2 , 0.01 ng/ μL plasmid, and 0.04 U/ μL Taq DNA polymerase.
3. Fill the reaction with ultrapure water to a final volume of 250 μL .
4. Use the following PCR program: (1) 5 min 95°C , (2) 30 s 95°C , (3) 30 s 52°C , (4) 10 s 72°C , (5) 10 min 72°C , (6) cooling to 4°C . Repeat **steps 2–4** $30 \times$. Analyze generated PCR products by agarose gel electrophoresis on a 2% agarose gel.
5. Purify amplified tRNA PCR products by using a PCR Purification Kit (*see Note 12*).
6. Apply 50 μL PCR product per column.
7. Elute in 20 μL ultrapure water.
8. Analyze DNA concentration using NanoDrop.
9. Thaw *in vitro* transcription components on ice. Mix all components before using.
10. Pipet the reactions at RT. The transcription reaction is composed of the following components: 100 μL $5 \times$ transcription buffer (f.c. $1 \times$), 100 μL $5 \times$ NTP mix (f.c. $1 \times$), 25 μL $20 \times$ enzyme mix (f.c. $1 \times$), and 8 ng/ μL (f.c.) template DNA. Fill the reaction with water to the final volume of 500 μL (*see Note 13*). Incubate the reaction for 3–6 h at 37°C and 500 rpm.
11. After completing the reaction, centrifuge tRNA transcripts at $12,000 \times g$ for 1 min and collect the supernatant.

12. Analyze tRNA transcripts by agarose gel electrophoresis (2%) using 2 μL of the tRNA transcript (*see Note 14*).
13. Treat the transcription reaction with 1 U DNaseI per 1 μg plasmid DNA for 10 min at 37 °C and 500 rpm.
14. Add a three-fold volume of TRIzol to the transcription reaction and mix carefully. TRIzol and chloroform shall be handled with care and under the fume hood. Incubate for 5 min at RT.
15. Add chloroform (200 μL per 1 mL TRIzol) and mix for 15 s by carefully inverting the tube. Incubate 2–3 min at RT. Centrifuge at $12,000 \times g$, 4 °C and 15 min.
16. Remove the aqueous phase and transfer to a fresh reaction tube (*see Note 15*).
17. Add isopropanol (HPLC grade, 500 μL per 1 mL TRIzol) and mix carefully, followed by incubation over night at 4 °C.
18. Centrifuge at $15,000 \times g$, 4 °C for a minimum of 1 h. Remove the supernatant and discard it.
19. Overlay the RNA pellet with ethanol (1 mL 75% per 1 mL TRIzol) and incubate at –20 °C for 30 min.
20. Centrifuge at $7.500 \times g$, 4 °C for 10 min. Remove the supernatant quantitatively and air dry the pellet.
21. Resuspend the pellet in water (80 μL per 0.5 mL transcription reaction).
22. Measure RNA concentration using NanoDrop and dilute to 100 μM .
23. To fold the RNA, use the following PCR program: 120 s 80 °C, 30 s 75 °C, 30 s 70 °C, 30 s 65 °C, 30 s 60 °C, 30 s 55 °C, 30 s 50 °C, 30 s 45 °C, 30 s 40 °C, 30 s 35 °C, 300 s 25 °C, 4 °C.
24. Shock-freeze tRNA in liquid nitrogen and store at –80 °C.

3.3.3 Cell-free Synthesis of IgG and Site-Specific Incorporation of Non-canonical Amino Acids (Ncaa)

1. For the site-specific incorporation of ncaa into *de novo* synthesized antibodies cell-free reactions as described in Chap. 3, Subheading 1.1 need to be additionally supplemented with an orthogonal tRNA/synthetase pair and the non-canonical amino acids (ncaa).
2. Pipet the components in the following order and mix after addition of each component: 5 μL 10 \times translation mix (f.c. 1 \times), 1.5 μL AzF (f.c. 3 mM), 20 μL CHO lysate (f.c. 40%), 2.5 μL tRNA^{Tyr}CUA (f.c. 5 μM), 1.5 μL eAzFRS (f.c. 3 μM), 2.5 μL of 200 dpm/pmol ¹⁴C-leucine (specific radioactivity of 66.67 dpm/pmol f.c.), 1 μL T7 RNA polymerase (f.c. 1 U/ μL), and 10 μL 5 \times energy mix (f.c. 1 \times) (*see Note 16*).

3. Add HC and LC encoding plasmid at a final concentration of 60 nM each.
4. Adjust the final volume of the reaction mix with ultrapure water to 50 μ L. Incubate the reaction for 3 h at 27 °C and 500 rpm and protect the reaction from light during incubation.
5. After completing the reaction, centrifuge the translation mixture at $16,000 \times g$ for 10 min at 4 °C. Remove and discard the resulting supernatant (SN1).
6. Wash the microsomal fraction (MF) with 200 μ L PBS, centrifuge for 3 min at $16,000 \times g$ at 4 °C, remove the supernatant and resuspend the pellet in PBS including 0.2% DDM. Incubate this solution for 45 min at RT under agitation.
7. Centrifuge for 10 min, $16,000 \times g$ and 4 °C. Remove the resulting supernatant (SN2) which contains solubilized antibodies and place it on ice.

**3.3.4 Labeling of IgG
with Fluorescent Dye
(Staudinger Ligation)**

1. Pipet the following components: 5 μ L SN2 fraction, 1 μ L DyLight 650-phosphine (f.c. 10 μ M), 4 μ L ultrapure water, resulting in 10 μ L final reaction volume. Incubate reactions for 1 h at 25 °C und 600 rpm.
2. Analyze labeling reaction by autoradiography and in-gel fluorescence as described in Chap. 3, Subheading 4. (Fig. 3).

**3.4 Qualitative
Protein Analysis**

**3.4.1 SDS-PAGE
and Autoradiography**

1. Take a 6 μ L aliquot of the sample (e.g., TM, MF, SN2) and mix with 6 μ L 2 \times non-reducing LDS sample buffer. Incubate on a shaker for 15 min at RT (*see* **Note 17**).
2. Use 3–8% Tris acetate gels for gel electrophoresis. Pipet 12 μ L per gel pocket and run electrophoresis at 150 V for 1 h.
3. After electrophoresis, remove the gel from the plastic cassette and put in 100 mL water for in-gel fluorescence analysis (3.4.2).

**3.4.2 In-Gel
Fluorescence Analysis
and Autoradiography**

1. Analyze the gel using Amersham Typhoon RGB Biomolecular Imager. Place the gel onto the scanning surface and scan using the following parameters: 633 nm extinction and 670 nm emission for DyLight-phosphine and 532 nm extinction and 580 nm emission for Bodipy-TMR-lysine.
2. Rinse the gel three times with deionized water to remove SDS and buffer salts.
3. Dry gels for 90 min at 70 °C using a vacuum filtration system.
4. Label the marker bands of the dried gel with radioactive ink.
5. Place dried gel into a phosphorimager cassette and incubate for at least 3 days.

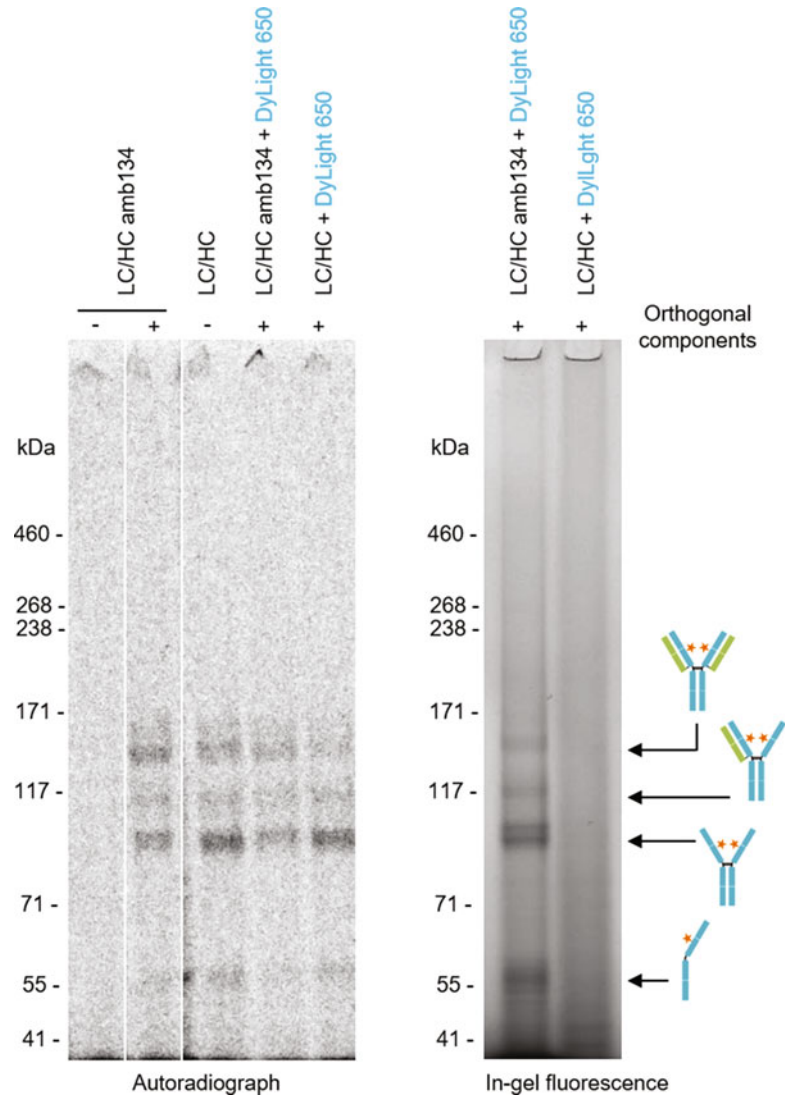


Fig. 3 Batch-based cell-free synthesis, site-specific introduction of ncaa by amber (amb) suppression and subsequent fluorescence labeling of antibodies with DyLight 650-phosphine. Qualitative analysis of cell-free synthesized antibody heavy chain (HC) and light chain (LC) was performed by SDS-Page followed by autoradiography (left side) and in-gel fluorescence (right side) (analysis of SN2). Cell-free synthesis was performed in the presence of ^{14}C -leucine and with (+) or without (–) supplementation of orthogonal components (synthetase eAzPheRS-SII, tRNATyrCUA and ncaa p-azido-L-phenylalanine (AzF)). The amb stop codon TAG was positioned in the CH1 domain (replacing S134, EU numbering) of HC. Orange asterisks indicate fluorescent dye conjugated to the incorporated ncaa. Unassembled LC (25.4 kDa) and termination product of HC amb134 (16.4 kDa) cannot be visualized in the autoradiograph because of its low molecular weight

6. Scan the screens using the Typhoon Trio + Variable Mode Imager or Amersham Typhoon RGB Biomolecular Imager.

4 Notes

1. CHO lysates were prepared as described previously [10]. In brief, CHO cells were grown in a Biostat B-DCU II bioreactor (Sartorius Stedium Biotech GmbH) at 37 °C using a chemically defined and serum-free cell medium. Cells were grown to a density of $3.5\text{--}5 \times 10^6$ cells/mL and harvested by centrifugation at $200 \times g$ for 5 min. Cells were washed twice and resuspended in a HEPES-based homogenization buffer (40 mM HEPES-KOH (pH 7.5), 100 mM NaOAc and 4 mM DTT). Resuspended CHO cells were lysed mechanically by applying a 20-gauge needle and a syringe. By using the syringe, cells were manually passed through the needle. After cell disruption, the homogenate was centrifuged at $6500 \times g$ for 10 min to remove cell nuclei and debris. The resulting supernatant was applied to a Sephadex G-25 column which was equilibrated with homogenization buffer. Elution fractions with the highest RNA/protein ratios were pooled. In order to remove the endogenous mRNA, cell lysates were mixed with S7 micrococcal nuclease (f.c. 10 U/mL) and CaCl_2 (f.c. 1 mM) and incubated for 2 min at RT. Inactivation of micrococcal nuclease was achieved by the addition of EGTA (f.c. 6.7 mM). Afterwards, creatine kinase (f.c. 100 µg/mL) was added to the lysate in order to ensure the regeneration of ATP out of creatine phosphate. Aliquots of the CHO lysate were immediately shock frozen in liquid nitrogen and subsequently stored at - 80 °C until further usage.
2. Coding sequences of HC and LC should be N-terminally fused to the melittin signal sequence to allow for the translocation of *de novo* synthesized polypeptide chains into the lumen of the microsomal vesicles [7]. Furthermore, HC and LC sequences should be fused to regulatory sequences necessary for CFPS (Fig. 1). The 5' untranslated region (5'UTR) of HC/LC templates contained a T7 promotor sequence and an internal ribosomal entry site (IRES from the intergenic region (IGR) of the Cricket paralysis virus (CrPV), (Genbank accession no. AF218039, nucleotides 6025–6216)) as regulatory sequences to allow for efficient transcription based on T7 RNA polymerase and factor-independent translation initiation, respectively. The 3'UTR contained a T7 terminator sequence and a multiple cloning site for subsequent cloning. HC/LC sequences were synthesized *de novo* by Geneart (Life technologies, Thermo Fisher) and cloned into pMA vector backbone.

The sequence of the variable domains was kindly provided by the lab of Michael Hust *et al.* (Technische Universität Braunschweig) [12]. The position of the amber stop codon in the CH1 domain at Serin 134 was chosen according to Zimmerman et al. (2014) [4].

3. The template used for protein synthesis should contain a T7 promotor, ribosomal binding site, and T7 terminator such as pIX3.0, pIVEX2.3d, and pIVEX2.4d vectors. Alternatively, a T5 promotor as contained in pQE2 vectors can be used.
4. The fluorescent dye Bodipy is susceptible to light. Protect it from light by using colored tubes or wrap the tubes with aluminum foil.
5. It is very important to work in an RNase-free environment and with RNase-free equipment. Use RNase-free filter tips for pipetting and RNase-free reaction vessels. Pipette the components in the listed order. Furthermore, it is recommended to avoid repeated freeze-thaw cycles of all components. After usage, shock-freeze the lysate in liquid nitrogen and store it at -80°C .
6. The DNA template used for CFPS should contain a T7 promotor. We found that the applied DNA template concentration is a potential parameter for optimization because template concentration influences protein synthesis efficiency. Different DNA templates may have different optimum concentrations within the cell-free reaction. For the synthesis of the chosen model antibody, we found that a 1:1 plasmid ratio of HC/LC, each added at 60 nM, worked best.
7. Cell-free reactions are scalable. You can adjust the final volume of the reaction according to the requirements of your experiment.
8. In general, the optimal temperature of the CHO cell-free system is 30°C [11] but the optimal incubation temperature may be different for different proteins. For antibody synthesis, 27°C was found to result in the highest yields of active antibodies.
9. Cell-free synthesized antibodies which have been translocated and trapped inside the lumen of the microsomes can be released by re-solubilization of the microsomal vesicles using PBS supplemented with 0.2% DDM. It is important to thoroughly resuspend the microsomes within the buffer in order to release translocated antibodies quantitatively.
10. We find that it is beneficial to collect all fractions throughout the purification, buffer exchange, and concentration procedure (e.g., flow-through, washing fractions, elution fractions). Aliquots of these solutions should be diluted in SDS-PAGE

sample buffer and analyzed by SDS-PAGE in order to monitor the purity of the aminoacyl-tRNA-synthetase during the preparation procedure.

11. We recommend to concentrate the synthetase up to a concentration of 5 g/L to ensure a minimal final concentration of 100 μ M. If necessary, repeat the concentration step.
12. We purify the PCR product using QIAquick PCR Purification Kit and determine the concentration of the PCR product by using a NanoDrop 2000c. For further analysis, prepare a 1% (w/v) agarose gel and load 1 μ L of the PCR product. The expected band size is 123 bps.
13. It is important to work in an RNase-free environment. Use RNase-free filter tips and reaction vessels.
14. Prepare a 2% (w/v) agarose gel. For sample preparation, mix 2 μ L of the RNA with 6 μ L MOPS sample buffer and load the sample to the agarose gel. Use an RNA ladder. The expected band size is around 200 bps.
15. After centrifugation, three phases will be visible: the aqueous phase on top with approximately 50% of the total volume containing the RNA; a middle interphase which is nearly invisible and below the red phenol/chloroform phase. Try to isolate only the aqueous phase.
16. Reactive groups of ncaa are often sensitive to light and might become instable upon light exposure. Thus, protect solutions from light by using colored tubes or wrap the tubes with aluminum foil.
17. The use of non-reducing sample buffer is important to maintain the disulfide bonds which connect the polypeptide chains of the antibodies. Heating of samples before gel electrophoresis is not necessary.

Acknowledgments

The authors would like to thank Prof. Dr. Michael Hust for providing the sequence of the antibody variable domains. Furthermore, we would like to thank Doreen Wüstenhagen and Dana Wenzel (Fraunhofer IZI-BB, Potsdam-Golm) for their excellent support regarding the preparation of the CHO lysates used in this study. This work is supported by the European Regional Development Fund (EFRE), the German Ministry of Education and Research (BMBF, No. 031B0078A), and the German Research Foundation (DFG Priority Programme 1623).

References

1. KEGG DRUG Database. [Online] 15.02.2021. <https://www.genome.jp/kegg/drug/br08328.html>
2. Davies SL, James DC (2009) Engineering mammalian cells for recombinant monoclonal antibody production. [Buchverf.] Mohamed Al-Rubeai. Cell line development. Springer Netherlands, Dordrecht. Bd. 6, S. 153–173
3. Axup JY (2012) Synthesis of site-specific antibody-drug conjugates using unnatural amino acids. *Proc Natl Acad Sci USA* 109:16101–16106
4. Feng T, Lu Y, Manibusan A et al (2014) A general approach to site-specific antibody drug conjugates. *Proc Natl Acad Sci U S A* 111:1766–1771
5. Zimmerman ES, Heibeck TH, Gill A et al (2014) Production of site-specific antibody-drug conjugates using optimized non-natural amino acids in a cell-free expression system. *Bioconjug Chem* 25:351–361
6. VanBrunt MP, Shanebeck K, Caldwell Z et al (2015) Genetically encoded Azide containing amino acid in mammalian cells enables site-specific antibody-drug conjugates using click cycloaddition chemistry. *Bioconjug Chem* 26:2249–2260
7. Stech M, Nikolaeva O, Thoring L et al (2017) Cell-free synthesis of functional antibodies using a coupled in vitro transcription-translation system based on CHO cell lysates. *Sci Rep* 7:S12030
8. Edwards H, Schimmel P (1990) A bacterial amber suppressor in *Saccharomyces cerevisiae* is selectively recognized by a bacterial aminoacyl-tRNA synthetase. *Mol Cell Biol* 10:1633–1641
9. Chin JW, Cropp TAS, Anderson JC et al (2003) An expanded eukaryotic genetic code. *Science* 301:964–967
10. Brödel AK, Wüstenhagen DA, Kubick S (2015) Cell-free protein synthesis systems derived from cultured mammalian cells. *Methods Mol Biol (Clifton, NJ)* 1261:129–140
11. Thoring L, Wüstenhagen DA, Borowiak M et al (2016) Cell-free systems based on CHO cell lysates: optimization strategies, synthesis of "difficult-to-express" proteins and future perspectives. *PLoS One* 11:e0163670
12. Thie H, Toleikis L, Li J et al (2011) Rise and fall of an anti-MUC1 specific antibody. *PLoS One* 6:e15921
13. Thoring L, Dondapati SK, Stech M et al (2017) High-yield production of "difficult-to-express" proteins in a continuous exchange cell-free system based on CHO cell lysates. *Sci Rep* 7:11710

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part III

Structure Determination



Chapter 10

Solid-State NMR Spectroscopy for Studying Microtubules and Microtubule-Associated Proteins

Yanzhang Luo, Shengqi Xiang, Alessandra Lucini Paioni, Agnes Adler, Peter Jan Hooikaas, A. S. Jijumon, Carsten Janke, Anna Akhmanova, and Marc Baldus

Abstract

In this chapter, we describe the preparatory and spectroscopic procedures for conducting solid-state NMR experiments on microtubules (MTs) obtained from human cells and their complexes with microtubule-associated proteins (MAPs). Next to labeling and functional assembly of MTs and MT-MAP complexes, we discuss solid-state NMR approaches, including fast MAS and hyperpolarization methods that can be used to examine these systems. Such studies can provide novel insight into the dynamic properties of MTs and MT-MAP complexes.

Key words Solid-state NMR spectroscopy, Microtubules, Microtubule-associated proteins, Protein dynamics, Protein interactions

1 Introduction

For more than four decades, solid-state Nuclear Magnetic Resonance (ssNMR) spectroscopy has been used to study complex biomolecular systems and recent advancements in ssNMR instrumentation and methodology have greatly expanded the utility and scope of such studies [1, 2]. Next to spectroscopic aspects, preparative procedures, in particular the generation of suitably (isotope-) labeled biomolecules have significantly enlarged the scope of research. Next to the well-established use of bacterial expression systems, significant progress has recently been made to produce functional and properly folded human protein targets for NMR studies, for example, by using yeast, insect, or mammalian cells [3]. Our group has previously shown that mammalian cell lines can indeed be used to conduct ssNMR studies on membrane proteins [4]. More recently, we introduced ssNMR approaches to study soluble proteins inside human cells that were pre-labeled

using bacterial expression systems [5]. In the latter case, we made use of electroporation procedures that provide an avenue for the delivery of selectively labeled specific proteins in an otherwise unlabeled cell background [6].

Here, we describe how to adapt these procedures to study microtubules (MTs) and associated proteins using isotope-labeled microtubules obtained from human cells. MTs are cytoskeletal polymers composed of tubulin subunits, which are essential for many biological processes, including cell division, migration, polarization, and intracellular trafficking. The dynamics and organization of MTs are regulated by nucleotide binding and many microtubule-associated proteins (MAPs). Previously, solid- and solution-state NMR have been used to study the interactions between MTs and isotope-labeled MAPs or small drugs [7–13]. In spite of the progress in producing isotope-labeled soluble tubulin for solution-state NMR studies, extending such experiments to using labeled MTs has failed so far. Firstly, production of recombinant, functional α/β -tubulin dimers from bacteria, which are commonly used to obtain labeled proteins, has not been possible, most likely due to the lack of chaperones and cofactors for tubulin folding and dimerization. In addition, previous protocols for tubulin purification from mammalian cells did not provide sufficient yields to obtain isotope-labeled MTs for NMR studies. Here, we describe how to prepare and label MTs and MAPs for ssNMR studies. In the latter case, we utilize the CKK domain that is important for the minus-end recognition of the Calmodulin-regulated spectrin-associated protein (CAMSAP) [9, 13]. Subsequently, we discuss solid-state NMR methods, including ultra-fast magic angle spinning (MAS) and hyperpolarization methods (such as Dynamic Nuclear Polarization, DNP) that can be used to study these systems. Such studies can provide novel insight into the dynamic interactions of MT tails [14]. As we have shown recently for the case of mRNA processing bodies [15], such experiments could also reveal dynamic interactions leading to the compartmentalization of the MT lattice by condensation of tau or other MAPs.

2 Materials

2.1 Cell Culture

1. HeLa S3 cell line (ATCC® CCL-2.2™).
2. Dulbecco's Modified Eagle Medium (DMEM, without amino acids, with 1 g/L glucose), [^{13}C , ^{15}N] labeled algal amino acid mixture, 200 mM stable L-glutamine, glucose, dialyzed fetal bovine serum (FBS), penicillin-streptomycin (P/S). Add 1 g/L algal amino acid mixture to DMEM and adjust the final concentration of glucose to 3.5 g/L. When the compounds are

completely dissolved, filter-sterile the medium by passing through the membrane filter with 0.2 μ m pore size. Include 10% dialyzed FBS and 1% P/S to the medium. Store at 4 °C.

3. 15 cm mammalian cell culture dishes, 1 L mammalian cell culture using Corning Erlenmeyer flasks and a cell culture incubator that includes a shaker.

2.2 Preparation of Isotope-Labeled MTs

1. Phosphate-buffered saline (PBS).
2. BRB80 buffer: 80 mM PIPES pH 6.8, 1 mM EGTA, 1 mM MgCl₂. Store at 4 °C.
3. Lysis buffer: 80 mM PIPES pH 6.8, 1 mM EGTA, 1 mM MgCl₂, 1 mM β -mercaptoethanol, 1 mM PMSF, protease inhibitors. Store at 4 °C.
4. High-molarity PIPES buffer: 1 M PIPES pH 6.8, 10 mM MgCl₂, 20 mM EGTA. Store at 4 °C.
5. 100 mM guanosine triphosphate (GTP) solution, 10 mM Pacitaxel dissolved in DMSO, glycerol.
6. Ultracentrifuges such as an optimal 1-90k ultracentrifuge (Beckman Coulter) (rotor type TLA-55) and an Optima™ LE-80K ultracentrifuge (Beckman Coulter) (rotor type 70.1 Ti) as well as a French-press homogenizer.

2.3 Purification of Isotope-Labeled CKK Domain from Bacteria

1. Washing buffer for His-tag purification: 50 mM phosphate buffer, 200 mM NaCl, 1 mM β -mercaptoethanol, 1 mM PMSF, 10 mM imidazole, protease inhibitors, pH 8.0. Store at 4 °C.
2. Elution buffer for His-tag purification: 50 mM phosphate buffer, 200 mM NaCl, 1 mM β -mercaptoethanol, 1 mM PMSF, 400 mM imidazole, protease inhibitors, pH 8.0. Store at 4 °C.
3. Buffer for size exclusion chromatography (SEC) for CKK: 40 mM phosphate buffer, 500 mM NaCl, 1 mM dithiothreitol (DTT), pH 7.0. Store at 4 °C.
4. Buffer for storage of CKK: 40 mM phosphate buffer, 150 mM NaCl, 1 mM dithiothreitol (DTT), pH 7.0. Store at 4 °C.

3 Methods

3.1 Suspension Cell Culture for Isotope Labeling

1. Culture HeLa S3 in two 15 cm cell culture dishes in labeled DMEM, and then transfer the cells into 12 dishes with the same medium to grow until a confluence of ~80–90%. (5–6 days).
2. Harvest the cells by trypsinization.
3. Count the number of cells with a cell counter.

4. Transfer the cells into 2.1 L labeled medium to a cell density of $\sim 150,000$ cells/mL. The cells are grown at 37°C , 5 % CO_2 and 120 rpm in 1L Corning Erlenmeyer flasks. Each flask contains 300 mL medium. Let the cells grow until the cell density reaches $\sim 1.2\text{--}1.5 \times 10^6$ cells/mL (4–5 days).
5. Harvest the cells by centrifugation at $500 \times g$ for 20 min at 4°C . Collect the cell pellet and resuspend in PBS, centrifuge the cells again at $500 \times g$ for 15 min at 4°C , and use the cell pellet for MT preparation.

3.2 Sample Preparation of Isotope-Labeled MTs (See also ref [16])

1. Resuspend the harvested cells with 1 g cell/mL lysis buffer. Lyse cells on ice by passing through a French-press homogenizer three times at 1000 psi. Spin down the cell lysate at $120,000 \times g$ at 4°C for 30 min and collect the supernatant containing the tubulin. Centrifuge the supernatant again at $5000 \times g$ at 4°C for 15 min to remove the remaining cell debris before the next step.
2. Add half volume of glycerol and 1 mM GTP to the supernatant and mix well. Incubating the mixture at 30°C for 30 min for MT polymerization. Subsequently, spin down the crude MT pellet at $150,000 \times g$ (Type 70.1 Ti, Beckman Coulter) at 30°C for 30 min. Discard the supernatant and keep the pellet on ice.
3. Resuspend the pellet in BRB80 supplemented with protease inhibitors and keep on ice for 30 min to allow for MT depolymerization. For a more efficient depolymerization, resuspend the solution frequently (once every 5 min). Subsequently, centrifuge the solution at $150,000 \times g$ (Type 70.1 Ti, Beckman Coulter) at 4°C for 30 min, and collect the supernatant containing the soluble tubulin (see **Note 1**).
4. Add an equal volume of high-molarity PIPES buffer, together with an equal volume of glycerol and 1 mM GTP to the supernatant and mix well. Polymerize the tubulin at 30°C for 30 min with the high-molarity PIPES buffer to inhibit residual MAPs binding to MTs. Subsequently, add $20\mu\text{M}$ Paclitaxel to the reaction and incubate for 20 min to generate Taxol-stabilized MTs. Spin down the Taxol-MTs at $150,000 \times g$ (TLA-55, Beckman Coulter) at 30°C for 30 min. Wash the pellet with BRB80 containing $20\mu\text{M}$ Paclitaxel and protease inhibitors and repeat the centrifugation to collect the MTs (Fig. 1) (see **Notes 2–4**).

3.3 Purification of Isotope-Labeled CKK from Bacteria

1. Express the protein of interest from *E. coli* in isotope-labeled M9 medium. Harvest the bacteria and resuspend in washing buffer for His-tag purification. Lyse the bacteria by sonication and collect the cell lysate by centrifugation at $40,000 \times g$ for 30 min at 4°C .

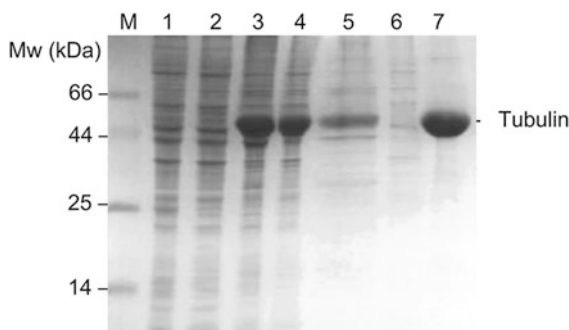


Fig. 1 Tubulin purification. SDS-PAGE analysis of tubulin purification from HeLa S3 cells. Lane M corresponds to the protein marker. From left to right the lanes show: the total cell lysate (lane 1); the supernatant of the first polymerization (lane 2); the pellet of the first polymerization (lane 3); the supernatant of the depolymerization (lane 4) as well as the associated pellet (lane 5); the supernatant of the second polymerization (lane 6) and the pellet of the second polymerization indicating the purified MTs (lane 7)

2. Filter the lysate by passing through a membrane filter with 0.45 μ m pore size. Load the lysate to a Ni²⁺ column and elute the bound proteins with the elution buffer.
3. Collect the eluate from the His-tag purification. Load the solution onto a Superdex 75 26/60 column equilibrated with the buffer for SEC.
4. Collect the eluted CKK from SEC and buffer-exchange to the storage buffer with ultra-filtration with a membrane filter with 3 kDa pore size. Concentrate the protein and store at 4 °C.

3.4 *ssNMR and DNP Experiments on MTs and MT-MAP Complexes*

1. Resuspend the MT pellet in warm BRB80 buffer with 20 μ M Paclitaxel. Add the MAPs of interest in the solution and incubate the reaction for 30 min at 30 °C. Centrifuge the solution at 150,000 \times g (TLA-55, Beckman Coulter) at 30 °C for 30 min. Wash the pellet with warm BRB80 without disturbing the pellet.
2. Transfer the pellet into the rotor filling tool and fill the rotor by centrifugation at 115,500 \times g (SW32 Ti, Beckman Coulter) at 30 °C for 30 min.
3. Use 1.3 mm rotor and spin samples to 44 kHz or higher and maintain set temperature around 270 K. For ³¹P experiments, larger MAS rotor dimensions and higher sample temperature (e.g., 290 K) are preferred.
4. For ¹H-¹⁵N, ¹H-¹³C, or ¹H-³¹P polarization transfer experiments, use cross-polarization schemes involving ramps such as a 100–50% ramp on the ¹H-channel of 95.4, and 71.7 kHz on the ³¹P-channel, with 1.2 ms CP contact times. For ¹H

(proton) decoupling, employ SPINAL decoupling [17] during evolution and detection times at around 90 kHz decoupling fields. Reference ^{13}C , ^{15}N , and ^{31}P spectra using adamantane (methylene, CH_2 peak), ^{15}N -labeled histidine (CO peak), and phosphate buffer, pH 7, by setting the respective peaks to 31.48 ppm and 0 ppm, respectively. The ^{15}N frequency was referenced indirectly via the ^{13}C signal.

5. For 2D ssNMR experiments on MT-MAP complexes, use dipolar correlation experiments such as PDSO (mixing time typically 10–200 ms), NCA (transfer time typically 3 ms), NCACX, (CC transfer times between 5 and 200 ms), DQ-SQ (with DQ excitation and reconversion times of 2.5 ms, ^{13}C - ^{13}C radio frequency-driven recoupling (RFDR) [18], NH, and CANH (using NH transfer times of 600 μs).
6. For ssNMR experiments that probe mobile MT and/or MAP segments, use ^{15}N -HSQC, ^{13}C HSQC, and 3D ^{15}N -edited ^1H - ^1H spin diffusion experiments. Use a ^1H - ^1H mixing time of 200 ms for ^{15}N -edited ^1H - ^1H spin diffusion experiments.
7. In order to improve the signal-to-noise ratio, perform Dynamic Nuclear Polarization (DNP) experiments (Fig. 2), in which polarization is transferred from free electrons to the nuclei of interest, therefore enhancing the ssNMR sensitivity.
8. For DNP experiments on labeled MT or MT-MAP complexes, modify procedure described under Subheading 3.4, **step 1** as follows. Wash the MT or the MT-MAP sample with BRB80 in D_2O containing 30% glycerol- d_8 , 20 μM Taxol and centrifuge at $150,000 \times g$ (TLA-55, Beckman Coulter) at 30 °C for 30 min.
9. Subsequently, resuspend with the DNP radical solution, obtained by dissolving the DNP agent AMUPol [19] in 60% glycerol- d_8 , 30% D_2O , and 10% BRB80 with a radical concentration of 15 mM.
10. Transfer the sample in a Bruker 3.2 mm sapphire rotor, snap-freeze, and store in liquid nitrogen until use.
11. In preparation of DNP experiments, cool DNP MAS probe down to 100 K. Measure T_1 to determine optimal recycle delay. If needed, optimize DNP enhancements by adjusting field values (requires sweepable NMR magnet). The enhancement is obtained from comparison of the signal amplitude with and without microwaves. DNP Signal enhancement factors should typically range between 19 (Fig. 2a) and 70 on the 800 MHz/527 GHz and 400 MHz/ 263 GHz DNP instruments, respectively.

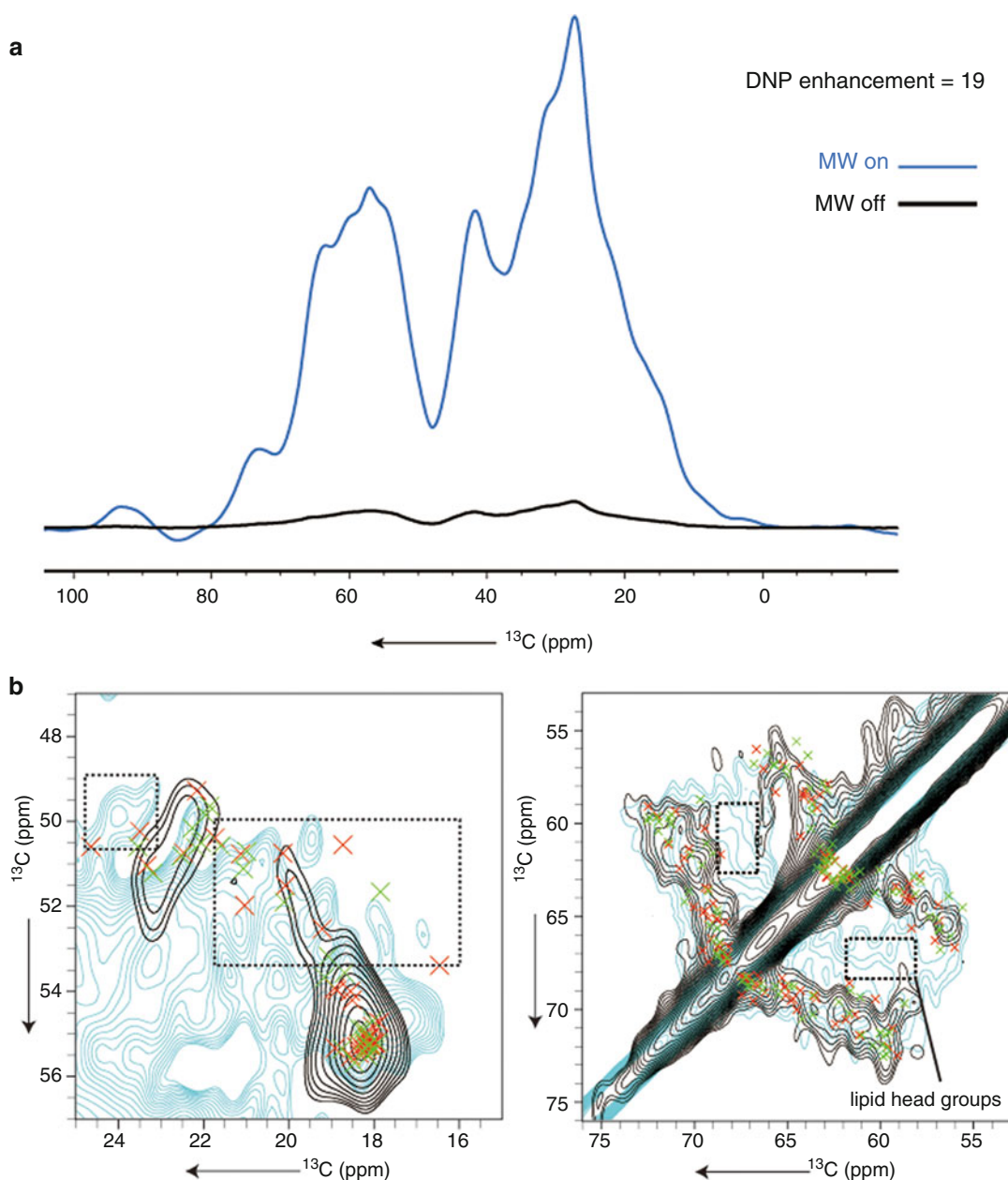


Fig. 2 DNP vs. standard ssNMR experiments. **(a)** DNP enhancement factor obtained on 800 MHz/527 GHz system. 1D ^1H - ^{13}C CP experiments of Taxol-stabilized MTs were measured on an 800 MHz spectrometer with DNP system. The black spectrum was recorded without microwave irradiation (therefore no DNP effect) and the blue spectrum was recorded with microwave irradiation (therefore enhanced by the DNP effect). An enhancement factor of 19 was obtained under the DNP conditions. **(b)** DNP-ssNMR experiments with [^{13}C , ^{15}N]-labeled, Taxol-MTs. Left Panel: Zoom-in on alanine $\text{C}\alpha$ - $\text{C}\beta$ regions on the ^{13}C - ^{13}C PDSD recorded with DNP at 100 K and MAS of 10.5 kHz (cyan) and RFDR at ambient temperature with MAS of 44 kHz (black). Chemical shift predictions [20] for alanine in random coil structures are highlighted in dash-line boxes. Right Panel: Zoom-in on serine and threonine $\text{C}\alpha$ - $\text{C}\beta$ regions on the same spectra. Extra signals that do not match prediction (stemming from co-purified lipids) were detected in the DNP experiment (dash-line boxes)

4 Notes

1. In the depolymerization step for MT preparation, the added BRB80 buffer should be kept at small volume (300–400 μ L buffer is used if starting with 8–10 mL harvested cell pellet). This helps obtaining a more efficient depolymerization resulting in more tubulin yield for the next steps. See also ref. [16].
2. Wear gloves when dissolving Paclitaxel to avoid exposing DMSO to skin and the Paclitaxel solution can be stored at $-20\text{ }^{\circ}\text{C}$ for 1 year. Prepare the GTP solution on the day of the preparation and keep the solution on ice. Avoid thawing and freezing of GTP. See also ref. [16].
3. We find that it is best to prepare the buffers fresh each time. (See ref. [14]).
4. When because of time limitations it is not possible to purify tubulin on the same day of harvesting cells, the purification can be stopped after the first polymerization. Store the polymerized MT pellet at $-80\text{ }^{\circ}\text{C}$ and continue the purification on the second day. See also ref. [16].

Acknowledgments

We thank Gert Folkers for helpful discussions and Johan van der Zwan for excellent technical support. This work was supported by the Dutch Science Foundation NWO (VENI grant 722.016.002 to SX, NWO-Groot grant 175.010.2009.002, and TOP-PUNT grant to MB) and by uNMR-NL, the National Roadmap Large-Scale NMR Facility of the Netherlands (grant 184.032.207). C.J. is supported by the grants ANR-10-IDEX-0001-02 PSL, ANR-11-LBX-0038, FRM DEQ20170336756. S.B. was supported by the FRM grant FDT201805005465, and CEFIPRA 5703-1, and J.A.S. by the European Union's Horizon 2020 Marie Skłodowska-Curie grant agreement No. 675737, and the FRM grant FDT201904008210.

References

1. Renault M, Cukkemane A, Baldus M (2010) Solid-state NMR spectroscopy on complex biomolecules. *Angew Chem Int Ed* 49:8346–8357
2. Quinn CM, Polenova T (2017) Structural biology of supramolecular assemblies by magic-angle spinning NMR spectroscopy. *Q Rev Biophys* 50:e1
3. Boisbouvier J, Kay LE (2018) Advanced isotopic labeling for the NMR investigation of challenging proteins and nucleic acids. *J Biomol NMR* 71:115–117
4. Kaplan M, Narasimhan S, de Heus C et al (2016) EGFR dynamics change during activation in native membranes as revealed by NMR. *Cell* 167:1241–1251
5. Narasimhan S, Scherpe S, Paioni A et al (2019) DNP supported solid-state NMR of proteins inside mammalian cells. *Angew Chem Int Ed Eng* 58:12969–12973

6. Theillet F-X, Binolfi A, Bekei B et al (2016) Structural disorder of monomeric α -synuclein persists in mammalian cells. *Nature* 530:45–50
7. Kumar A, Blommers MJ, Krastel P et al (2010) Interaction of Epothilone B (Patupilone) with microtubules as detected by two-dimensional solid-state NMR spectroscopy. *Angew Chem Int Ed Eng* 49:7504–7507
8. Yan S, Guo C, Hou G et al (2015) Atomic-resolution structure of the CAP-Gly domain of dynactin on polymeric microtubules determined by magic angle spinning NMR spectroscopy. *Proc Natl Acad Sci U S A* 112:14611–14616
9. Atherton J, Jiang K, Stangier MM et al (2017) A structural model for microtubule minus-end recognition and protection by CAMSAP proteins. *Nat Struct Mol Biol* 24:931–943
10. Kubo S, Nishida N, Udagawa Y et al (2013) A gel-encapsulated bioreactor system for NMR studies of protein–protein interactions in living mammalian cells. *Angew Chem Int Ed Eng* 52:1208–1211
11. Kesten C, Wallmann A, Schneider R et al (2019) The companion of cellulose synthase I confers salt tolerance through a Tau-like mechanism in plants. *Nat Commun* 10:857
12. Kadavath H, Fontela YC, Jaremko M et al (2018) The binding mode of a tau peptide with tubulin. *Angew Chem Int Ed Eng* 57:3246–3250
13. Atherton J, Luo Y, Xiang S et al (2019) Structural determinants of microtubule minus end preference in CAMSAP CKK domains. *Nat Commun* 10:5236
14. Luo Y, Xiang S, Hooikaas PJ et al (2020) Direct observation of dynamic protein interactions involving human microtubules using solid-state NMR spectroscopy. *Nat Commun* 11:18
15. Damman R, Schutz S, Luo Y et al (2019) Atomic-level insight into mRNA processing bodies by combining solid and solution-state NMR spectroscopy. *Nat Commun* 10:4536
16. Souphron, J, Bodakuntla, S, Jijumon, AS et al (2019) Purification of tubulin with controlled post-translational modifications by polymerization–depolymerization cycles *Nature Protocols*, 14, 1634–1660
17. Fung B, Khitrin AK, Ermolaev K (2000) An improved broadband decoupling sequence for liquid crystals and solids. *J Magn Reson* 142:97–101
18. Bennett AE, Rienstra CM, Griffiths J (1998) Homonuclear radio frequency-driven recoupling in rotating solids. *J Chem Phys* 108:9463–9479
19. Sauvée C, Rosay M, Casano G et al (2013) Highly efficient, water-soluble polarizing agents for dynamic nuclear polarization at high frequency. *Angew Chem Int Ed Eng* 52:10858–10861
20. Gradmann S, Ader C, Heinrich I et al (2012) Rapid prediction of multi-dimensional NMR data sets. *J Biomol NMR* 54:377–387



Chapter 11

Dynamic Structural Biology Experiments at XFEL or Synchrotron Sources

Pierre Aller and Allen M. Orville

Abstract

Macromolecular crystallography (MX) leverages the methods of physics and the language of chemistry to reveal fundamental insights into biology. Often beautifully artistic images present MX results to support profound functional hypotheses that are vital to entire life science research community. Over the past several decades, synchrotrons around the world have been the workhorses for X-ray diffraction data collection at many highly automated beamlines. The newest tools include X-ray-free electron lasers (XFELs) located at facilities in the USA, Japan, Korea, Switzerland, and Germany that deliver about nine orders of magnitude higher brightness in discrete femtosecond long pulses. At each of these facilities, new serial femtosecond crystallography (SFX) strategies exploit slurries of micron-size crystals by rapidly delivering individual crystals into the XFEL X-ray interaction region, from which one diffraction pattern is collected per crystal before it is destroyed by the intense X-ray pulse. Relatively simple adaptations to SFX methods produce time-resolved data collection strategies wherein reactions are triggered by visible light illumination or by chemical diffusion/mixing. Thus, XFELs provide new opportunities for high temporal and spatial resolution studies of systems engaged in function at physiological temperature. In this chapter, we summarize various issues related to microcrystal slurry preparation, sample delivery into the X-ray interaction region, and some emerging strategies for time-resolved SFX data collection.

Key words Serial femtosecond crystallography (SFX), X-ray-free electron laser (XFEL), Microcrystal slurry, Time-resolved macromolecular crystallography, X-ray emission spectroscopy (XES), Metalloenzymes

1 Introduction

Macromolecular crystallography (MX) is one of the most important techniques used by a large cross section of the research community and with almost 150,000 atomic models based upon X-ray diffraction released by the Protein Data Bank as of September 2020 [1, 2]. The vast majority of the MX datasets are collected from one to a few crystal(s) measuring $\sim 10\text{s} - 100\text{s } \mu\text{m}^3$, held at 100 K in a cryostream, and rotated about one or more axis during data collection [3]. Similarly, nearly all of the cryo-EM atomic models are collected from samples plunge-cooled to form a vitrified ice and

then held at ~100 K. But life is dynamic, and function is not compatible with cryogenic conditions. Indeed, dynamics and time scales in biology (Fig. 1) range from femtoseconds for electronic transitions and bond vibrations to picoseconds for light-induced charge separation, photo-isomerization, and amino acid side chain rotation, to microseconds for domain motion and ion transport and fast enzyme reactions, to milliseconds for most enzyme reaction rates and fast protein folding, and to seconds for protein synthesis and DNA or RNA replication/synthesis. An unmet grand challenge in structural biology is to routinely create molecular movies of macromolecular systems engaged in catalysis/function under physiological conditions and with high spatial and temporal resolutions.

We are experiencing a step-change in macromolecular crystallography. This is derived from serial MX and time-resolved functional studies that are directly linked to XFELs and the use of micron-size samples [4–7]. One important hypothesis is that micron-sized crystals minimize or eliminate barriers to relieving strain that builds up when conformational changes propagate across molecules and unit cells. A dramatic case in point is an aptamer of messenger RNA, a riboswitch that binds adenosine and undergoes a large conformational change that ultimately regulates gene expression [8–10]. Using a slurry of micron-size RNA crystals and mix-and-inject methods, Stagno et al. reported time-resolved SFX results that demonstrate ligand binding-induced conformational changes so significant that the crystals changed symmetry, but they did not shatter [11, 12]. This remarkable result has also recently been observed via atomic force microscopy wherein the probe-tip is scanned across the outer layers of microcrystals. These experiments demand the intensity and tight focus of the XFEL beam; without either, these experiments would not be feasible.

Crystallographers frequently observe microcrystal showers measuring only a few microns on a side that arise from initial sparse matrix screens used early in most projects. These conditions are then “optimized” to yield large single crystals typically measuring ~25–100s μm or more on at least two sides. During typical data collection at 100 K, the X-ray dose is distributed throughout the entire crystal volume, by combinations of rotation and translation [13–20]. And so, although microcrystals are ubiquitous, they are frequently overlooked because they can be difficult to use and/or perceived to be inappropriate for structural analysis. However, the characteristics of XFELs change the sample requirements for MX. Showers of microcrystals are ideal for XFELs that deliver hard X-rays with high flux density in a very well-focused femtoseconds-long pulse. Microcrystals of membrane proteins—an under-represented class in the Protein Data Bank (PDB)—are often produced and well suited for serial femtosecond crystallography (SFX)

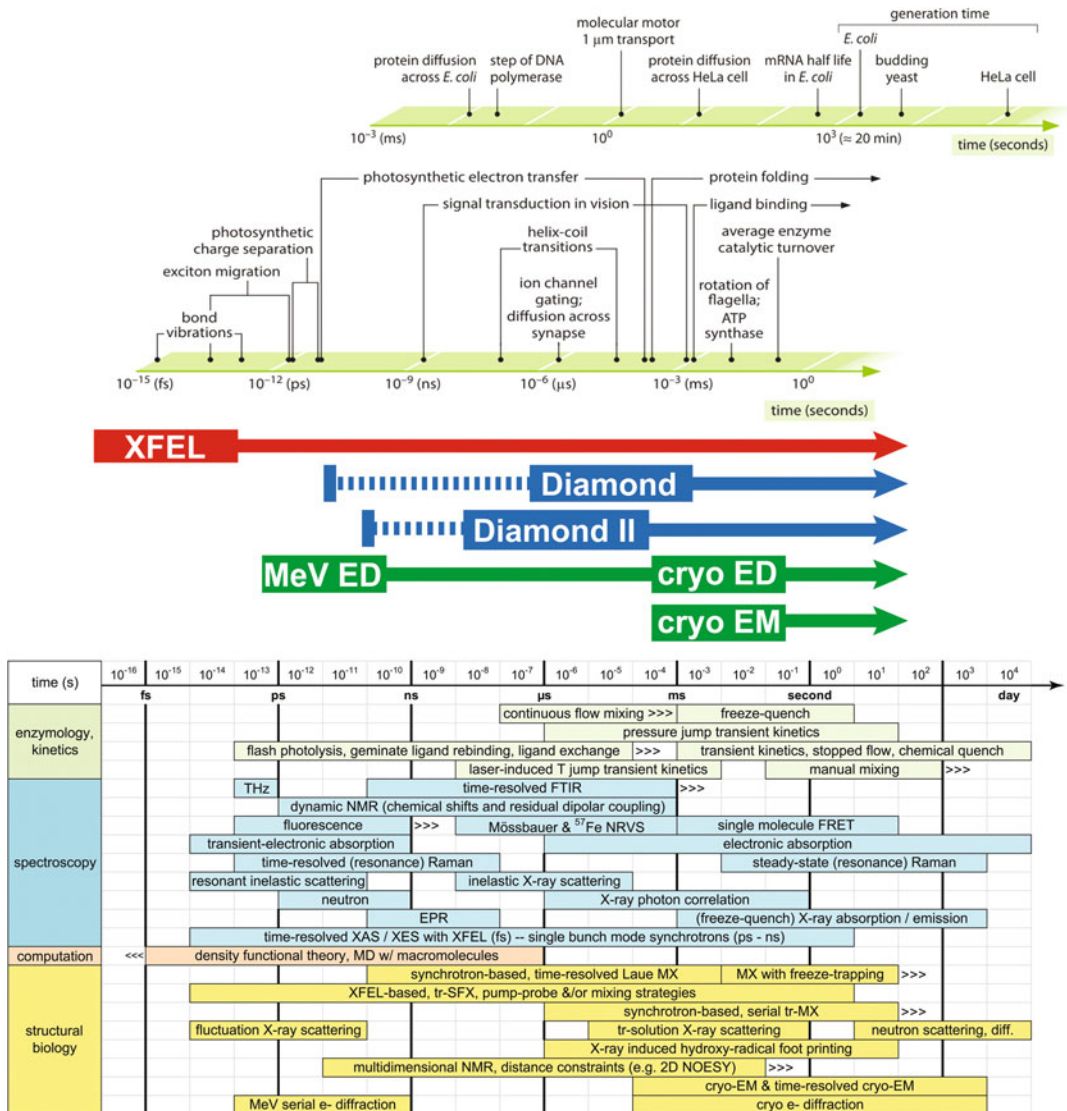


Fig. 1 Biophysicists use many tools to study a wide range of biological systems. A comparison of major events in biochemistry (top), X-ray, and electron sources for structural biology experiments (middle), and major techniques used in biophysical analysis (bottom). The biological processes span orders of magnitude in time and space. The time required to elicit a spectroscopic signal and/or an X-ray diffraction pattern to high resolution at a modern synchrotron (such as Diamond Light Source (Diamond) and its anticipated upgrade to a diffraction-limited lattice (Diamond II)) or XFEL range from microseconds to femtoseconds at synchrotrons and XFELs, respectively. Serial electron diffraction at a MeV source (MeV ED) is emerging as a method to exploit slurries of submicron samples at room temperature to provide data with high temporal and atomic resolution. Cryogenic methods have an inherent greater than hundreds of μ s time resolution limit that is imposed by the freeze-quench process (e.g., cryo-EM and cryo-ED). For comparison, several examples of major complementary methods are shown along the bottom. Scientists often use several of these methods in their research to build up as comprehensive an understanding as possible of structure, dynamics, and mechanism. Abbreviations: THz, terahertz spectroscopy; FTIR, Fourier-transform infrared spectroscopy; NMR, nuclear magnetic resonance; 2D-NOESY, two-dimensional Nuclear Overhauser Effect; NRVs, nuclear resonance vibrational

methods [21–26]. Moreover, SFX strategies have translated to analogous serial MX methods at synchrotrons around the world including Diamond Light Source (UK), and anticipated for Diamond II.

Serial femtosecond crystallography (SFX) is a new technique developed to exploit the femtoseconds long pulses from XFELs and to use thousands of micron-size crystals or smaller [4–7]. It is the dominant method in life sciences at all five XFELs since it was first reported in 2011 by Chapman and colleagues at the LCLS [27, 28]. SFX enables time-resolved experiments at XFELs with almost a complete lack of radiation-induced alteration, exploitation of micron to submicron-size crystals, and exquisitely sharp temporal resolution in time-resolved studies.

2 Sample Preparation (See Note 1)

When teaching serial crystallography, we find it useful to use numerical simulations and virtual labs to help illustrate the concepts. To this end, *XRayView* is an interactive program that allows one to visualize the overall experimental scheme, change various experimental parameters, and simulate the impact on the observed diffraction pattern at the detector [29, 30]. The program enables students to easily compare rotation-based experiments with polychromatic pink-beam Laue methods, and to SFX experiments similar to those executed at the LCLS. For more advanced explorations, the program *nanoBragg* calculates the absolute-scale scattering from a nanocrystal and creates realistic images, with and without photon-counting noise included, of the anticipated diffraction pattern in photons per pixel on the detector [31, 32]. In particular for SFX methods with slurries of very small samples, the program can also add “noise” in the form of scattering contributions approximated from the solvent (e.g., water) as a microdroplet or jet surrounding the nanocrystal.

Preparing samples for time-resolved SFX experiment requires a great quantity of microcrystals with high density. Depending on the sample delivery method chosen, tens of microliters to several milliliters of sample solutions with density up to 10^{11} microcrystals/ml will be needed. Indeed, SFX experiments are consuming a large

Fig. 1 (continued) spectroscopy; FRET, fluorescence resonance energy transfer; EPR, electron paramagnetic resonance; XAS X-ray absorption spectroscopy; XES, X-ray emission spectroscopy; MD, molecular dynamics; MX, macromolecular crystallography; tr-SFX, time-resolved serial femtosecond crystallography; tr-MX, time-resolved macromolecular crystallography; cryo-EM, cryo-electron microscopy; cryo-ED, cryo-electron diffraction; MeV, mega-electron volt; e-diffraction, electron diffraction. Top portion adapted from *Cell Biology by Numbers* by Ron Milo, Rob Phillips, Copyright 2015. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc

amount of protein from few hundreds of micrograms to tens of grams [33]. Research and developments are focusing to improve sample consumption for sample delivery systems in order to make serial crystallography experiment more accessible [34].

Protein crystallization needs to be optimized to produce a relatively large amount of homogenous microcrystals slurry. It is crucial to grow crystals having similar size for time-resolved SFX experiment for two main reasons: (a) Light excitation or diffusion of a substrate will be different into a small crystal ($\sim 1\ \mu\text{m}$) and a bigger one ($\sim 10\ \mu\text{m}$). (b) In serial crystallography, one crystal gives one diffraction pattern from an essentially stationary orientation, and to obtain a full dataset we need to merge diffraction images from thousands of crystals (usually about 10,000 or more). Needless to say that for the best results all the crystals need to be as identical as possible; a term typically referred to as isomorphous.

Vapor diffusion technique largely used to grow single crystals for cryo-crystallography is poorly suited to produce hundreds of microliters to few millilitre of microcrystal slurry. It is quite obvious that harvesting enough crystals from 24 or 96-well plates will be very time consuming. On the other hand, crystallization conditions defined from vapor diffusion can be used to extrapolate starting conditions for either free interface diffusion (FID) or batch method [35–37]. FID and batch method will allow production of large scale of microcrystal slurry well suited for SFX experiment.

2.1 Free Interface Diffusion (FID)

The production of homogenous nano/microcrystals for SFX experiments using FID was first described by Kupitz and collaborators in 2014 [37]. More practical details on FID can be found in the book chapter from Coe and Ros [36]. To briefly describe this technique, the protein solution (usually less dense) is transferred in 1.5 ml centrifuge tube. The reservoir solution containing the precipitant (more dense than protein solution) is added to the protein solution without mixing. It will create naturally a linear gradient allowing nucleation at some place along the gradient. Once the crystals take form, they will become denser and then, sink into the bottom of the tube where there is no protein. It will have for consequence to stop the crystal growth and to keep the crystals small and homogenous. One can gently centrifuge to help and speed up the process. Crystals should be harvested before both solutions (protein and reservoir) mix completely.

2.2 Batch Method

A recent paper from Beale and collaborators in 2019 explains in detail how to move from vapor diffusion crystallization conditions to a large-scale production with the batch method [35]. As described in the publication, the vapor diffusion conditions will be used to define the phase diagram of the protein. It is a prerequisite, as the phase diagram will give the conditions (protein and precipitant concentration) for the nucleation zone. Contrary to

FID, in the batch method the protein and the reservoir solutions are well mixed together. One can add crystal seeds to help to produce extra-nucleation. Several conditions need to be tested to have a good number of crystals with homogenous in size. Once the right size of crystal is achieved, the experimenter will have to stop the crystal growth by diluting them into high precipitant concentration buffer.

These two techniques will allow large-scale production of microcrystal slurry. But there is always possibility to have few outlier bigger crystals. Depending on the sample delivery method, these outliers can clog the system and be responsible for hours of down-time during the experiment. Then, it is a good practice to filter the crystals by attaching a capillary of similar or smaller inner diameter than the one used with the sample delivery system to the syringe used for harvesting the microcrystal slurry.

3 Sample Delivery and Data Collection

SFX studies are very often conducted at room temperature, from which one diffraction pattern is recorded from each stationary microcrystal in a random orientation. XFEL beams are typically well focused to deliver submicron or 1–3 μm spot size at the sample, and are about nine orders of magnitude brighter than synchrotrons like Diamond Light Source. Because so much energy is deposited into the sample, it explodes [38]. Consequently, SFX methods require a unique sample for each diffraction pattern and the whole dataset is merged from thousands of still images. To this end, sample delivery methods for serial MX is a very active R&D effort in the field (Fig. 2). The aim is to rapidly and efficiently deliver sample, without wasting precious material, at a rate that matches the XFEL pulse frequency and/or the detector characteristics.

A variety of methods have been developed to deliver a slurry of microcrystals into the XFEL interaction region (Fig. 2), which include liquid flow-focusing gas dynamic virtual nozzles (ff-GDVN) or viscous media jets and extruders [25, 27, 39–45], a concentric-flow electrokinetic injector [46–48], on-demand microdroplets that may be coupled to a conveyor belt transport system [34, 49–56], fixed targets that raster a sample array through the X-ray beam [34, 57–66], and even goniometer-based methods [67–71]. Many of these methods are readily adaptable to time-resolved studies in which the reaction is triggered by either light or mixing as discussed below [72].

SFX data collection times are driven by a combination of the X-ray pulse frequency, the detector capabilities, and the hit ratio defined as the number of indexed lattice(s) divided by total number of X-ray pulses delivered (Table 1). Crystal hit ratios range from

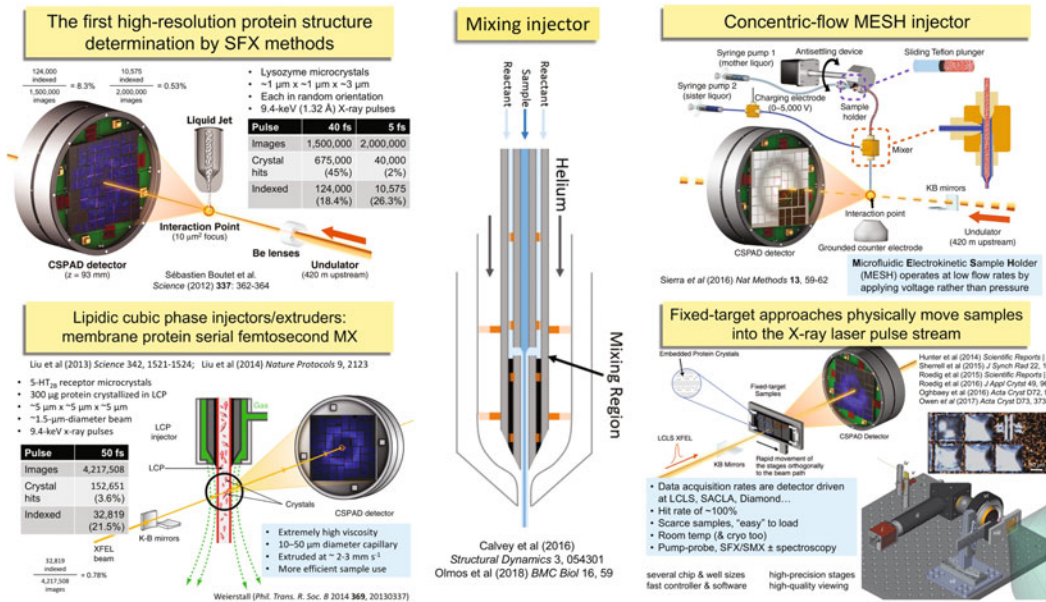


Fig. 2 Examples of sample delivery methods in serial femtosecond crystallography studies (see Note 1). Top left: A gas dynamic virtual nozzle (GDVN) used to produce fast liquid jets were among the first methods developed for SFX experiments [27, 28]. Viscous media extruders have been used at all XFELs and many synchrotrons since their slower flow rate provides for more efficient sample delivery, especially for lower frequency XFEL sources and/or X-ray detectors (bottom left) [25, 200, 201]. The concentric-flow electrokinetic injector (CoMESH) is an efficient sample delivery method that also enables ligand additions for time-resolved SFX experiments (top right) [46, 47, 202]. Fixed targets are among the most sample efficient methods and can achieve very high crystal lattice hit ratios (bottom right) [34, 57, 59–61, 63–67, 69, 203–205]. Mix-inject strategies blend GDVN liquid jets with substrate addition capabilities for more general time-resolved applications (central portion); see also references [40, 93, 173, 175, 176, 206–209]

much less than 1% to greater than 100% in some fixed target systems with more than one crystal in a fraction of the wells. Ideally, each crystal lattice will be in a random orientation and thus, the whole SFX dataset is built up in a stochastic process. Collecting as many as 25,000 indexed lattice dataset can take significantly more than a 12-h shift with a low hit ratio and/or a slow detector or pulse frequency. In contrast, only a few seconds are needed under conditions that yield a high indexing ratio with a fast detector. It is important to point out that in the fastest cases to date, the detector becomes a limiting factor rather than the X-ray pulse frequency. Moreover, in every case, the scientists must evaluate each SFX dataset in as near to real time as possible. This is clearly more challenging with faster data collection, larger image files, and larger unit cells with many more reflections that must be merged. Consequently, highly parallelized algorithms will be critical as well as access to large clusters and/or supercomputer centers for the most intensive calculations [73–78].

Table 1
Minimum time required for one SFX dataset with 25,000 indexed crystal lattices

Facility	Detector	Image rate Hz	Time to Collect Dataset with			
			5% hit ratio		80% hit ratio	
			Seconds	Minutes	Seconds	Minutes
LCLS	ePIX	120	4167	69.4	260	4.3
LCLS	Rayonix ^a	10	50,000	833.3	3125	52.1
LCLS	Rayonix ^a	30	16,667	277.8	1042	17.4
SACLA / PAL-XFEL	MPCCD	60	8333	138.9	521	8.7
SwissFEL	Jungfrau ^b	100	5000	83.3	313	5.2
European XFEL	Jungfrau ^b	160	3125	52.1	195	3.3
Diamond VMXi	Eiger2 ^c	500	1000	16.7	63	1.1
European XFEL	AGPID ^d	3520	142	2.4	9	0.15
LCLS-II-HE	ePIX ^c	10,000	50	0.83	3	0.05
SHINE	tbd ^f	17,000	29	0.48	2	0.03

^aRayonix MX340-XFEL detector with different on-chip pixel binning modes impact pixel size and image readout rate, e.g., $2 \times 2 = 10$ Hz, $4 \times 4 = 30$ Hz

^bA JUNGFRAU detector with 16 “on-board” memory cells that can collect 16 images across a 300 μ s pulse train arriving at 10 Hz (equivalent to 53.3 kHz intra-train frequency) at the European XFEL or at 100 Hz at the SwissFEL

^cThe Dectris Eiger2 is a counting detector and is shown for comparison at the VMXi beamline at Diamond Light Source. The detector does permit electronic gating to collect fractions of the 2 ms image rate

^dThe Eu.XFEL can deliver up to 27,000 pulses in 600 μ s duration trains at 10 Hz, to achieve an intra-train frequency of 4.5 MHz. The Adaptive Gain Integrating Pixel Detector has “on-board” memory sufficient to save 352 images per train at 10 Hz

^eUnder development and likely to collect at ~ 10 kHz evenly spaced across the 1 MHz pulse

^fUnder development and likely to collect at ~ 17 kHz evenly spaced across the 1 MHz pulse

Although SFX was first demonstrated in 2011 at the LCLS [27, 28], serial MX methods are now also available at nearly all major synchrotron facilities around the world [33, 79–84]. Moreover, beamlines dedicated to serial MX are often identified as a flagship capability at facilities undergoing upgrades to diffraction-limited lattice configurations, including Diamond II [85–87]. Thus, room temperature, serial MX methods that couple dynamics and functional studies with structural analysis will become routine around the world. This trend will continue to expand and impact all of life sciences.

3.1 Photosensitive Systems

Time-resolved serial MX naturally blends functional and structural analyses into the same sample and experiment. This is sometimes described as *Dynamic Structural Biology*. It exploits slurries of microcrystals at room temperature and is increasingly prevalent at XFELs and synchrotrons. Light-activated systems are relatively rare

in biology, but they are easier to initiate, and many have the potential to probe isomerization rates from femtosecond(s) and slower. General time-resolved structural biology at slower time scales with diverse biochemistry applications is separated from ultrafast applications by fundamental time scales of decoherence, which is typically completed within few picoseconds.

Heretofore, the *status quo* in structural biology has been to use many separate samples, from which different types of data are collected under different conditions. Some of the experimental conditions may be very far from physiological. For example, traditional MX or cryo-EM methods usually provide insightful atomic models of a particular ground state or a trapped intermediate state molecule from a sample held at 100 K. In contrast, most functional studies are conducted in solution, at room temperature, and often include binding, conformational changes, and/or other types of dynamics.

Ultrafast time-resolved spectroscopic experiments performed on photo-active proteins in solution have provided many important insights into the very early events after absorbing a photon. Because XFELs are still new, analogous time-resolved SFX experiments are still emerging. In most reports, authors compare time-resolved SFX and spectroscopic results; although they may have used similar pump-probe delay times, the comparisons frequently do not yield one-to-one correlations. One possible explanation is that the crystal versus solution conditions are sufficiently different that ultrafast spectroscopic experiments have not been performed at the same pH, viscosity, ionic strength, among many other potential variables that were present in the SFX experiments. Furthermore, the desire of structural biologists to observe and maximize illumination-dependent differences in electron density maps has often pushed experimental conditions into the multiphoton regime. Recently, it has become clear that visible light power dependence studies are critically important to nearly all light-activated time-resolved SFX experiments [88]. Unfortunately, in part because beamtime is so rare, scientists may underprioritize these “control” experiments, and as a result these types of data may be sacrificed under time-pressure situations.

Examples of some recent XFEL experiments involving light-activated systems include photosystem I [28, 89–93], photosystem II [46, 49, 51, 52, 69, 94–106], photo-active yellow protein [107–109], human rhodopsin [110, 111], bacteriorhodopsins [42, 112–118], jumping spider rhodopsin [119, 120], light-activated ion channels [121], fluorescent proteins [122–125], several phytochromes [51, 126, 127], DNA photolyase, and photo-dissociation studies of myoglobin-CO [128] or cytochrome c oxidase-CO [129, 130]. Of these systems, photosystems I and II have evolved mechanisms to diffuse the excess energy absorbed in multiphoton events through the network of internal chromophores. Therefore,

they are less susceptible to the impact of single versus multiphoton time-resolved SFX studies. The photo-dissociation of CO from metalloproteins can be considered outside their normal function and less constrained by illumination conditions. The remaining systems evolved to react to visible photons, and single photon methods are likely to be most physiological. Moreover, many of these systems experience photo-driven conformational changes that are often linked to photoisomerization of the chromophore. Such events are often very fast and require timing tools to help coordinate the pump-probe experiment [131–134].

3.1.1 Photosystem II as an Example System for Time-Resolved SFX

Photosystem II (PS-II) is an important benchmark system for time-resolved SFX studies with results coming principally from three international teams of researchers in the USA, Asia, and Europe [49, 89, 95–97, 99, 100, 103–106]. The enzyme is responsible for the “great oxidation event” approximately 2.4 billion years ago that transformed the earth from an anaerobic reducing atmosphere to the O₂ rich and oxidizing atmosphere today. PS-II is a large integral membrane protein expressed in all plant and most photosynthetic microorganisms. The protein uses four photons to catalyze the four electron oxidation of two water molecules that form one O₂ molecule plus four H⁺ that help establish a proton gradient used for ATP generation by ATP synthase. The PS-II reaction is driven by visible light photons that initiate very rapid charge separation events (fs-ps), and much slower electron transfer events (hundreds of ms) to reduce the quinone pool.

Thus, PS-II is an ideal system for time-resolved SFX studies; but challenging too since the entire reaction cycle spans more than 12 orders of magnitude in time. A large collaborative group developed an on-demand, acoustic droplet ejection onto tape system (Fig. 3) [34, 49, 51, 53, 97, 101, 102, 135]. It transports discrete nanoliter droplets across a laser illumination platform before they reach the X-ray interaction region. This allows for multiple illumination / equilibration perturbations to advance the reaction cycle for time-resolved SFX experiments. An added benefit of the design is that the region between the droplet ejection and the X-ray interaction region can also be fitted with either an O₂ reaction chamber or a second picoliter droplet ejector system, both of which enable a wide range of mixing-based time-resolved reactions.

O₂ bond formation within PS-II is catalyzed by the oxygen-evolving complex (OEC) that include a Mn₄O₅Ca cluster (Fig. 4) [136, 137]. When a photon is absorbed by the P680 chromophore, it results in charge separation and electron transfer to the quinone site. The oxidized P680 is then reduced by the OEC, which advances one oxidation state with each photon absorption event. In the Kok cycle, the first photon promotes the S₁ to S₂ transition, the second photon S₂ to S₃, and the third photon from S₃ to S₄. The S₄ state catalyzes the conversion of two water molecules into

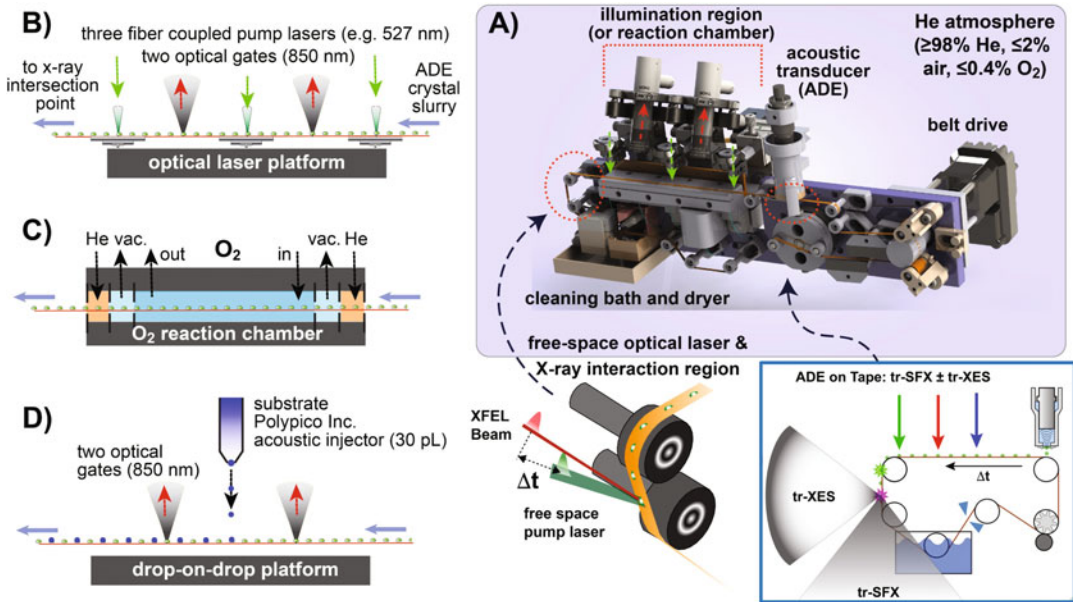


Fig. 3 Acoustic tape-drive system for pump-probe time-resolved SFX and XES measurements of photosystem II (a, b) and options for O_2 -dependent reactions (c) or drop-on-drop ligand additions (d). This on-demand sample delivery strategy is very efficient and flexible supporting several types of time-resolved SFX experiments. See also references [49, 51–53, 103, 135, 182, 210]

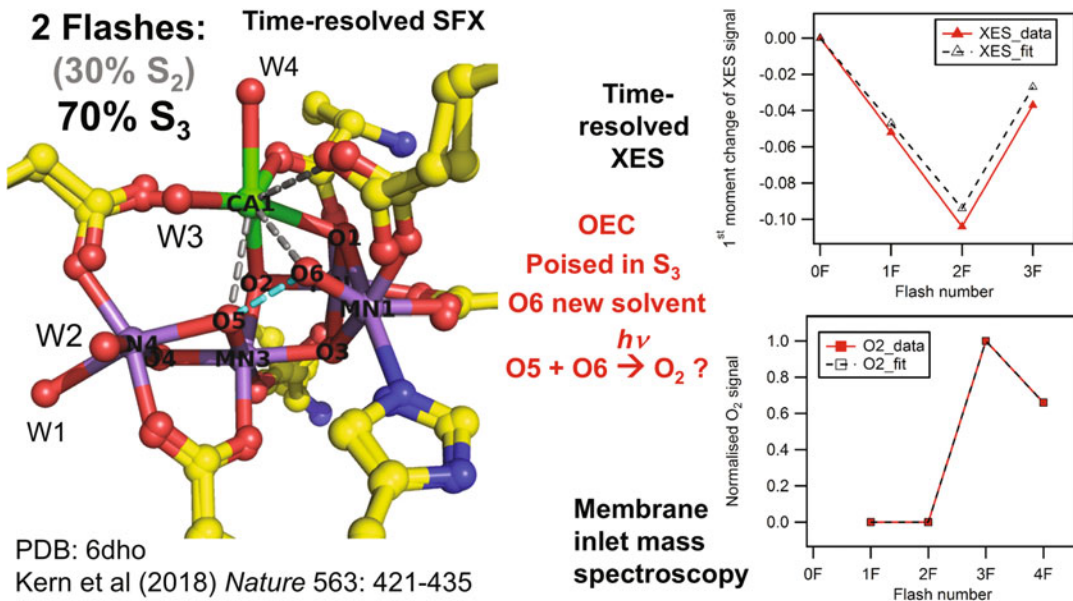


Fig. 4 Correlated time-resolved SFX and XES results from PS-II. The 2.09 Å resolution atomic model for S_3 (left). The first moment change of the Mn $K\beta_{1,3}$ XES spectra from PS-II crystals obtained in situ simultaneously with SFX (right, top). Flash-induced O_2 yield in crystal slurry as a function of flash number (right, bottom). See also references [49, 51, 97, 103]

O₂ resulting in the S₀ state of the OEC. Absorption of a fourth photon produces the stable S₁ starting state.

The Mn atoms within the cluster act as a redox “buffer” and each oxidation state has a unique K β _{1,3} X-ray emission spectrum. Consequently, time-resolved XES collected simultaneously with time-resolved SFX provides complementary data on the electronic and atomic structures of the catalytic center [49, 97, 99–102, 138]. Groups working principally at the LCLS and LBNL developed methods to simultaneously collect time-resolved SFX data in the forward direction and time-resolved X-ray emission spectroscopy (XES) at 90°, from each sample and each X-ray pulse [49, 51, 52, 97, 98, 100–102, 138–140]. Some of the results for PS-II are shown in Fig. 4 for the OEC poised in the S₃ state where μ s time-resolved structures demonstrate motion of the Mn atoms and entry of a new solvent atom (label O6 in the image) [49]. The atomic model is correlated with oxidation of the Mn atoms within the OEC and the O₂ generation assays from crystal slurries. Details for the time-resolved reaction that forms the O₂ molecule remain ill-defined; however, the deepest mechanistic insights will likely depend upon correlated studies and benefit from XFEL sources.

3.1.2 Caged Protein

Light-driven strategies to initiate catalysis in systems that are not naturally light sensitive include: (a) caged compounds, (b) caged proteins, (c) ligand exchange, or (d) temperature jump methods [94, 141–157]. Each of these must satisfy the requirements of high selectivity, high quantum yield, and temporal resolution. To these ends, *o*-nitrobenzyl moieties are among the most common photocaging groups for substrates and amino acids, but their decaging photochemistry may not be as fast nor as clean as *p*-hydroxyphenacyl or coumarylmethyl derivatives. Caged substrates are either co-crystallized with the target macromolecule or soaked into slurries of microcrystals. Incorporation of non-natural amino acids that convert a given protein into a caged protein is more difficult (e.g.,

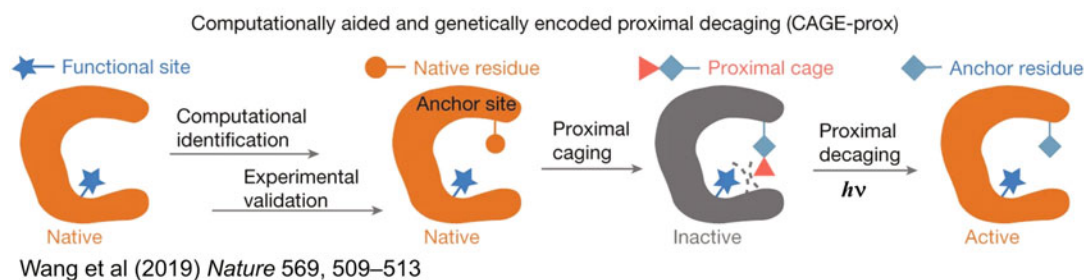


Fig. 5 Schematic concept of the CAGE-prox strategy. A proximal cage can be genetically incorporated at an anchor site in close proximity to the protein functional site for temporal blockage of its activity. Proximal decaging can lead to rapid rescue of protein functions, as long as the de-caged anchor residue has negligible influence on protein activity. Reproduced with permission from Wang et al. (2019) *Nature* 569, 509–513

Fig. 5). Consequently, caged protein approaches typically include a computational evaluation stage, followed by protein translation using *amber* TAG codon-suppression methods, or post-translational modification strategies. The photo-active moiety is most often linked to thio, amino, carboxy, or hydroxy groups of proteins or substrate ligands, and are then cleaved by irradiation with UV to visible light. Photo-cleavage of the caging group varies in rate (ps– μ s) and quantum yield (<0.2 – 1), which then generates: (a) authentic substrate in the active site vicinity, (b) a rapid pH shift, (c) a temperature jump, (d) removes an active site barrier and enables substrate binding, or (e) eliminates a dynamic or conformational restraint required for catalysis. Ligand exchange methods include photolabile metal-CO or NO complexes that mimic a metal-O₂ intermediates in a reaction cycle. Photo-dissociation of the blocking diatomic molecules then allows for O₂ binding and the ensuing reaction. All of these methods are experiencing a resurgence of research and development activity and provide important opportunities for time-resolved structural biology at XFELs. Some of these techniques will require careful coordination of the timing between the visible light pump laser and the XFEL probe pulse.

3.2 Enzyme Catalysis

The concept of dynamic structural biology is illustrated in Fig. 6 and links reaction dynamics with sample preparation, reaction triggering, and data collection. The KEGG database lists more than 6.75 million genes for enzymes that are derived from individual genomes or metagenomes [158–160]. Macromolecular crystals are

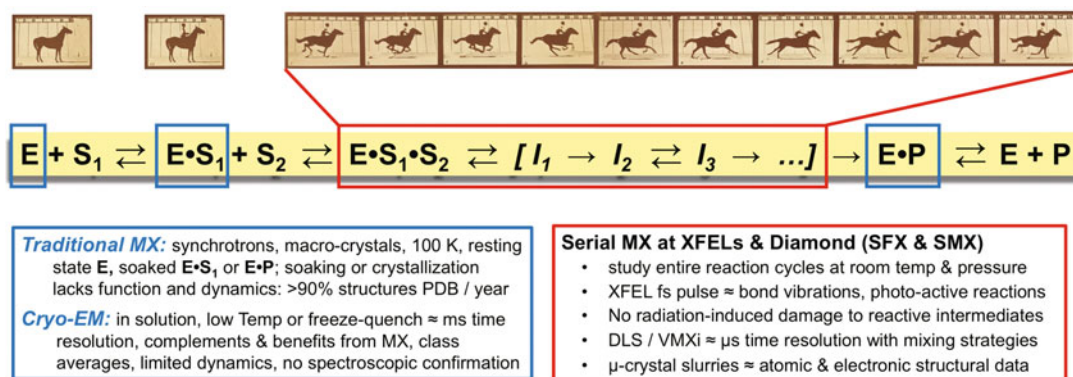


Fig. 6 Concepts of time-resolved structural biology. Snapshots of data inspired by Muybridge’s “horse in motion” photographs (Palo Alto CA, June 1878, 500 μ s shutter speed). Stable, ground state complexes are studied by traditional MX and cryo-EM methods, most often at 100 K. Time-resolved cryo-EM methods use freeze-quench techniques, typically with millisecond time resolution. Serial MX methods with microcrystals and monochromatic X-rays at synchrotrons or XFELs collect still diffraction images at room temperature with microsecond to femtosecond exposures, respectively, at different delay time intervals after optical laser flash (es) or ligand mixing

typically about 50% protein and 50% solvent, which is similar to the overall protein concentration inside cells. Analyzing more than 36,000 enzymes listed in the *BRENDA* database reveals that (a) the median turnover time for catalysis in solution is about 70 ms, (b) more than 60% exhibit a k_{cat} value between 1 and 100 s^{-1} , and (c) enzymes catalyzing reactions related to secondary metabolism are typically 30-fold slower than those of central metabolism [161–165]. Dynamics play important but often ill-defined roles in enzyme catalysis [166–168]. A driving hypothesis for time-resolved serial MX is that because small molecule substrates diffuse relatively fast (e.g., molecules in water at 310 K: O_2 (32 g/mol) = $2 \mu\text{m}^2/\text{ms}$; glycine (75 g/mole) = $1 \mu\text{m}^2/\text{ms}$; glucose (180 g/mole) = $0.6 \mu\text{m}^2/\text{ms}$; sucrose (342 g/mole) = $0.52 \mu\text{m}^2/\text{ms}$), enzyme microcrystals will equilibrate with substrates faster than catalytic turnover [169–172]. This implies that time-resolved structural biology methods are generalizable when exploiting micron-size crystals or smaller. However, due to the very limited availability of XFEL beamtime, relatively few time-resolved SFX experiments exploiting mixing strategies have been conducted at fully commissioned XFEL facilities [11, 51, 173, 174].

3.2.1 Mix-Inject Sample delivery for Time-Resolved SFX

To date, most time-resolved SFX studies that exploit mixing methods used liquid gas dynamic virtual nozzle (GDVN) style jets or microfluidic devices that have been developed by several groups (Fig. 2) [11, 12, 40, 41, 93, 173–177]. For the majority of cases, GDVN liquid jets typically achieve sample injection speeds of $10\text{--}30 \text{ m s}^{-1}$ and are suitable for $10\text{--}120 \text{ Hz}$ repetition rate XFEL sources. Many of these mixing-injectors are slow and need on the order of seconds to fully mix substrate with crystals. Other types of devices mix on the millisecond time scale, but require high-dilution ratios to infuse substrate into the crystal slurry stream. These types consume large amounts of substrate or ligand and dilute the crystal concentration, which also reduces the overall SFX data collection rates. All of the GDVN nozzle methods are prone to clogging or freezing when used in vacuum chambers and benefit from cleaning between samples.

For MHz sources, sample injection speeds need to be on the order of $50\text{--}100 \text{ m s}^{-1}$ so that fresh material is presented to each XFEL pulse [38, 90, 178–181]. Mixing-injectors that produce such high jet velocities must have very small orifices and therefore can only accommodate microcrystals. This matches the desire to use less than $\sim 3 \mu\text{m}$ crystals and is consistent with rapid equilibration of ligands throughout the crystal. These constraints also impact GDVN nozzle fabrication. For instance, a recent example developed for MHz data collection at the European XFEL used a high-end 3D printer (Nanoscribe GmbH) and two photon stereolithography methods to achieve free-form geometries with

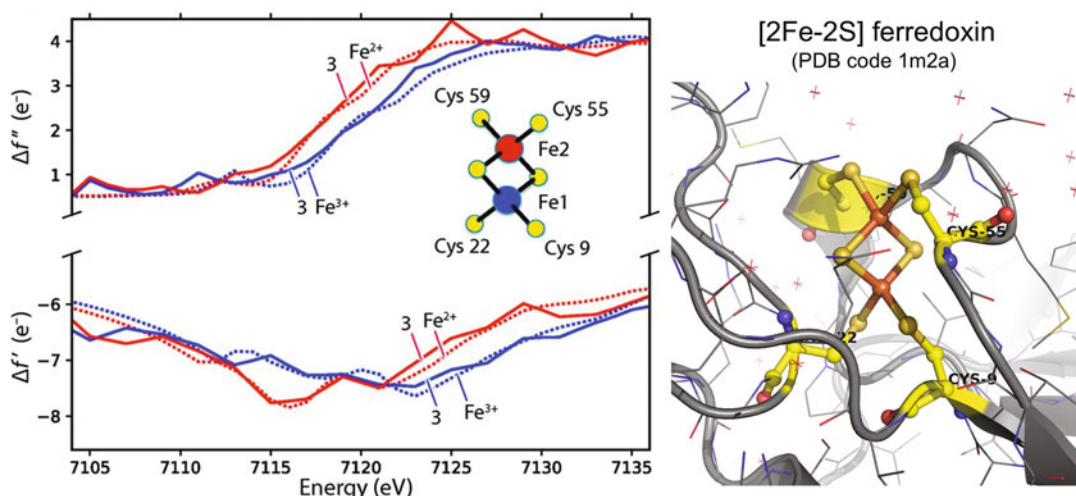


Fig. 7 Spatially resolved anomalous dispersion (SPREAD) and crystal structure enable redox assignment to metal ions. (Left) Simulated anomalous scattering curves for the two iron centers converge to the ground truth showing that the oxidation state difference between Fe²⁺ and Fe³⁺ is clearly revealed by the refined 3-macrocycle models. (Right) The 1.5 Å resolution crystal structure of the [2Fe-2S] ferredoxin from *Aquifex aeolicus* reveals the atomic position, but not the redox status of the metal atoms. Inspired by Sauter et al. (2020) *Acta Crystallogr D76*, 176–192 and PDB code 1m2a

submicron precision [40]. Devices like those illustrated in Fig. 7 produce 50–225 μm jet lengths, diameters as low as 536 ± 35 nm with a liquid flow rates of 2.4 ± 0.12 μl min⁻¹ and a gas flow rates of 22.5 ± 0.2 mg min⁻¹. At the SPB/SFX instrument at the European XFEL, a similar design with a gas and liquid orifice size of 60 and 50 μm diameter, respectively, were used to inject microcrystals of up to 6–8 μm in size at velocities of up to 100 m s⁻¹ [180, 181]. The delay time point(s) achievable for high velocity mix-inject jets is principally a function of the flow rate and the distance the sample-ligand mixture travels to the nozzle exit; the time of flight to the beam will only be ~1–3 μs and thus very short compared to most enzyme reaction times (average turnover time of ~60 ms).

3.3 “Multimessenger techniques” Help Reduce Ambiguity in Time-Resolved SFX Experiments

Time-resolved SFX that is correlated with time-resolved XES has been used extensively to study PS-II [49, 51, 97–102, 140]. This strategy provides data relevant to the electronic and atomic structures of metal centers. For these types of experiments, it is important to evaluate the impact of the XFEL pulse intensity and duration upon the metal center since the measurement itself will alter the electronic structure of the center. Under some circumstances, nonlinear excitation processes may disturb the spectroscopic signature of the metalloenzyme without significantly altering the recorded diffraction pattern.

The acoustic tape-drive system described above can also be used for mixing-based studies from metalloenzyme reactions, including those that react with O_2 (see Fig. 4). For example, iron-dependent enzymes are ubiquitous in biology. But it is difficult to study oxygen-Fe intermediates by traditional, synchrotron-based MX methods because their reactivity also makes them very sensitive to photoreduction by the X-ray beam. In many cases, an $Fe(IV)=O$ intermediate is produced in the reaction cycle; therefore, it is also important to include X-ray emission spectroscopy to help differentiate it from $Fe(II)-OH_2$ and $Fe(III)-OH$ species. The electron density maps for these three moieties are nearly identical except at extraordinarily high resolution. These types of compound are also very sensitive to X-ray radiation-induced artifacts, but the spectroscopic signatures are very different. This strategy provides critical correlations between atomic and electronic structures (especially first row transition metal centers) from the same sample and X-ray pulse. For instance, this also represents an opportunity to include time-resolved XES measurements into the time-resolved SFX data processing pipelines so that only diffraction patterns that also exhibit the appropriate spectroscopic signature are merged together into a time point dataset [102]. This has been applied recently to several metalloenzymes that react with O_2 including the soluble methane monooxygenase (sMMO) [182], ribonucleotide reductases (RNR) [51]; several hydrogenases, heme-based P450 enzymes, and several nonheme iron oxygenases or model systems [183]. Applications of this strategy will provide more confidence in the electron density map interpretation and the resulting atomic models deposited to the PDB that are released to the whole structural biology community.

The electronic structure of a metal center is an essential factor in reactivity and is profoundly influenced by its first and second shell coordination geometry, as well as the local electrostatic environment [184–187]. X-ray absorption spectroscopy (XAS) techniques complement crystallography, especially the analysis of the X-ray absorption near-edge structure (XANES) and the extended X-ray absorption fine structure (EXAFS) regions of the spectrum. The XANES portion provides insights into the oxidation state and coordination geometry, whereas XAFS region provides accurate measurements of metal–metal and metal–ligand distances. Indeed, the metric parameters that come from fits to the XAFS spectral region are often far more accurate than the errors in crystallographic methods will allow.

Nearly all types of spectroscopic measurements, including XES and XAS, record a signal from all of the particular metal atoms in the sample, whether it is ordered in the crystal lattice and/or enzyme active site (likely catalytically relevant) or disordered in the surrounding mother liquor (probably not mechanistically important). Thus, spectral overlap can be a challenge and is

particularly difficult for enzymes that contain multinuclear metal clusters wherein each metal atom may have a discreet redox and electronic structure (Fig. 7). An emerging method to resolve this ambiguity is to measure the XAS edge in the crystallographic diffraction data; a technique referred to as spatially resolved anomalous dispersion (SPREAD) [31, 188–190]. The method typically includes a series of complete diffraction datasets collected at several monochromatic energies across the K-absorption edge of the particular metal of interest. The electronic structure is then assigned to each individual metal center by refining the wavelength-dependent anomalous correction parameters in the datasets. This data collection strategy has been attempted at LCLS beamtime using seeded beam and SASE mode; and the preliminary results suggest that it is possible. In the meantime, recent theoretical calculations using 50,000–100,000 simulated diffraction patterns with *nanoBragg* demonstrate that SASE-based XFEL pulses with a 30 eV bandpass are suitable for analysis by SPREAD analysis of the [2Fe:2S] reduced ferredoxin containing an Fe(III)-Fe(II) center [31, 32]. The simulations also indicate that the incident spectrum for each XFEL pulse must also be included in the analysis [191]. Furthermore, radial streaks are observed in Bragg reflections collected with a long crystal to detector distance and are derived from the combined effects of the broad XFEL bandpass, crystal mosaicity, and energy-dependent structure factors [192].

In parallel and complementary to XFEL efforts, time-resolved MX methods are under development at synchrotron facilities too, especially those that provide pink-beam and microfocus capabilities [81, 82, 84, 193]. At XFEL or synchrotron facilities, the reactions in crystals must be synchronized throughout all unit cells in order to observe high resolution diffraction from reaction cycle intermediates. For mixing strategies, this will depend upon viscosity and the size of the substrate molecules traversing channels within the crystal lattice and macromolecules themselves [194–196]. Therefore, best practices dictate that when possible scientists should measure reactivity in the solid state and from more than one space group with different lattice packing. The homogeneity of the microcrystal slurry is also an important optimization parameter for mixing-based time-resolved SFX experiments [197, 198].

4 Notes

1. XFEL experiments and especially time-resolved SFX experiments are still at the cutting-edge for structural biology compared to routine and often automated data collection at synchrotron sources. The complexity of SFX and time-resolved SFX presents numerous challenges that all must be overcome for a successful outcome. A recent review by Kupitz and Sierra

outlines many important lessons learned from first-hand experience gained from many XFEL data collection opportunities [199]. One of the most important pieces of advice is to talk with instrument scientists and others in the field to help evaluate what is possible now, soon, or in the future. It is clear that the situation will continue to rapidly evolve and will soon yield methods, strategies, and instruments to produce time-resolved molecular movies of macromolecular systems engaged in function.

Acknowledgments

The authors acknowledge the financial support of this work from a Wellcome Investigator Award in Science 210734/Z/18/Z (to A. M.O.) and a Royal Society Wolfson Fellowship RSWF\R2\182017 (to A.M.O).

References

1. Goodsell DS, Zardecki C, Di Costanzo L et al (2020) RCSB protein data Bank: enabling biomedical research and drug discovery. *Protein Sci* 29:52–65
2. Consortium PDB (2019) Protein data bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* 47: D520–D528
3. Berman HM, Burley SK, Kleywegt GJ et al (2016) The archiving and dissemination of biological structure data. *Curr Opin Struct Biol* 40:17–22
4. Schlichting I (2015) Serial femtosecond crystallography: the first five years. *IUCrJ* 2 (Pt 2):246–255
5. Fromme P (2015) XFELs open a new era in structural chemical biology. *Nat Chem Biol* 11:895–899
6. Chapman HN (2019) X-ray free-electron lasers for the structure and dynamics of macromolecules. *Annu Rev Biochem* 88:35–58
7. Spence JCH (2017) XFELs for structure and dynamics in biology. *IUCrJ* 4:322–339
8. Breaker RR (2018) Riboswitches and translation control. *Cold Spring Harb Perspect Biol* 10(11)
9. Breaker RR (2012) Riboswitches and the RNA world. *Cold Spring Harb Perspect Biol* 4(2):a003566
10. Bhandari YR, Fan L, Fang X et al (2017) Topological structure determination of RNA using small-angle X-ray scattering. *J Mol Biol* 429:3635–3649
11. Stagno JR, Liu Y, Bhandari YR et al (2017) Structures of riboswitch RNA reaction states by mix-and-inject XFEL serial crystallography. *Nature* 541:242–246
12. Stagno JR, Bhandari YR, Conrad CE et al (2017) Real-time crystallographic studies of the adenine riboswitch using an X-ray free-electron laser. *FEBS J* 284:3374–3380
13. Henderson R (1995) The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q Rev Biophys* 28:171–193
14. Henderson R (1990) Cryoprotection of protein crystals against radiation-damage in electron and X-ray-diffraction. *Proc R Soc B Biol Sci* 241:6–8
15. de la Mora E, Coquelle N, Bury CS et al (2020) Radiation damage and dose limits in serial synchrotron crystallography at cryo- and room temperatures. *Proc Natl Acad Sci U S A* 117:4142–4151
16. Bury CS, Brookes-Bartlett C, Walsh SP et al (2018) Estimate your dose: RADDOS-3D. *Protein Sci* 27:217–228
17. Garman EF, Weik M (2017) Radiation damage in macromolecular crystallography. *Methods Mol Biol* 1607:467–489
18. Zeldin OB, Brockhauser S, Brembridge J et al (2013) Predicting the X-ray lifetime of

- protein crystals. *Proc Natl Acad Sci U S A* 110:20551–20556
19. Holton JM, Frankel KA (2010) The minimum crystal size needed for a complete diffraction data set. *Acta Crystallogr D Biol Crystallogr* 66:393–408
 20. Holton JM (2009) A beginner's guide to radiation damage. *J Synchrotron Radiat* 16:133–142
 21. Li D, Caffrey M (2020) Structure and functional characterization of membrane integral proteins in the lipid cubic phase. *J Mol Biol* 432:5104–5123
 22. Zhang Q, Cherezov V (2019) Chemical tools for membrane protein structural biology. *Curr Opin Struct Biol* 58:278–285
 23. Mishin A, Gusach A, Luginina A et al (2019) An outlook on using serial femtosecond crystallography in drug discovery. *Expert Opin Drug Discovery* 14:933–945
 24. Neutze R, Branden G, Schertler GF (2015) Membrane protein structural biology using X-ray free electron lasers. *Curr Opin Struct Biol* 33:115–125
 25. Weierstall U, James D, Wang C et al (2014) Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nat Commun* 5:3309
 26. Caffrey M (2015) A comprehensive review of the lipid cubic phase or in meso method for crystallizing membrane and soluble proteins and complexes. *Acta Crystallogr F Struct Biol Commun* 71:3–18
 27. Boutet S, Lomb L, Williams GJ et al (2012) High-resolution protein structure determination by serial femtosecond crystallography. *Science* 337:362–364
 28. Chapman HN, Fromme P, Barty A et al (2011) Femtosecond X-ray protein nanocrystallography. *Nature* 470:73–77
 29. Phillips GN (1995) XRayView: a teaching aid for X-ray crystallography. *Biophys J* 69:1281–1283
 30. Phillips GN (2011) XRayView, a virtual X-ray crystallography laboratory <http://www.phillipslab.org/downloads>, Access Data 14 Sep 2020
 31. Sauter NK, Kern J, Yano J, Holton JM (2020) Towards the spatial resolution of metalloprotein charge states by detailed modeling of XFEL crystallographic diffraction. *Acta Crystallogr D Struct Biol* 76:176–192
 32. Holton JM, Frankel KA (2020) SnanoBragg, a short program for calculation of absolute scattering from molecules and small crystals <https://bl831.als.lbl.gov/~jamesh/nanoBragg/>, Access Data 14 Sep 2020
 33. Zhao FZ, Zhang B, Yan E-K et al (2019) A guide to sample delivery systems for serial crystallography. *FEBS J* 286:4402–4417
 34. Davy B, Axford D, Beale JH et al (2019) Reducing sample consumption for serial crystallography using acoustic drop ejection. *J Synchrotron Radiat* 26:1820–1825
 35. Beale JH, Bolton R, Marshall SA et al (2019) Successful sample preparation for serial crystallography experiments. *J Appl Crystallogr* 52:1385–1396
 36. Coe J, Ros A (2018) Small is beautiful: growth and detection of nanocrystals. In: Boutet S, Fromme P, Hunter M (eds) *X-ray free electron lasers*. Springer, Cham, pp 59–85
 37. Kupitz C, Grotjohann I, Conrad CE et al (2014) Microcrystallization techniques for serial femtosecond crystallography using photosystem II from *Thermosynechococcus elongatus* as a model system. *Philos Trans R Soc Lond Ser B Biol Sci* 369:20130316
 38. Stan CA, Milathianaki D, Laksmono H et al (2016) Liquid explosions induced by X-ray laser pulses. *Nat Phys* 12:966–971
 39. Kim D, Echelmeir A, Villarreal J et al (2019) Electric triggering for enhanced control of droplet generation. *Anal Chem* 91:9792–9799
 40. Knoska J, Adriano L, Awel S et al (2020) Ultracompact 3D microfluidics for time-resolved structural biology. *Nat Commun* 11:657
 41. Oberthuer D, Knoška J, Wiedorn MO et al (2017) Double-flow focused liquid injector for efficient serial femtosecond crystallography. *Sci Rep* 7:44628
 42. Nogly P, Panneels V, Nelson G et al (2016) Lipidic cubic phase injector is a viable crystal delivery system for time-resolved serial crystallography. *Nat Commun* 7:12314
 43. DePonte DP, Weierstall U, Schmidt K et al (2008) Gas dynamic virtual nozzle for generation of microscopic droplet streams. *J Phys D Appl Phys* 41:195505
 44. Kovacsova G, Grünbein ML, Kloos M et al (2017) Viscous hydrophilic injection matrices for serial crystallography. *IUCrJ* 4:400–410
 45. Conrad CE, Basu S, James D et al (2015) A novel inert crystal delivery medium for serial femtosecond crystallography. *IUCrJ* 2:421–430
 46. Sierra RG, Gati C, Laksmono H et al (2016) Concentric-flow electrokinetic injector enables serial crystallography of ribosome and photosystem II. *Nat Methods* 13:59–62

47. Dao EH, Poitvin F, Sierra RG et al (2018) Structure of the 30S ribosomal decoding complex at ambient temperature. *RNA* 24:1667–1676
48. Tetreau G, Banneville A-S, Andreeva EA et al (2020) Serial femtosecond crystallography on in vivo-grown crystals drives elucidation of mosquitoicidal Cyt1Aa bioactivation cascade. *Nat Commun* 11:1153
49. Kern J, Chatterjee R, Young ID et al (2018) Structures of the intermediates of Kok's photosynthetic water oxidation clock. *Nature* 563:421–425
50. Orville AM (2017) Acoustic methods for on-demand sample injection into XFEL beams. In: *X-ray free electron lasers: applications in materials, chemistry and biology*. The Royal Society of Chemistry, pp 348–364
51. Fuller FD, Gul S, Chatterjee R et al (2017) Drop-on-demand sample delivery for studying biocatalysts in action at X-ray free-electron lasers. *Nat Methods* 14:443–449
52. Roessler CG, Agarwal R, Allaire M et al (2016) Acoustic injectors for drop-on-demand serial femtosecond crystallography. *Structure* 24:631–640
53. Soares AS, Engel MA, Stearns R et al (2011) Acoustically mounted microcrystals yield high-resolution X-ray structures. *Biochemistry* 50:4399–4401
54. Wu P, Noland C, Ultsch M et al (2016) Developments in the implementation of acoustic droplet ejection for protein crystallography. *J Lab Autom* 21:97–106
55. Mafune F, Miyajima K, Tono K et al (2016) Microcrystal delivery by pulsed liquid droplet for serial femtosecond crystallography. *Acta Crystallogr D Struct Biol* 72:520–523
56. Hadimioglu B, Stearns R, Ellson R (2016) Moving liquids with sound: the physics of acoustic droplet ejection for robust laboratory automation in life sciences. *J Lab Autom* 21:4–18
57. Hunter MS, Segelke B, Messerschmidt M et al (2014) Fixed-target protein serial microcrystallography with an x-ray free electron laser. *Sci Rep* 4:6026
58. Doak RB, Kovacs GN, Gorel A et al (2018) Crystallography on a chip—without the chip: sheet-on-sheet sandwich. *Acta Crystallogr Sect D Struct Biol* 74:1000–1007
59. Oghbaei S, Sarracini A, Ginn HM et al (2016) Fixed target combined with spectral mapping: approaching 100% hit rates for serial crystallography. *Acta Crystallogr D Struct Biol* 72:944–955
60. Sherrell DA, Foster AJ, Hudson L et al (2015) A modular and compact portable mini-endstation for high-precision, high-speed fixed target serial crystallography at FEL and synchrotron sources. *J Synchrotron Radiat* 22:1372–1378
61. Mueller C, Marx A, Epp SW et al (2015) Fixed target matrix for femtosecond time-resolved and in situ serial microcrystallography. *Struct Dyn* 2:054302
62. Aller P, Sanchez-Weatherby J, Foadi J et al (2015) Application of in situ diffraction in high-throughput structure determination platforms. *Methods Mol Biol* 1261:233–253
63. Lieske J, Cerv M, Kreida S et al (2019) On-chip crystallization for serial crystallography experiments and on-chip ligand-binding studies. *IUCrJ* 6:714–728
64. Roedig P, Ginn HM, Pakendorf T et al (2017) High-speed fixed-target serial virus crystallography. *Nat Methods* 14:805–810
65. Roedig P, Duman R, Weatherby-Sanchez J et al (2016) Room-temperature macromolecular crystallography using a micro-patterned silicon chip with minimal background scattering. *J Appl Crystallogr* 49:968–975
66. Roedig P, Vartiainen I, Duman R et al (2015) A micro-patterned silicon chip as sample holder for macromolecular crystallography experiments with minimal background scattering. *Sci Rep* 5:10451
67. Shelby ML, Gilbille D, Grant TD et al (2020) A fixed-target platform for serial femtosecond crystallography in a hydrated environment. *IUCrJ* 7:30–41
68. Chreifi G, Baxter EL, Doukov T et al (2016) Crystal structure of the pristine peroxidase ferryl center and its relevance to proton-coupled electron transfer. *Proc Natl Acad Sci U S A* 113:1226–1231
69. Baxter EL, Aguila L, Alonso-Mori R et al (2016) High-density grids for efficient data collection from multiple crystals. *Acta Crystallogr D Struct Biol* 72:2–11
70. Cohen AE, Soltis M, González A et al (2014) Goniometer-based femtosecond crystallography with X-ray free electron lasers. *Proc Natl Acad Sci U S A* 111:17122–17127
71. Zander U, Bourenkov G, Popov A et al (2015) MeshAndCollect: an automated multi-crystal data-collection workflow for synchrotron macromolecular crystallography beamlines. *Acta Crystallogr D Biol Crystallogr* 71:2328–2343
72. Orville AM (2020) Recent results in time resolved serial femtosecond crystallography at XFELs. *Curr Opin Struct Biol* 65:193–208

73. Sauter NK, Rose JP, Bhat TN (2020) Transactions from the 69th Annual Meeting of the American Crystallographic Association: data best practices-current state and future needs. *Struct Dyn* 7:021301
74. Bernstein HJ, Andrews LC, Diaz J et al (2020) Best practices for high data-rate macromolecular crystallography (HDRMX). *Struct Dyn* 7:014302
75. Leonarski F, Mozzanica A, Brückner M et al (2020) JUNGFR AU detector for brighter x-ray sources: solutions for IT and data science challenges in macromolecular crystallography. *Struct Dyn* 7:014305
76. Meisburger SP, Case DA, Ando N (2020) Diffuse X-ray scattering from correlated motions in a protein crystal. *Nat Commun* 11:1271
77. Forster A, Schulze-Bries C (2019) A shared vision for macromolecular crystallography over the next five years. *Struct Dyn* 6:064302
78. Helliwell JR, McMahon B, Guss JM et al (2017) The science is in the data. *IUCrJ* 4:714–722
79. Grimes JM, Hall DR, Ashton AW et al (2018) Where is crystallography going? *Acta Crystallogr D Struct Biol* 74:152–166
80. Martiel I, Muller-Werkmeister HM, Cohen AE (2019) Strategies for sample delivery for femtosecond crystallography. *Acta Crystallogr D Struct Biol* 75:160–177
81. Meents A, Wiedorn MO, Srajer V et al (2017) Pink-beam serial crystallography. *Nat Commun* 8:1281
82. Mehrabi P, Schultz EC, Agthe M et al (2019) Liquid application method for time-resolved analyses by serial synchrotron crystallography. *Nat Methods* 16:979–982
83. Martin-Garcia JM, Conrad CE, Nelson G et al (2017) Serial millisecond crystallography of membrane and soluble protein microcrystals using synchrotron radiation. *IUCrJ* 4:439–454
84. Martin-Garcia JM, Zhu L, Mendez D et al (2019) High-viscosity injector-based pink-beam serial crystallography of microcrystals at a synchrotron radiation source. *IUCrJ* 6:412–425
85. Chenevier D, Joly A (2018) ESRF: inside the extremely brilliant source upgrade. *Synchrotron Radiat News* 31:32–35
86. Chapon LC, Boscaro-Clarke I, Dent AJ, et al (2019) Diamond-II — Conceptual Design Report. Diamond Light Source Ltd.: Harwell Science & Innovation Campus, Didcot, Oxfordshire, OX11 0DE, UK
87. Boscaro-Clarke I, Evans G, Rambo R et al (2019) Diamond-II—Advancing Science. 2019, Diamond Light Source Ltd: Harwell Science & Innovation Campus, Didcot, Oxfordshire OX11 0DE, UK
88. Grünbein ML, Stricker M, Kovacs GN et al (2020) Illumination guidelines for ultrafast pump-probe experiments by serial femtosecond crystallography. *Nat Methods* 17:681–684
89. Johansson LC, Arnlund D, Katona G et al (2013) Structure of a photosynthetic reaction Centre determined by serial femtosecond crystallography. *Nat Commun* 4:2911
90. Wiedorn MO, Awel S, Morgan AJ et al (2018) Rapid sample delivery for megahertz serial crystallography at X-ray FELs. *IUCrJ* 5:574–584
91. Echelmeier A, Kim D, Villareal JC et al (2019) 3D printed droplet generation devices for serial femtosecond crystallography enabled by surface coating. *J Appl Crystallogr* 52:997–1008
92. Gisriel C, Coe J, Letrun R et al (2019) Membrane protein megahertz crystallography at the European XFEL. *Nat Commun* 10:5021
93. Aquila A et al (2012) Time-resolved protein nanocrystallography using an X-ray free-electron laser. *Opt Express* 20:2706–2716
94. Suga M, Hunter MS, Doak RB et al (2020) Time-resolved studies of metalloproteins using X-ray free electron laser radiation at SACLA. *Biochim Biophys Acta Gen Subj* 1864:129466
95. Suga M, Akita F, Yamashita K et al (2019) An oxyl/oxo mechanism for oxygen-oxygen coupling in PSII revealed by an x-ray free-electron laser. *Science* 366:334–338
96. Suga M, Akita F, Hirata K et al (2015) Native structure of photosystem II at 1.95 Å resolution viewed by femtosecond X-ray pulses. *Nature* 517:99–103
97. Young ID, Ibrahim M, Chatterjee R et al (2016) Structure of photosystem II and substrate binding at room temperature. *Nature* 540:453–457
98. Kern J, Yachandra VK, Yano J (2015) Metalloprotein structures at ambient conditions and in real-time: biological crystallography and spectroscopy using X-ray free electron lasers. *Curr Opin Struct Biol* 34:87–98
99. Kern J, Tran R, Alonos-Mori R et al (2014) Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy. *Nat Commun* 5:4371
100. Kern J, Alonso-Mori R, Tran R et al (2013) Simultaneous femtosecond X-ray

- spectroscopy and diffraction of photosystem II at room temperature. *Science* 340:491–495
101. Alonso-Mori R, Kern J, Gildea RJ et al (2012) Energy-dispersive X-ray emission spectroscopy using an X-ray free-electron laser in a shot-by-shot mode. *Proc Natl Acad Sci U S A* 109:19103–19107
 102. Fransson T, Chatterjee R, Fuller FD et al (2018) X-ray emission spectroscopy as an in situ diagnostic tool for X-ray crystallography of metalloproteins using an X-ray free-electron laser. *Biochemistry* 57:4629–4637
 103. Ibrahim M, Fransson T, Chatterjee R et al (2020) Untangling the sequence of events during the S2 → S3 transition in photosystem II and implications for the water oxidation mechanism. *Proc Natl Acad Sci U S A* 117:12624–12635
 104. Kupitz C, Bsu S, Grotjohann I et al (2014) Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. *Nature* 513:261–265
 105. Ayer K, Yefanov OM, Oberthür D et al (2016) Macromolecular diffractive imaging using imperfect crystals. *Nature* 530:202–206
 106. Suga M, Akita F, Sugahara M et al (2017) Light-induced structural changes and the site of O=O bond formation in PSII caught by XFEL. *Nature* 543:131–135
 107. Tenboer J, Basu S, Zatsepin N et al (2014) Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science* 346:1242–1246
 108. Pande K, Hutchinson CD, Groenhof G et al (2016) Femtosecond structural dynamics drives the trans/cis isomerization in photoactive yellow protein. *Science* 352:725–729
 109. Pandey S, Bean R, et al ST (2019) Time-resolved serial femtosecond crystallography at the European XFEL. *Nat Methods* 17:73–78
 110. Kang Y, Zhou XE, Gao X et al (2015) Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. *Nature* 523:561–567
 111. Zhou XE, Goa X, Barty A et al (2016) X-ray laser diffraction for structure determination of the rhodopsin-arrestin complex. *Sci Data* 3:160021
 112. Nakane T, Hanashima S, Suzuki M et al (2016) Membrane protein structure determination by SAD, SIR, or SIRAS phasing in serial femtosecond crystallography using an iododetergent. *Proc Natl Acad Sci U S A* 113:13039–13044
 113. Nango E, Royant A, Kubo M et al (2016) A three-dimensional movie of structural changes in bacteriorhodopsin. *Science* 354:1552–1557
 114. Nogly P, Weinert T, James D et al (2018) Retinal isomerization in bacteriorhodopsin captured by a femtosecond x-ray laser. *Science* 361:145
 115. Wickstrand C, Nogly P, Nango E et al (2019) Bacteriorhodopsin: structural insights revealed using X-ray lasers and synchrotron radiation. *Annu Rev Biochem* 88:59–83
 116. Weinert T, Skopintsev P, James D et al (2019) Proton uptake mechanism in bacteriorhodopsin captured by serial synchrotron crystallography. *Science* 365:61–65
 117. Panneels V, Wu W, Tsai C-J et al (2015) Time-resolved structural studies with serial crystallography: a new light on retinal proteins. *Struct Dyn* 2:041718
 118. Nass Kovacs G, Colletier J, Grünbein ML et al (2019) Three-dimensional view of ultrafast dynamics in photoexcited bacteriorhodopsin. *Nat Commun* 10:3177
 119. Varma N, Mutt E, Mühle J et al (2019) Crystal structure of jumping spider rhodopsin-1 as a light sensitive GPCR. *Proc Natl Acad Sci U S A* 116:14547–14556
 120. Nagata T, Koyanagi M, Tsukamoto H et al (2019) The counterion-retinylidene Schiff base interaction of an invertebrate rhodopsin rearranges upon light activation. *Commun Biol* 2:180
 121. Yun JH, Li X, Park J-H et al (2019) Non-cryogenic structure of a chloride pump provides crucial clues to temperature-dependent channel transport efficiency. *J Biol Chem* 294:794–804
 122. Hutchison CD, Cordon-Preciado V, Morgan RM et al (2017) X-ray free electron laser determination of crystal structures of dark and light states of a reversibly photoswitching fluorescent protein at room temperature. *Int J Mol Sci* 18:1918
 123. Woodhouse J, Nass Kovacs G, Coquelle N et al (2020) Photoswitching mechanism of a fluorescent protein revealed by time-resolved crystallography and transient absorption spectroscopy. *Nat Commun* 11:741
 124. Colletier JP, Sliwa M, Gallat F-X et al (2016) Serial femtosecond crystallography and ultrafast absorption spectroscopy of the photoswitchable fluorescent protein IrisFP. *J Phys Chem Lett* 7:882–887
 125. Coquelle N, Sliwa M, Woodhouse J et al (2018) Chromophore twisting in the excited state of a photoswitchable fluorescent protein

- captured by time-resolved serial femtosecond crystallography. *Nat Chem* 10:31–37
126. Edlund P, Takala H, Claesson E et al (2016) The room temperature crystal structure of a bacterial phytochrome determined by serial femtosecond crystallography. *Sci Rep* 6:35279
 127. Claesson E, Wahlgren WY, Takal H et al (2020) The primary structural photoresponse of phytochrome proteins captured by a femtosecond X-ray laser. *elife* 9:e53514
 128. Barends TR, Foucar L, Ardevol A et al (2015) Direct observation of ultrafast collective motions in CO myoglobin upon ligand dissociation. *Science* 350:445–450
 129. Ishigami I, Zatsepin NA, Hikita M et al (2017) Crystal structure of CO-bound cytochrome c oxidase determined by serial femtosecond X-ray crystallography at room temperature. *Proc Natl Acad Sci U S A* 114:8011–8016
 130. Shimada A, Kubo M, Baba S et al (2017) A nanosecond time-resolved XFEL analysis of structural changes associated with CO release from cytochrome c oxidase. *Sci Adv* 3:e1603042
 131. Nakajima K, Joti Y, Katayama T et al (2018) Software for the data analysis of the arrival-timing monitor at SACLA. *J Synchrotron Radiat* 25:592–603
 132. Katayama T, Owada S, Togashi T et al (2016) A beam branching method for timing and spectral characterization of hard X-ray free-electron lasers. *Struct Dyn* 3:034301
 133. Sanchez-Gonzalez A, Johnson AS, Fitzpatrick A et al (2017) Coincidence timing of femtosecond optical pulses in an X-ray free electron laser. *J Appl Phys* 122:203105
 134. Yabuuchi T, Kon A, Inubushi Y et al (2019) An experimental platform using high-power, high-intensity optical lasers with the hard X-ray free-electron laser at SACLA. *J Synchrotron Radiat* 26:585–594
 135. Roessler CG, Kuczewski A, Stearns R et al (2013) Acoustic methods for high-throughput protein crystal mounting at next-generation macromolecular crystallographic beamlines. *J Synchrotron Radiat* 20:805–808
 136. Yano J, Yachandra V (2014) Mn4Ca cluster in photosynthesis: where and how water is oxidized to dioxygen. *Chem Rev* 114:4175–4205
 137. Hillier W, Wydrzynski T (2008) O-18-water exchange in photosystem II: substrate binding and intermediates of the water splitting cycle. *Coord Chem Rev* 252:306–317
 138. Alonso-Mori R, Sokaras D, Zhu D et al (2015) Photon-in photon-out hard X-ray spectroscopy at the Linac coherent light source. *J Synchrotron Radiat* 22:612–620
 139. Alonso-Mori R, Asa K, Bergmann U et al (2016) Towards characterization of photo-excited electron transfer and catalysis in natural and artificial systems using XFELs. *Faraday Discuss* 194:621–638
 140. Jensen SC, Sullivan B, Hartzler DA et al (2019) X-ray emission spectroscopy at X-ray free electron lasers: limits to observation of the classical spectroscopic response for electronic structure analysis. *J Phys Chem Lett* 10:441–446
 141. Tosha T, Nomura T, Nishida T et al (2017) Capturing an initial intermediate during the P450_{nor} enzymatic reaction using time-resolved XFEL crystallography and caged-substrate. *Nat Commun* 8:1585
 142. Deiters A, Groff D, Ryu Y et al (2006) A genetically encoded photocaged tyrosine. *Angew Chem Int Ed Engl* 45:2728–2731
 143. Wang J, Liu Y, Liu Y et al (2019) Time-resolved protein activation by proximal decaging in living systems. *Nature* 569:509–513
 144. Givens RS, Rubina M, Wirz J (2012) Applications of p-hydroxyphenacyl (pHP) and coumarin-4-ylmethyl photoremovable protecting groups. *Photochem Photobiol Sci* 11:472–488
 145. Johnson LN (1992) Time-resolved protein crystallography. *Protein Sci* 1:1237–1243
 146. Austin RH, Beeson KW, Eisenstein L et al (1975) Dynamics of ligand binding to myoglobin. *Biochemistry* 14:5355–5373
 147. Mondal P, Meuwly M (2018) Solvent composition drives the rebinding kinetics of nitric oxide to microperoxidase. *Sci Rep* 8:5281
 148. Murakawa Y, Nagai M, Mizutani Y (2012) Differences between protein dynamics of hemoglobin upon dissociation of oxygen and carbon monoxide. *J Am Chem Soc* 134:1434–1437
 149. Beece D, Eisenstein J, Frauenfelder D et al (1979) Dioxygen replacement reaction in myoglobin. *Biochemistry* 18:3421–3423
 150. Flanagan JC, Baiz CR (2019) Ultrafast pH-jump two-dimensional infrared spectroscopy. *Opt Lett* 44:4937–4940
 151. Abbruzzetti S, Sottini S, Viappiani C, Corrie JE (2005) Kinetics of proton release after flash photolysis of 1-(2-nitrophenyl)ethyl sulfate (caged sulfate) in aqueous solution. *J Am Chem Soc* 127:9865–9874
 152. Thompson MC, Barad BA, Wolff AM et al (2019) Temperature-jump solution X-ray

- scattering reveals distinct motions in a dynamic enzyme. *Nat Chem* 11:1058–1066
153. Keedy DA, Kenner LR, Warkentin M et al (2015) Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *elife* 4: e07574
 154. Reddish MJ, Callender R, Dyer RB (2017) Resolution of submillisecond kinetics of multiple reaction pathways for lactate dehydrogenase. *Biophys J* 112:1852–1862
 155. Winter MB, Herzik MA, Kuriyan J, Marletta MA (2011) Tunnels modulate ligand flux in a heme nitric oxide/oxygen binding (H-NOX) domain. *Proc Natl Acad Sci U S A* 108: E881–E889
 156. Alberding N, Frauenfelder H, Hanggi P (1978) Stochastic theory of ligand migration in biomolecules. *Proc Natl Acad Sci U S A* 75:26–29
 157. Alberding N, Austin RH, Chan SS et al (1978) Fast reactions in carbon monoxide binding to heme proteins. *Biophys J* 24:319–334
 158. Kanehisa M, Sato Y, Furumichi M et al (2019) New approach for understanding genome variations in KEGG. *Nucleic Acids Res* 47:D590–D595
 159. Du J, Yuan Z, Ma Z et al (2014) KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a PATH analysis model. *Mol BioSyst* 10:2441–2447
 160. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27–30
 161. Davidi D, Longo LM, Jabłońska J et al (2018) A Bird's-eye view of enzyme evolution: chemical, physicochemical, and physiological considerations. *Chem Rev* 118:8786–8797
 162. Bar-Even A, Noor E, Savir Y et al (2011) The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* 50:4402–4410
 163. Jeske L, Placzek S, Schomburg I et al (2019) BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res* 47: D542–D549
 164. Walsh C (2001) Enabling the chemistry of life. *Nature* 409:226–231
 165. Benkovic SJ, Hammes-Schiffer S (2003) A perspective on enzyme catalysis. *Science* 301:1196–1202
 166. Warshel A, Bora RP (2016) Perspective: defining and quantifying the role of dynamics in enzyme catalysis. *J Chem Phys* 144:180901
 167. Agarwal PK (2019) A biophysical perspective on enzyme catalysis. *Biochemistry* 58:438–449
 168. Henzler-Wildman KA, Lei M, Thai V et al (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* 450:913–916
 169. Milo R, Jorgensen P, Moran U et al (2010) BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res* 38(Database issue):D750–D753
 170. Schmidt M (2020) Reaction initiation in enzyme crystals by diffusion of substrate. *Crystals* 10:116
 171. Schmidt M, Saldin DK (2014) Enzyme transient state kinetics in crystal and solution from the perspective of a time-resolved crystallographer. *Struct Dyn* 1:024701
 172. Schmidt M (2013) Mix and inject: reaction initiation by diffusion for time-resolved macromolecular crystallography. *Adv Condens Matter Phys* 5-6:1–10
 173. Olmos JL, Pandey S, Martin-Garcia JM et al (2018) Enzyme intermediates captured "on the fly" by mix-and-inject serial crystallography. *BMC Biol* 16:59
 174. Kupitz C, Olmos JL, Holl M et al (2017) Structural enzymology using X-ray free electron lasers. *Struct Dyn* 4:044003
 175. Calvey GD, Katz AM, Pollack L (2019) Microfluidic mixing injector holder enables routine structural enzymology measurements with mix-and-inject serial crystallography using X-ray free electron lasers. *Anal Chem* 91:7139–7144
 176. Calvey GD, Katz AM, Schaffer CB, Pollack L (2016) Mixing injector enables time-resolved crystallography with high hit rate at X-ray free electron lasers. *Struct Dyn* 3:054301
 177. Monteiro DCF, Vakili M, Harich J et al (2019) A microfluidic flow-focusing device for low sample consumption serial synchrotron crystallography experiments in liquid flow. *J Synchrotron Radiat* 26:406–412
 178. Grunbein ML, Nass Kovacs G (2019) Sample delivery for serial crystallography at free-electron lasers and synchrotrons. *Acta Crystallogr D Struct Biol* 75:178–191
 179. Grunbein ML, Shoeman RL, Doak RB (2018) Velocimetry of fast microscopic liquid jets by nanosecond dual-pulse laser illumination for megahertz X-ray free-electron lasers. *Opt Express* 26:7190–7203
 180. Grünbein ML, Bielecki J, Gorel A et al (2018) Megahertz data collection from protein microcrystals at an X-ray free-electron laser. *Nat Commun* 9:3487

181. Wiedorn MO, Oberthür D, Bean R et al (2018) Megahertz serial crystallography. *Nat Commun* 9:4025
182. Srinivas V, Banerjee R, Lebrette H et al (2020) High-resolution XFEL structure of the soluble methane monooxygenase hydroxylase complex with its regulatory component at ambient temperature in two oxidation states. *J Am Chem Soc* 142:14249–14266
183. Miller KR, Paretsky JD, Follmer AH et al (2020) Artificial iron proteins: modeling the active sites in non-heme dioxxygenases. *Inorg Chem* 59:6000–6009
184. Mara MW, Hadt RG, Reinhard ME et al (2017) Metalloprotein entatic control of ligand-metal bonds quantified by ultrafast x-ray spectroscopy. *Science* 356:1276–1280
185. Holm RH, Solomon EI (2014) Introduction: bioinorganic enzymology II. *Chem Rev* 114:3367–3368
186. Solomon EI, Szilagyi RK, George S et al (2004) Electronic structures of metal sites in proteins and models: contributions to function in blue copper proteins. *Chem Rev* 104:419–458
187. Holm RH, Kennepohl P, Solomon EIS (1996) Structural and functional aspects of metal sites in biology. *Chem Rev* 96:2239–2314
188. Einsle O, Andrade SL, Dobbek H et al (2007) Assignment of individual metal redox states in a metalloprotein by crystallographic refinement at multiple X-ray wavelengths. *J Am Chem Soc* 129:2210–2211
189. Spatzal T, Schlesier J, Burger E-M et al (2016) Nitrogenase FeMoc investigated by spatially resolved anomalous dispersion refinement. *Nat Commun* 7:10902
190. Zhang L, Kaiser JT, Meloni G et al (2013) The sixteenth iron in the nitrogenase MoFe protein. *Angew Chem Int Ed Engl* 52:10529–10532
191. Zhu DL, Cammarata M, Feldkamp JM et al (2012) A single-shot transmissive spectrometer for hard x-ray free electron lasers. *Appl Phys Lett* 101:034103
192. Hattne J, Echols N, Tran R et al (2014) Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers. *Nat Methods* 11:545–548
193. Sanchez-Weatherby J, Sandy J, Mikolajek H et al (2019) VMXi: a fully automated, fully remote, high-flux in situ macromolecular crystallography beamline. *J Synchrotron Radiat* 26:291–301
194. Pravda L, Berka K, Svobodová R et al (2014) Anatomy of enzyme channels. *BMC Bioinformatics* 15:379
195. Juers DH, Ruffin J (2014) MAP_CHANNELS: a computation tool to aid in the visualization and characterization of solvent channels in macromolecular crystals. *J Appl Crystallogr* 47:2105–2108
196. Coleman RG, Sharp KA (2009) Finding and characterizing tunnels in macromolecules with application to ion channels and pores. *Biophys J* 96:632–645
197. Heymann M, Opatthalage A, Wierman JL et al (2014) Room-temperature serial crystallography using a kinetically optimized microfluidic device for protein crystallization and on-chip X-ray diffraction. *IUCrJ* 1:349–360
198. Abdallah BG, Zatespin NA, Roy-Chowdhury S et al (2015) Microfluidic sorting of protein nanocrystals by size for X-ray free-electron laser diffraction. *Struct Dyn* 2:041719
199. Kupitz C, Sierra RG (2020) Preventing bio-bloopers and XFEL follies: best practices from your friendly instrument staff. *Crystals* 10:251
200. Weierstall U (2014) Liquid sample delivery techniques for serial femtosecond crystallography. *Philos Trans R Soc Lond Ser B Biol Sci* 369:20130337
201. Liu W, Wacker D, Gati C et al (2013) Serial femtosecond crystallography of G protein-coupled receptors. *Science* 342:1521–1524
202. Dasgupta M, Budday D, de Oliveira SH et al (2019) Mix-and-inject XFEL crystallography reveals gated conformational dynamics during enzyme catalysis. *Proc Natl Acad Sci U S A* 116:25634–25640
203. Rabe P, Beale JH, Butryn A et al (2020) Anaerobic fixed-target serial crystallography. *IUCrJ* 7:901–912
204. Ebrahim A, Moreno-Chicano T, Appleby MV et al (2019) Dose-resolved serial synchrotron and XFEL structures of radiation-sensitive metalloproteins. *IUCrJ* 6:543–551
205. Owen RL, Axford D, Sherrell DA et al (2017) Low-dose fixed-target serial synchrotron crystallography. *Acta Crystallogr D Struct Biol* 73:373–378
206. Wang D, Weierstall U, Pollack L, Spence J (2014) Double-focusing mixing jet for XFEL study of chemical kinetics. *J Synchrotron Radiat* 21:1364–1366
207. Bohne S, Heymann M, Chapman HN et al (2019) 3D printed nozzles on a silicon fluidic chip. *Rev Sci Instrum* 90:035108
208. Nelson G, Kirian RA, Weierstall U et al (2016) Three-dimensional-printed gas

- dynamic virtual nozzles for x-ray laser sample delivery. *Opt Express* 24:11515–11530
209. Trebbin M, Krüger K, DePonte D et al (2014) Microfluidic liquid jet system with compatibility for atmospheric and high-vacuum conditions. *Lab Chip* 14:1733–1745
210. Burgie ES, Clinger JA, Miller MD et al (2020) Photoreversible interconversion of a phytochrome photosensory module in the crystalline state. *Proc Natl Acad Sci U S A* 117:300–307



Chapter 12

From Tube to Structure: SPA Cryo-EM Workflow Using Apoferritin as an Example

Christoph A. Diebolder, Rebecca S. Dillard, and Ludovic Renault

Abstract

In this chapter, we present an overview of a standard protocol to achieve structure determination at high resolution by Single Particle Analysis cryogenic Electron Microscopy using apoferritin as a standard sample. The purified apoferritin is applied to a glow-discharged support and then flash frozen in liquid ethane. The prepared grids are loaded into the electron microscope and checked for particle spreading and ice thickness. The microscope alignments are performed and the data collection session is setup for an overnight data collection. The collected movies containing two-dimensional images of the apoferritin sample are then processed to obtain a high-resolution three-dimensional reconstruction.

Key words Single Particle Analysis, Standard Protocol, Workflow, Cryo-EM, Cryogenic Electron Microscopy, Sample Preparation

1 Introduction

The field of single particle cryogenic electron microscopy (cryo-EM) has seen tremendous developments over the last decade. Prior to 2010 it was a rare occasion to resolve cryo-EM structure at resolutions better than 4 Angstroms and those were mostly limited to large icosahedral viruses. Recent developments in sample preparation, data collection strategies, cameras, and data processing software have opened the path to resolving high-resolution structure of a larger group of samples such as small asymmetrical objects [1, 2].

Samples for single particle cryo-EM are typically prepared by first applying a 3–5 μ l volume of a purified protein to a hydrophilic grid. After application, the grid is blotted to remove excess liquid and plunged into a liquid cryogen (ethane or a mixture of ethane and propane) for rapid vitrification. The grid is maintained at liquid nitrogen temperatures, preserving the sample in a thin film of vitreous ice [3–6]. Here, we will describe this workflow in detail using an apoferritin sample and the Thermo Fisher Scientific Vitrobot Mk IV as an example.

Many exciting new sample preparation methods have recently been developed that have the potential to further improve the cryo-EM workflow. These include alterations to the grids, such as all gold grids for reducing beam-induced motion of the sample [7], graphene substrates as support films [8–12] and affinity grids, which use a grid substrate such as Ni-NTA lipid monolayers [13, 14], streptavidin crystals [15–20], or bound antibodies [21–23] to purify sample particles or immobilize them and therefore prevent interactions with the air–water interface [24]. There has also been recent success with new instruments for the vitrification process that require smaller sample volumes and provide reproducibly thin ice on a faster and more controlled timescale [25–31]. These developments will likely be very useful for many samples when traditional methods produce suboptimal results.

With the newest developments in Electron Microscope software operations (“auto-alignments”) as well as a diversification and standardization of “on-the-fly” processing pipelines, it has now become possible to obtain robust and reliable fully automated reconstructions for standard samples such as apoferritin. The newest generation cameras are also now bigger and faster and can collect up to 250 movies per hour, allowing “sample to structure” in less than half a day. This changes the way researchers approach the data collection time at high-end instruments and opens up new possibilities for drug design and conformational studies using cryo-EM.

In this chapter, we present a detailed protocol on how to prepare apoferritin samples with a Vitrobot freezing device and how to collect high-resolution data on a Titan Krios equipped with a K2 bioquantum detector. General steps of the image processing workflow are briefly described.

2 Materials

2.1 Reagents

1. Purified apoferritin from equine spleen (Sigma A3660).
2. Liquid nitrogen.
3. Ethane.

2.2 Equipment

1. Quantifoil R 2/2200 mesh copper grids (Quantifoil Micro Tools GmbH, Jena, Germany).
2. PELCO easiGlow Glow Discharge Cleaning System (Ted Pella, Inc., Redding, CA, USA).
3. Thermo Fisher Scientific Vitrobot (Thermo Fisher Scientific, Waltham, MA, USA).
4. Standard Vitrobot filter paper, 55/20 mm, Grade 595 (Thermo Fisher Scientific, Waltham, MA, USA) or

Whatman Filter Paper, 55 mm, Grade 1 (GE Healthcare Life Sciences, Chicago, IL, USA).

5. 60 ml syringe.
6. Tweezers for grid handling, e.g., Dumont #5 (Electron Microscopy Sciences, Hatfield, PA, USA).
7. Dressing forceps (Ted Pella, Inc., Product Number 13268).
8. Cryo grid boxes with lid, round (Agar Scientific, Code AGG3727).
9. Screw driver.
10. C-clips (Thermo Fisher Scientific, Part Number 9432 909 97551).
11. C-clip rings (Thermo Fisher Scientific, Part Number 9432 909 97561).
12. C-clip insertion tools (Thermo Fisher Scientific, Part Number 9432 909 97571).
13. AutoGrid assembly workstation (Thermo Fisher Scientific, Part Number 1000068).
14. AutoGrid tweezers (Thermo Fisher Scientific, Part Number 9432 909 97631).
15. AutoGrid containers (Thermo Fisher Scientific, Part Number 9432 909 97621).
16. Gripper tool for AutoGrid containers (Thermo Fisher Scientific, Part Number 9432 909 97671).
17. Cryo dewars.
18. Cryo containers.
19. 10 μ l pipette and tips.
20. Cassette loading station (Thermo Fisher Scientific, Part Number 9432 909 97601).
21. AutoGrid cassette (Thermo Fisher Scientific, Part Number 9432 909 97581).
22. Cassette tweezers (Thermo Fisher Scientific, Part Number 9432 909 97651).
23. NanoCab (Thermo Fisher Scientific, Part Number 9432 909 97591).
24. Surgical masks.
25. ThermoFisher Titan Krios Transmission electron microscope with SFEG electron source, three condenser lens illumination system, Volta phase plate, Gatan K2 summit direct electron detector with Bioquantum energy filter for zero loss filtering.
26. Calibration grid (e.g., Sigma S106 cross grating with latex beads).

2.3 Software

1. TEM User Interface (TUI) 2.7.1.20333REL
2. FLuCam viewer (FluCam) 2.7.1.20333FEL
3. Tecnai Imaging & Analysis (TIA) 4.15
4. EPU Automated Single Particles Acquisition Software (EPU) 1.10
5. Digital Micrograph (DM) 2.33.1084.0
6. AutoCTF 0.6.7

3 Methods

3.1 Preparation of Sample Grids

3.1.1 Glow Discharge

1. Using grid handling tweezers, transfer grids from the box to a glass slide. Handle the grids only by the rim and place them with the carbon film facing up. If you have difficulty distinguishing between the two faces of the grid, it may help to note that the support film typically faces the center of a new box of grids (*see* **Note 1**).
2. Place the glass slide with grids on the pedestal of the PELCO easiGlow and cover with the glass chamber.
3. Turn on the PELCO easiGlow using the power switch.
4. Go to the Main menu, then to Protocols, and select a protocol.
5. To set up or adjust an existing protocol, press “Program Screen”. See Table 1 for an example protocol (*see* **Note 2**).
6. To run the protocol, return to the protocol screen and press “Run Screen”, then “Auto Run”. The status of each step of the protocol can be monitored on the display screen.

Table 1
Example protocol for glow discharging grids using the PELCO easiGlow Glow Discharge

Cleaning system	Glow polarity	Discharge current (mA)	Time (s)	Vacuum	Gas inlet
1 Ultimate pressure	Negative	0	30	0.39 mBar	Gas 1 (open to air)
2 Stable pressure	Negative	0	30	0.39 mBar	Gas 1 (open to air)
3 Hold pressure	Negative	0	30	0.39 mBar	Gas 1 (open to air)
4 Glow discharge	Negative	25	60	0.39 mBar	Gas 1 (open to air)
5 Hold pressure	Negative	0	5	0.39 mBar	Gas 1 (open to air)
6 Vent w/ pump off	Negative	0	30	ATM	Gas 1 (open to air)
7 End protocol	Negative	0	30	ATM	Gas 1 (open to air)

7. After the venting has completed, remove the glass chamber and retrieve the glass slide with grids, then place them aside until needed.
8. Replace the glass chamber and turn off the unit using the power switch.

3.1.2 Vitrification

1. Turn on the Vitrobot [33] using the power switch on the back of the unit and allow it to startup.
2. To attach the humidifier, first check that the O-ring is correctly placed to ensure that a proper seal will form.
3. Align the black dot on the humidifier with the dot on the Vitrobot just beneath the main chamber, then insert it upwards and rotate to lock it in place.
4. Draw up 50 ml of water using a 60 ml syringe, then attach the syringe to the tube at the bottom of the humidifier.
5. Use the syringe to fill the humidifier with 40–50 ml of water. Be careful not to overfill, as this may cause water to pool within the main chamber of the Vitrobot or below the humidifier. Draw out the syringe until water begins to come back out of the humidifier. The humidifier is then properly filled and the syringe can be detached.
6. In the Console tab of the Vitrobot display, set the temperature to 4 °C. Set the humidity to 95% by touching the black arrows and select the “On” radio button (*see Note 3*). These parameters may need to be adjusted based on differences in the experiment, sample, and instrument. Make sure to allow sufficient time for the instrument to reach the desired conditions while setting up the rest of the experiment.
7. In the Options tab of the Vitrobot display, adjust the Blot Time (s), Wait Time(s), Drain Times(s), Blot Force, and Blot Total (*see Note 4*).
8. Select options in the Miscellaneous panel of the Options tab based on personal preferences.
9. Prepare the blotting pads by securing blotting paper to them using the white attachment rings. It is advised to do this far enough in advance of the vitrification process to ensure that the paper has time to equilibrate with the temperature and humidity settings. This provides better reproducibility as the saturation of the blotting paper should then be consistent between grids.

3.1.3 Prepare the Cryogen Cup

1. Fill a 4 L cryodewar with liquid nitrogen. If a 4 L dewar is difficult to handle, several smaller cryo containers can also be used.

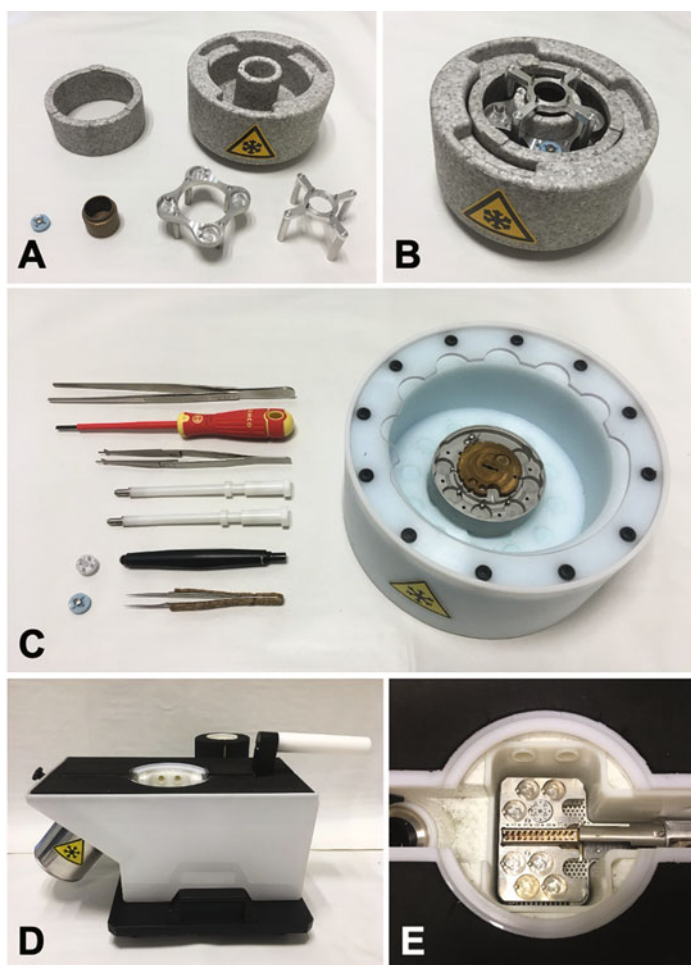


Fig. 1 Tools for sample vitrification, AutoGrid clipping, and cassette loading. **(a)** The components of the Vitrobot cryogen cup, clockwise from the top left: anticontamination ring, cryogen cup, metal spindle, grid box platform, brass ethane cup, and cryo grid box. The cryogen cup is shown assembled in **(b)**. **(c)** AutoGrid clipping station (right) and associated tools, from top to bottom: dressing forceps, screwdriver, AutoGrid tweezers, c-clip insertion tools, gripper tool, and grid handling tweezers. Shown on the far left are an AutoGrid box (top) and a cryo grid box (bottom). **(d)** An assembled cassette loading station with the nanocab attached. **(e)** View of the cassette loading station chamber illustrating the orientation of the cassette

2. Assemble the cryogen cup, including the brass ethane cup, grid box platform, grid box, metal spindle, and anticontamination ring, as shown in Fig. 1a, b.
3. From this point forward, it may be helpful to wear a surgical facemask to prevent condensation from your breath from contaminating the liquid nitrogen.

4. Cool the assembled cryogen cup with liquid nitrogen from either the cryodewar or cryo containers. First fill the brass ethane cup with liquid nitrogen, then the surrounding outer ring. The liquid nitrogen will initially boil as the components of the cup are cooled. Allow the liquid nitrogen to evaporate from the brass ethane cup, but continue to fill the outer ring of the cryogen cup to create a liquid nitrogen bath that covers the grid box platform. Allow the temperature to stabilize for a few minutes until the level of liquid nitrogen is maintained with minimal bubbling.
5. Insert the tip of a tube connected to an ethane tank into the brass ethane cup, touching the tip to the bottom of the cup, then open the regulator valve on the ethane tank. If the cup has reached the appropriate temperature, ethane will quickly begin to condense in the cup. This can be observed by both liquid formation and a bubbling sound. Fill the brass ethane cup with liquid ethane by holding the tip of the ethane tube against the bottom or side of the cup. Once the bubbles from the ethane begin to touch the metal spindle at the top of the cup, slowly remove the tip of the ethane tube from the condensed ethane while simultaneously closing the regulator valve on the ethane tank. This will prevent liquid ethane from being drawn back into the ethane tube.
6. Ensure the ethane has properly cooled by waiting for a thin white layer of frozen ethane to form on the inner surfaces of the brass ethane cup.
7. Remove the metal spindle from the cryogen cup with dressing forceps. If it has frozen to the brass ethane cup, use the warm tongs or another upside-down ethane cup to help release it.

3.1.4 Plunge Grids

1. Pick up a glow-discharged grid with the Vitrobot tweezers, only grasping the grid at the rim to prevent damaging the support film. Secure the grid by sliding the tweezer lock down to the first notch.
2. On the Vitrobot display, press “Place New Grid” to position the plunging rod for loading the tweezers. Gently attach the tweezers to the rod. The grid can be picked up and positioned with the carbon film facing either direction, depending on the handedness of the user. Be aware of the position of the grid at all times, so that it is not accidentally damaged. Press “Continue” to bring the rod with attached tweezers and grid into the main chamber of the Vitrobot.
3. Place the cooled cryogen cup onto the pedestal of the Vitrobot. Press “Continue” to raise the cryogen cup to the main chamber.

4. Press “Continue” to lower the grid into the sample loading position. Open the window on the side of the Vitrobot to which the carbon film of the grid faces and apply ~3 μL of apoferritin sample to the carbon film of the grid (*see Note 3*).
5. Press “Continue” to initiate blotting and plunging of the grid into the liquid ethane within the cryogen cup. After the plunging has occurred, the rod and cryogen cup will be lowered from the main chamber to allow user access.
6. Carefully remove the tweezers from the rod and transfer the cryogen cup to the benchtop, keeping the grid submerged in the liquid ethane.
7. Release the lock on the tweezers while maintaining pressure to keep the grid from dropping into the liquid ethane.
8. Use the tweezers to quickly transfer the grid to the surrounding liquid nitrogen bath and release it into a storage space in the cryo grid box.
9. Repeat this process for each grid to be vitrified.
10. Close the cryo grid box lid with a precooled screwdriver. A pair of dressing forceps can be used to keep the grid box lid opening aligned with the notch in the grid box while closing to prevent grids from falling out.
11. Transfer the cryo grid box to liquid nitrogen for storage.
12. Shut down the Vitrobot.
13. Place the cryogen cup in a safe space to allow the ethane to evaporate.
14. Remove the blotting papers from the pads in the Vitrobot and discard.
15. Detach the humidifier and pour out the remaining water, then set it aside to allow it to dry.
16. Press “Exit” on the display screen and confirm that the tweezers have been removed. Once the shutdown procedure has completed, turn off the unit using the power switch on the back.

3.1.5 Clipping

1. Ensure that there is enough liquid nitrogen available in a 4-liter cryodewar or several small cryo containers to keep the Auto-Grid assembly workstation cold throughout the clipping process.
2. Load c-clips into the opening at the end of the c-clip insertion tools using a pair of tweezers. Press the end of each tool to align the c-clips at the tip. Be careful not to completely eject the c-clips from the insertion tools.
3. Cool the AutoGrid assembly workstation with liquid nitrogen. Once cooled, maintain the liquid nitrogen just below the

clipping level. Working in dry nitrogen prevents the grid from floating away during the clipping process.

4. Cool the loaded c-clip insertion tools, AutoGrid tweezers, grid handling tweezers, screwdriver, an AutoGrid container, and dressing forceps in the surrounding liquid nitrogen bath. Figure 1c, d show the various tools used in the clipping process.
5. Use the precooled dressing forceps to transfer the grid box containing the grids from liquid nitrogen storage. Place it into one of the grid box spaces in the AutoGrid assembly workstation and open the lid using the precooled screwdriver.
6. Place a c-clip ring into a slot in the AutoGrid assembly workstation and allow it to cool.
7. Using precooled grid handling tweezers, retrieve a grid from the grid box and gently place it into the c-clip ring.
8. Rotate the AutoGrid assembly into the clipping position using the AutoGrid tweezers.
9. Place the c-clip insertion tool into the slot above the grid and gently press the top of the tool to lock the c-clip into the c-clip ring containing the grid.
10. Rotate the AutoGrid assembly back into the loading position and transfer the clipped grid (AutoGrid) to the AutoGrid container using the AutoGrid tweezers.
11. Repeat for all grids.
12. Store AutoGrids in liquid nitrogen until loading.

3.1.6 Loading Cassette

1. Assemble the cassette loading station with the cassette gripper arm and an empty cassette. The fully assembled loading station is shown in Fig. 1e. During the initial cooling of the cassette loading station and cassette loading, it may be helpful to cool the NanoCab separately and set it aside with its cover rather than leaving it attached to the loading station. This will reduce the possibility of contamination of the nitrogen in the NanoCab.
2. Close the opening on the side of the loading station with the cover.
3. Cool the cassette loading station with liquid nitrogen and place covers over the cooled chamber to prevent contamination.
4. Once the loading station has sufficiently cooled, maintain the level of liquid nitrogen to keep the cassette submerged.
5. Transfer the AutoGrid box containing AutoGrids from liquid nitrogen storage directly into the loading station. Place it into one of the grid box holders within the loading station and remove the AutoGrid box lid using a precooled gripper tool.

6. Precool the AutoGrid tweezers in the liquid nitrogen and then use them to retrieve the first AutoGrid from the AutoGrid box. Place the AutoGrid into the cassette with the c-clip facing the bottom of the cassette (toward the position labeled “1”). Figure 1f shows the numbering system for the AutoGrids within the cassette. In this orientation, the c-clip for each AutoGrid should face left when inserted into the cassette.
7. Ensure that the AutoGrid is properly held in place by the cassette spring by gently tapping it with the AutoGrid tweezers.
8. Repeat for all AutoGrids to be loaded.
9. Remove the cover on the side of the loading station and attach the precooled NanoCab, as shown in Fig. 1e. This is done by retracting the spring at the top, holding the NanoCab in place, and releasing the spring to secure it.
10. Insert the gripper arm into the cassette by sliding it over and covering the cassette opening. Use the button on the side of the arm to grip the cassette, then quickly pick it up and transfer it to the NanoCab. Release the button and retract the gripper arm, leaving the cassette in the NanoCab.
11. Remove the NanoCab containing the loaded cassette from the cassette loading station by releasing the spring at the top of the loading station.
12. Use your fingers to check that the pin on the top of the Nanocab can be pulled up and that it springs back into place once released. This ensures that the pin has not frozen in place.
13. Replace the cover on the NanoCab and transport it to the Titan Krios.
14. Open the microscope door and check that there is no cassette currently loaded (“Cassette Loaded” light will not be illuminated).
15. Insert the NanoCab and press the “Load” button.
16. Monitor the loading progress in the Autoloader panel of the TEM User Interface.
17. Once the loading process has completed, remove the NanoCab and close the microscope door.
18. Place all tools, the AutoGrid assembly workstation, cassette loading station, and NanoCab in a safe space to warm up and dry.
19. Invert all cryo dewars and containers to allow them to warm up and dry.

3.2 Microscopy

3.2.1 Microscope

Alignment

and Performance Check

(Also See Videos

on [EM-learning.com](https://www.em-learning.com))

Subsequent protocol is written specifically for the hard- and software configuration mentioned in the Subheading 2. The here described combination of a ThermoFisher Titan Krios transmission electron microscope with a Gatan bioquantum energy filter and K2 direct electron detector are the currently most common setup for high-end data collection in the field of single particle cryo-EM and might therefore be of immediate use for many microscopists (Fig. 2). However, there are alternatives, e.g., the JEOL cryo ARM 300 electron microscope or the ThermoFisher Falcon direct electron detectors. While it is beyond the scope of this protocol to give a comprehensive overview on the options available or even describe their functionality in detail, more experienced users will be able to generalize and easily apply it to other configurations.

This step-by-step guide is aimed to help beginners in cryo-EM to get started in data collection on high-end instrumentation. It is assumed that the microscope and all components have good basic alignments and are functional and stable. Basic alignments and advanced troubleshooting are not covered by this manual. Separate notes, however, help during troubleshooting of common problems that might appear during execution of this workflow and occasionally point out alternative options.

Each individual step in the following manual is followed by a *[string of instructions]* using following frequent abbreviations:

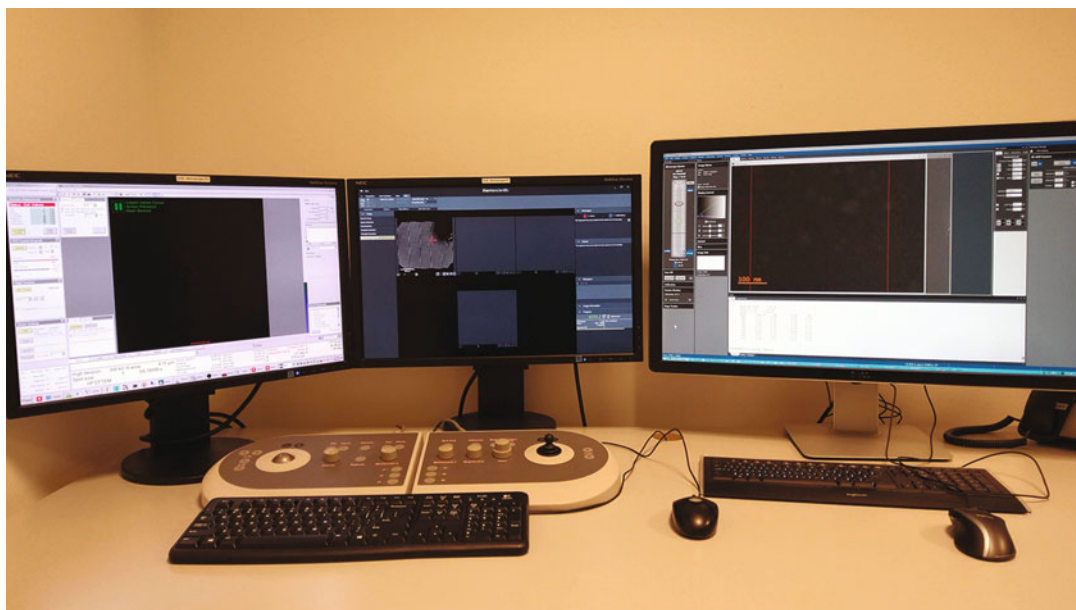


Fig. 2 Overview of the screens of the Microscope PC (M-PC) with the TEM user Interface (TUI), log window, and the embedded FluCam Viewer (left screen), EPU (middle screen), and K2-Camera control PC (K2-PC) with Digital Micrograph (DM) (right screen) as well and the Hand Panels (HP). Within TUI, individual Control Panels (CP) can be selected from the dropdown list in the lower right of the screen

M-PC: Microscope control PC; **K2-PC:** Gatan K2 Camera Control PC; **TUI:** TEM User Interface; **HP:** Hand Panels; **CP:** Control Panel (*see* **Notes 5** and **6**).

1. Make sure that the TEM server is connected [*M-PC>Microscope Software Launcher>Start Server and Applications*] and that TEM User interface (TUI), TEM imaging & Analysis (TIA), Digital Micrograph (DM), and FluCam Viewer are running, they should all be listed in the Windows Task Bar.
2. On the camera control PC, make sure DM is running, switch to Power User setup [*K2-PC>DM>Help>User>Power User*] and load relevant floating windows [*K2-PC>DM>window>-Floating Window layout>PowerUser*] to access the filter and camera controls.
3. In TUI, check microscope error logs [*M-PC>TUI>log window*] as well as status of cooling [*M-PC>TUI>Temperature Control CP>all fields should be green and both Dewars full*], vacuum [*M-PC>TUI>Vacuum (Supervisor) CP>all fields should be green*], and status of subcomponents such as energy filter [*K2-PC>DM>Auto Filter>the slit should be insertable/retractable*] and K2 camera [*K2-PC>DM>K2 Direct Detection>Health Status>wrench icon>Quick Scan, no light should be orange or red*].
4. Check that Microscope PC and K2 PC are communicating properly: the K2 needs to be listed as “EF-CCD” [*M-PC>TUI>CCD/TV Camera CP>Drop down*]. Further, make sure that the “Gatan Remote Tem” Software is running and transferring data [*M-PC>Gatan remote TEM>show log*], change magnification [*HP>Magnification*] and check if the magnification change is listed in the log. Additionally, the “Remote Digital Micrograph” Software needs to be running, make sure that camera and filter are connected. Finally, on the K2-Control PC check that information sent from the microscope is received and vice versa [*K2-PC>DM>Microscope>Setup>Test>Get Spot size*], the correct Spot size should be shown.
5. Load the most recent Alignments [*M-PC>TUI>Alignments CP>Flap out>File tab>select from list>move desired alignments from ‘Available’ to ‘Selected’>Apply*] and FEG registers for the voltage intended to use [*M-PC>FEG Registers CP>Select from list>Set*]. Make sure the Field Emission Gun (FEG) is powered and operational [*M-PC>TUI>FEG Control CP>flapout>Power; FEG*] and the high tension is at the voltage intended to be used [*M-PC>TUI>High Tension CP>drop down>High Tension*].

6. Perform a grid inventory of the Autoloader [*M-PC>TUI>Autoloader CP>flap-out>Inventory*] and check whether all grids are found at the expected position.
7. If it did not remain on stage, load cross-grating grid [*M-PC>TUI>Autoloader CP>position 1>Load*]. While loading and stabilizing, start EPU software, import previously used beam settings and set a low SA magnification beam [*M-PC>EPU>Preparations>Beam Presets>Hole/Eucentric Height>Set*], typically npEFTEM 6.500 × magnification [*HP>Magnification*], Spot size 5 [*HP>L3 and R3*], apertures C1:2000, C2:150, Obj: retracted, SA: retracted [*M-PC>TUI>Aperture CP*].
8. Open column valve [*M-PC>TUI>Setup CP>Col. Valves Closed*], switch off Autoloader turbo pump [*M-PC>TUI>Autoloader (User) CP>Flap-out>Turbo Auto off radio button*], insert fluscreen [*typically HP>R1, the FluCam will start automatically*], zoom in on the FluCam [*M-PC>FluCam Viewer>High resolution*], and activate the circular marker for the GIF entrance aperture [*M-PC>FluCam viewer>GIF*].
9. Move the stage to the center of an intact square [*HP>Joystick*] (*see Note 5* on what to do if you cannot see a beam), reset beam [*M-PC>TUI>Beam Settings CP>Reset beam*], set objective lens to eucentric focus [*HP>Eucentric focus*], reset defocus [*usually HP>L2*], normalize all [*usually HP>R2*], spread the beam to the size of the marker for the GIF entrance aperture [*HP>Intensity, clock-wise from crossover*], align it with the circle [*M-PC>TUI>Direct Alignments>Beam shift>HP>MultifunctionX and Y*], and bring the sample to eucentric height [*HP>Z axis up and down*] (e.g., using the caustic ring method, *see Note 6* on other methods for setting the eucentric height). Store the current stage position [*M-PC>TUI>Stage2 CP>Add*] and rename it “film”.
10. Verify for SA and LM modes [*HP>Magnification*] in microprobe and nanoprobe [*M-PC>TUI>Beam settings CP>Nanoprobe/Microprobe*] that there is neither a magnification nor spot size-dependent beam or image shift [*HP>Magnification; HP>L3/R3*].
11. Similarly, find a broken square on the grid [*HP>Joystick, if necessary lower the Magnification for speedup*] and store this position as “Hole”, it will be needed later to estimate dose rate, to acquire gain references, etc. Return to “film” [*M-PC>Stage2 CP>“film”>Go*].
12. Go to intermediate SA magnification [*HP>Magnification*], e.g., np EFTEM 42.000 x, if necessary recenter and resize beam to match the circle [*HP>Multifunction X and Y*];

HP>Intensity], refine eucentric height and lower stage by a few micron [*HP>Z axis down*].

13. Center condenser aperture 2 [*M-PC>TUI>Apertures CP>Adjust Condenser 2>HP>Multifunction X and Y*] and stigmatize condenser lens system [*M-PC>TUI>Stigmator>condenser>HP>Multifunction X and Y*] (on a three condenser system, first align the third condenser lense [*M-PC>TUI>Direct alignments CP>Condenser center TEM>HP>Multifunction X and Y*; the center of the focussed beam should be aligned with the center of the parallel beam, if necessary modulate the amplitude with [*HP>Focusstep*]] and then stigmatize the parallel beam as described before). The beam should now be concentric and spread uniformly when going through the cross over [*HP>Intensity*].
14. Execute following direct alignments (consult the help menu for more information on the individual tasks [*key board F1*]): nP Beam tilt pivot points in X and Y direction [*M-PC>TUI>Direct alignments CP>nP Beam tilt pp X or Y>HP>Multifunction X and Y*; image movement should be minimized, keep in mind that this is usually very stable, but focus dependent alignment], beam shift [*M-PC>Direct Alignments CP>Beam shift>HP>Multifunction X and Y*; the parallel beam should now be aligned with the marker of the GIF entrance aperture], and rotation center [*M-PC>Direct Alignments>Rotation center>HP>Multifunction A and Y*; the center of the image should show minimal movement when wobbling through focus of the objective lens, if necessary change wobbling amplitude [*HP>Focus step*]]. Other (direct) alignments, e.g., gun alignments, are typically very stable and are only performed on demand (e.g., Gun tilt and Gun shift if strong changes in beam intensity or spot size-dependent beam shift are observed).
15. Insert the 100-micron objective aperture [*M-PC>TUI>Aperture CP>Objective*, if necessary, select the correct size from the dropdown], and center it. To do so, first take action to protect the K2 camera by making sure that the fluscreen is inserted [*HP>R1, check status on the FluCam*] and the K2 camera is retracted [*K2-PC>DM>Camera>retract*], go to diffraction mode [*HP>Diffraction, red light goes on*], if necessary focus and stigmatize diffraction beam [*HP>Focus*; *M-PC>TUI>-Stigmator CP>Diffraction>HP>Multifunction X and Y*], insert the beam stop [*M-PC>TUI>FluCam viewer>insert Beamstop icon*], shift the central beam behind the beam stop [*HP>Multifunction X and Y*], adapt the FluCam histogram to make the diffraction pattern visible [*M-PC>FluCam viewer>histogram*; alternatively, adapt the histogram by left clicking into the image and turning the middle mouse wheel], and

- align the aperture with the diffraction pattern [*M-PC>Aperture CP>Adjust Objective>HP>Multifunction X and Y*].
16. Confirm that the beam is parallel. This is the case if simultaneously the diffraction pattern and the aperture edge are focused and sharp, respectively [*HP>Focus; HP>Intensity*].
 17. Leave diffraction mode [*HP>Diffraction, red light disappears*], retract the beam stop [*M-PC>TUI>FluCam viewer>retract Beamstop icon*], go back to Eucentric height [*HP>Z axis up*], and tune objective lens system. To do so, go to high SA magnification, e.g., 250.000 x [*HP>Magnification*] (if necessary, recenter and resize the beam, the screen current should now be ~0.5 nA), defocus by a few hundred nm [*HP>Focus, turn counter-clock wise from cross over*], insert the K2 camera [*K2-PC>DM>Camera>insert*], and confirm that there is an unobstructed image [*K2-PC>DM>Camera view>Start view*].
 18. Defocus the image by a few hundred nm to underfocus [*HP>Focus, counter-clockwise from focus; at underfocus, the images shows increased contrast. Contrary, at overfocus, low frequency contrast gets inverted while in focus, the image almost disappears because no contrast is transferred*].
 19. Stigmatize objective lens system and perform coma free alignment. Because it might obstruct the tilted beam, the objective aperture needs to be retracted first [*M-PC>Apertures CP>Obj. retracted*]. Lens alignment can be done manually or automatically:
 20. Option 1 (manually): put camera on binned live view with high-frame rate [*K-2PC>DM>Camera View>Select search>Start view*], calculate live FFT [*K2-PC>DM>Process>live>FFT*], defocus in order clearly resolve 3–5 Thon rings, stigmata lens [*M-PC>Stigmator CP>Objective>multifunction X and Y, rings should be concentric*]. Perform coma-free alignment [*M-PC>Direct alignments CP>Coma free alignment X and Y>Multifunction X and Y; for all beam tilts, all images should show the same defocus. Keep in mind that for higher beam tilts the FFT of the will become elliptical due to spherical aberration of the objective lens*].
 21. Option 2 (automatically), using AutoCTF software. First, calibrate the system [*M-PC>AutoCTF>Calibration tab>Get from system*] and define parameter limits [*M-PC>AutoCTF>Settings Tab; limits for astigmatism should be 5 nm and for coma 500 nm*]. Then measure defocus and astigmatism of the untilted beam [*M-PC>AutoCTF>AutoCTF tab>tick counting, select 3 s exposure time, no binning, full readout>measure*], if necessary, adjust defocus/height to get

reliable fitting of Thon rings. Next, stigmata image [*M-PC>AutoCTF>AutoCTF tab>Auto-stigmata*] iteratively until reaching 5 nm. Create Zemlin tableau [*M-PC>AutoCTF>AutoCTF tab>Measure coma*], and correct coma iteratively until reaching 500 nm threshold [*M-PC>AutoCTF>AutoCTF tab>Auto-coma*].

22. Perform the Young fringe resolution test: In this test, two shifted images at high magnification and high dose are taken and compared. This allows to estimate the optical resolution of the system at conditions that are not limited by the electron dose. For this, change C1 lens to a low spot size, e.g., 2 [*HP>L3*], go to the highest available SA magnification, e.g., $250.000 \times$ [*HP>Magnification*], center [*HP>track ball*] and condense [*HP>Intensity*] the beam to get a high screen current, e.g., 2 nA, and take an un-binned, full area, 3 s exposure on the K2 in linear mode [*M-PC>CCD/TV Camera CP>Acquire*]. Now, shift the image by less than half of the field of view by either applying stage [*HP>Joystick*] or image shift [*M-PC>select Image shift as Multifunction X and Y>HP>Multifunction X and Y*], wait a few seconds to stabilize and take another image with similar settings. Within TIA, select and copy the second image, click on the first image, and paste it to the right, calculate the sum of both images [*M-PC>TIA>Primary Processing>+ icon>Data1=image1, Data2=image2*], select the summed image, and calculate its Fourier transform [*M-PC>TIA>FFT/IFFT>FFT*], and remove the red square from the summed image to use the FFT of the full image [*mouse>left click on square>Keyboard Delete*]. Finally, in the Fourier transform, read out the highest frequency at which interference lines can be seen in shift direction [*mouse>left click on frequency>M-PC>TIA>Data Info>R*]. This frequency represents the resolution according to the young fringe test. The test may also be performed in perpendicular direction in order to detect anisotropy. The system specifications guarantee an isotropic resolution of at least 1.4 Å (corresponding to the third diffraction ring of the gold cross-grating sample). In practice, for a well aligned and stable system, typically a resolution well beyond 1 Å can be achieved.
23. Acquire a reference image at experimental conditions: This image allows an estimate to be made of the practical information limit for an almost ideal sample. Insert and center objective aperture, stigmatize objective lens, and move stage to “hole” [*see previous instructions on these tasks*]. The magnification should be chosen to obtain a pixel size of about 1/3–1/4 of the desired resolution, typically, $130.000 \times$ [*HP>Magnification*], resulting in a pixel size of 0.88 Å and a pixel area of

0.77 Å². Next, the illumination settings are adjusted so that the beam size is at least $1.5 \times$ and not more than $2 \times$ the diameter of the field of view of the detector which corresponds to the size of the marker for the GIF entrance aperture on the FluCam viewer [*HP>Intensity*] and check that at these settings the beam remains parallel [*M-PC>TUI>Beam Settings CP>Illumination “parallel”*]. Increase the spot size until the dose rate on the detector reaches values between 5 and 10 e⁻/pix s in counted mode [*HP>R3>K2-PC>DM>Camera view>Start view>read out dose rate, too low dose rates result in unnecessary long exposure times and higher dose rates in coincidence loss in electron counting; also account for the loss of intensity by electron scattering of thicker samples.*]. Store these beam settings as “Acquisition” preset in EPU [*M-PC>EPU>Preparation>Acquisition and Optics Settings>Presets>Data Acquisition>get; select ‘Counted’ mode*]. Calculate the exposure time needed to reach a wanted total electron dose on the sample, here 50 e⁻/Å². For a measured dose rate of 8 e⁻/pix s, this would be (50/8/0.77) s = 8.11 s. For dose fraction fractions, aim for a signal of ~2 e⁻/pix. In our example, this would result in ~32 fractions. Store the calculated exposure time and number of fractions in EPU [*M-PC>EPU>Preparation>Acquisition and Optics Settings>Data Acquisition>fill in time and number of fractions*]. Move stage back to “film” [*M-PC>TUI>Stage2 CP>“film”>Go*] and acquire an image with given preset [*M-PC>EPU>Preparation>Acquisition and Optics Settings>Data Acquisition>Preview*]. Calculate Fourier transform and estimate to which frequency Thon rings can be seen [*M-PC>EPU>FFT button*]. This frequency corresponds to the theoretical image information limit using the chosen acquisition parameters. Keep in mind that the parameters calculated above are highly depend in the type of detector used and their mode of operation. For an in-depth comparison of other set-ups, particularly using the Falcon 3 direct detector, see [34].

24. Tune the Bioquantum energy filter. To do so, first move stage to “hole” [*M-PC>Stage2 CP>“hole”>Go*], load the acquisition preset [*M-PC>EPU>Preparation>Acquisition and Optics Settings>Data Acquisition>Set*], insert [*K-2PC>DM>tick the “slit” tick box*], set [*K-2PC>DM>set slit size to 20 eV*] and pre-center the slit [*K2-PC>DM>Align ZLP*], lower spot size by three steps [*HP>L3*] and run a full filter tuning [*K2-PC>DM>tune button>full tune*].
25. Acquire gain and dark references of the K2 camera. [*K2-PC>DM>Camera>Prepare gain reference; follow on screen instructions for acquisition of linear and counted mode gain and dark references*].

3.2.2 Grid Screening
and Setup of Automated
SPA Data collection
Using EPU

1. Sequentially load and screen the grids that have been loaded to the autoloader. Many criteria will impact the decision on which grid to use for data collection which is beyond the scope of this manual. See, e.g., [35] on how to judge the quality of and optimize a sample for cryo single particle data collection. Here, we assume that the samples have been optimized and pre-screened previously.
2. Load favorite grid [*M-PC>TUI>Autoloader CP>position x>Load*], open column valve [*M-PC>TUI>Setup CP>Col. Valves Closed*], and make sure the AL turbo pump is switched off [*M-PC>TUI>Autoloader (User) CP>Flap-out>Turbo Auto off radio button*].
3. Load Atlas preset in EPU [*M-PC>EPU>Preparation>Acquisition and Optics Settings>Atlas>Set*], insert fluscreen [*HP>RI*], bring grid roughly to eucentric height [*see Note 6*], retract the objective aperture [*M-PC>TUI>Aperture CP>Objective>none*], and collect a full atlas [*M-PC>EPU>Atlas>Session Setup>Create new sample>type name>Acquire*].
4. Link a new EPU session to the collected atlas [*M-PC>EPU>EPU>Session Setup>New Session>type name; choose MRC file format, manual session and regular hole size*] and store coordinates [*M-PC>EPU>EPU>Square Selection>Unselect all>right click over desired coordinate>move stage to coordinate; M-PC>TUI>Stage2 CP>Add*] for a “hole” position (typically a broken square), a flat and thin “sample” position (representative hole with presumably thin ice), and a “feature” position (typically ice contamination or a crack).
5. Move stage to “hole” [*M-PC>Stage2 CP>“hole”>Go*], set acquisition beam [*M-PC>EPU>Preparation>Acquisition and Optics Settings>Data Acquisition>Set*], and take a preview image to confirm beam centering, dose rate, and gain reference quality [*M-PC>EPU>Preparation>Acquisition and Optics Settings>Data Acquisition>Preview*].
6. Move to “sample” [], refine eucentric height [*see Note 6*], slightly defocus image to a value close to imaging condition (here 1–2 μm) [*HP>Focus*], and take another preview image. Check again dose rate (it should be at least 80% of the input flux, otherwise the sample is too thick), if necessary adjust beam shift [*M-PC>Direct Alignments CP>Beam shift>HP>Multifunction X and Y*] and beam tilt pivot points for this defocus [*M-PC>TUI>Direct alignments CP>nP Beam tilt pp X or Y>HP>Multifunction X and Y*], and calculate the FFT of the preview image to check for amounts of signal, defocus, astigmatism, drift, vibration, magnetic fields, etc.

7. Go through all beam presets in EPU [*M-PC>EPU>Preparation>Acquisition and Optics Settings>...>Set*] and make sure that all of them produce preview images with sufficient signal with full illumination of the detector [*M-PC>EPU>Preparation>Acquisition and Optics Settings>...>Preview*]. Test the auto functions for Eucentric height and Focusing [*M-PC>EPU>Auto Functions>Auto Eucentric by beam tilt / Autofocus>Start*].
8. Move to “feature” position [*M-PC>Stage2 CP>“hole”>Go*] and calibrate image shift [*M-PC>EPU>Preparation>Calibrate Image Shifts>Start Calibration; follow the on screen instructions*].
9. Insert [*M-PC>TUI>Aperture-CP>Objective*] and if necessary, recenter [*M-PC>TUI>Apertures CP>Objective>Adjust; HP>Multifunction X and Y*] the objective aperture.
10. Add good squares to the EPU session [*M-PC>EPU>EPU>Square Selection>right click over square>add>right click>move stage to grid square*]. For each square set, the eucentric height [*using the auto function*] and take square images for hole selection [*M-PC>EPU>EPU>Hole Selection>Acquire*].
11. For each square, select holes with thin ice. Make use of the histogram and other tools provided to reliably select holes with constant ice thickness [*M-PC>EPU>EPU>Hole Selection>Acquire>measure hole size>place the yellow circles accurately over two neighboring holes>find holes; adjust the histogram sliders to select holes of ideal thickness*]. Holes may also be added or removed manually [*Ctrl+left click; Shift+left click*].
12. Go to a representative hole on the first selected square [*right click on hole>move stage to location*] and design a template for automated data collection [*M-PC>EPU>EPU>Template Definition>Acquire>Find and Center Hole*]. Place four green Acquisition positions [*Add acquisition area>left click on target area*] in a 2 μm hole making sure that the detector area (rectangle) will be taken fully in the hole while the exposure beam (round) should additionally touch the carbon [*left click and drag acquisition area to move o another position*]. Assign one defocus value to each exposure area, ranging between $-0.7 \mu\text{m}$ and $-2 \mu\text{m}$ defocus [*left click on each green field>type defocus value*]. Overlapping exposure beams are to be avoided to prevent double exposures. Add a blue focus area to the carbon next to the hole [*Add autofocus area>left click*] and select to focus every 10 μm using the objective lens without auto stigmation.

13. Execute the template on this and a couple of other holes to compare ice thickness, particle distribution, and reproducibility of auto functions [*Template Execution*].
14. Add enough squares and holes for an overnight data collection. For the settings discussed here (estimated speed of ~50 movies per hour) ~1.000 exposure areas (and thus ~250 holes) will be sufficient.
15. Perform final checks [*Is the AL turbo pump switched off? Is the Krios enclosure closed? Does the nitrogen tank still hold sufficient liquid nitrogen (>50 l) and is it fully pressurized (1.5 bar over pressure)? If applicable, is the magnetic field cancellation system active? Is the objective aperture inserted and centered? Is the exposure beam still centered? Did the slit of the energy filter drift since tuning? Is sufficient local storage space available or are the scripts for automated movie transfer to the file server as well as for on-the-fly processing, etc. running?*].
16. Start automated data collection [*M-PC>EPU>EPU>Automated Acquisition>Start*]. Monitor the EPU session and live pre-processing for at least 30 min to make sure that parameters stay constant and to get initial statistics which will help to fine tune important parameters, such as defocus range, exposure positions, beam shift, or objective stigmators.

3.2.3 Data Processing

The following paragraph is a general description on how to approach image processing in cryo-EM. The different standard steps of a typical processing workflow are presented in Fig. 3.

Motion Correction

Since the year 2011, a new generation of complementary metal-oxide-semiconductor (CMOS) cameras has made it possible to directly detect incoming electrons without the need for a scintillator, which was the case for charge-coupled device (CCD) cameras. These CMOS cameras also referred as “direct detectors” have a much-improved Detective Quantum Efficiency (DQE) resulting in

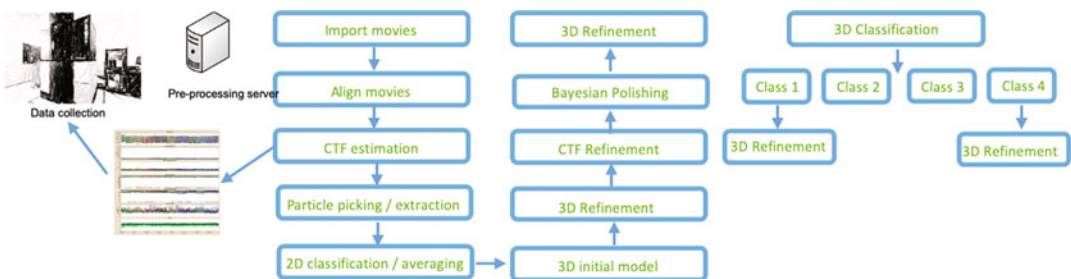


Fig. 3 Image processing workflow. A typical workflow for SPA image processing comprising a “pre-processing” loop providing live feedback to the microscope operator on the quality of the data being collected

higher contrast images as well as a high-frame rate, making it possible to record fast movies [36].

When the electron beam hits the sample embedded in vitreous ice, it creates a radiation-induced motion and blurriness of the images resulting in loss of information. The recording of movies makes it possible to correct (to some extent) for this beam-induced movement at the micrograph or particle level [37]. Typical programs to correct for beam-induced movement are MotionCorr [38, 39] and Unblur [40]. Both programs will also carryout a dose-weighting approach in order to take into account the radiation damage generated while collecting the movie. The output of the movie correction programs will be a sum of the aligned stack of frames (aligned average) and the aligned stack itself.

CTF Estimation

The aligned average (we will use the term micrograph in the following text) can then be used to estimate the Contrast Transfer Function (CTF) of the microscope that is affecting the signal to noise ratio (SNR) of the Fourier components of each micrograph [41]. Typical programs used at this step are CTFFIND [42] and gCTF [43]. These programs fit CTFs to the Thon rings visible in the power spectra of (patches of) micrographs allowing estimation of defocus and astigmatism angle as well as trying to give some information on the quality of the fit and the maximum resolution of the fit (thus providing a first estimation on the quality of the collected data).

Live Processing

The above two steps, movie alignment and CTF estimation can be fully automated. This has resulted in the development of live processing programs allowing a real-time feedback to the user of the microscope in order to optimize the data collection time.

Homemade scripts or open source software can perform such tasks [44–47].

At the NeCEN facility, we have developed a homemade system tailored to the microscopist's needs.

A background script continuously monitors data collection at the microscope and when a new project is started, the script picks it up and creates a pre-processing pipeline associated to the data collection happening at the microscope.

The collected data is moved from the microscope PC to a longer term storage. At the same time, the data is analyzed and results are provided to the microscope operator in a private website format that is accessible from the NeCEN private network (Fig. 4).

From the above two steps, useful information can be gathered such as plotting the motion at the micrograph level, defocus values, estimated resolution from the micrograph, phase shift when phase plate is used. In particular, the information limits and astigmatism values reported by the CTF programs are valuable numbers. If the reported values are inconsistent or get degraded, it is a good

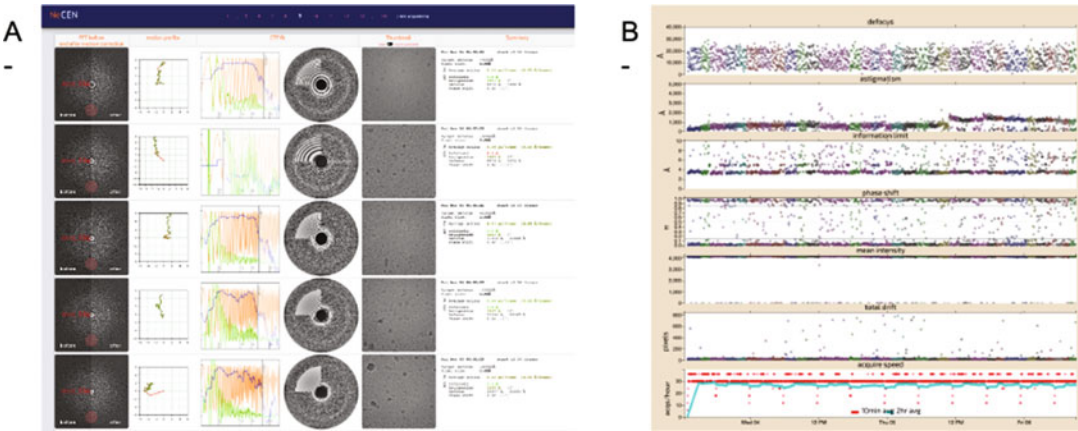


Fig. 4 On-the-fly pre-processing at NeCEN. **(a)** An example of the pre-processing output from NeCEN's internal website. It does movie alignment and CTF estimation of generated aligned micrograph and provides output values for each micrograph (including, astigmatism, CTF fit, information limit, defocus value). **(b)** A different website show time series plot of defocus, astigmatism, phase shift (for vpp data collection), mean intensity, total drift, and speed of data collection

indication that the user should investigate if the microscope behaves as it should or if the grid area chosen for data collection is suitable to collect high-resolution images.

Particle Picking

The particle-picking step is self-describing; the scientist has to select features in the micrographs that look like the object of interest (“particles”). Different software exists offering different approaches to the problem of picking particles. It is a complex step to be performed fully automatically and different programs offer different approaches to the problem.

Most of the software available for particle picking have a manual mode where the user manually selects the particles. This approach works well as the scientist knows best the sample, but it is very time consuming and repetitive. Manually picking a full micrograph can take 2–3 min or more. When a typical dataset is 2000 micrographs, a minimum of 100 h of manual picking is needed in the best cases. It is not a very realistic approach when you have to quickly process datasets on a regular basis.

Some software like XMIPP and EMAN2 [48, 49] offer a supervised approach using machine learning approaches. While the scientist picks manually a few micrographs, the program is self-training itself and offers to automatically pick the remaining of the micrographs.

Others like RELION offer template-matching approaches where different templates of the object of interest are used to automatically pick micrographs [50]. Templates can be generated from a small subset of particles picked manually and then averaged or from projections of a known 3D model.

Recently developed programs like WARP [47] and crYOLO [51] use neural networks and a database of trained models to automatically pick particles without any human intervention.

Datasets obtained from (semi-) automated particle selection procedures often contain more false positives than those selected manually. These comprise aggregated particles, contamination in the ice containing the sample, background features, etc.

Once particles have been selected from all the micrographs, they are extracted (boxed) from the images. The user needs to choose the size of the box. It should be larger than the particle to be extracted; it is common practice to use a box size 1.5 to two times larger than the longest axis of the particle to be extracted.

2D Classification

The 2D classification step can help cleaning the data by removing noisy particles, suboptimal particles, and junk in general. Usually, one would do 2–3 rounds of 2D classification cleansing before moving into *ab initio* 3D model generation.

2D classification procedures like the one adopted in RELION are based on the maximum-likelihood method [52–54].

Reference-free class averages are obtained in a completely unsupervised manner by starting multi-reference alignments from average images of random subsets of the unaligned data.

Mostly, the only parameter that is needed to input by the user is the number of classes. Each class should in the end contain a minimum of 300 particles to make sure the signal to noise ratio is high enough to obtain good particle alignments. A typical dataset of 100,000 extracted particles could be divided into 300 classes. In practice, the higher the number of classes the longer the computation will take. The user has to perform some trial and error runs to obtain optimal results.

Initial Model Generation

After 2D classification, a 3D model needs to be generated. Different approaches exist; EMAN2 [49] uses 2D class averages to find the orientation in 3D while software like cryoSPARC [55] and RELION [56] use a stochastic gradient approach procedure based on the raw particles. It is thus even more important to have a “clean” dataset containing good particles. In case a 3D structure of a similar molecule already exists, it can be faster to use this information to generate a low-resolution initial model to refine with the new dataset. Another approach is to generate experiment maps obtained by subtomogram averaging or random conical tilt [57, 58].

3D Classification

Once an initial model has been generated, it is necessary to perform 3D classification.

No sample is completely homogeneous and heterogeneity is always present and will reduce the quality and resolution of the final structure.

Different software offers the option to perform unsupervised 3D classification where a low-resolution initial model is used as a reference for generating a small number of 3D models. In the initial round of refinement, each particle is randomly assigned to a 3D class.

The number of 3D classes chosen by the user depends on the size of the dataset but most importantly by the available computing power. Typically, less than 10 classes are used to reduce computing time while still being able to differentiate good and bad structures (either in terms of resolution or in terms of structural arrangement). To assess the consistency of the classification, one usually runs several 3D classifications using different number of classes [54]. More advanced 3D classification runs can also be performed with more exhaustive angular searches, as well as with the use of specific soft-edge mask to identify small structural differences among the classes [54].

3D Refinement

After 3D classification, each interesting class can be individually refined to high resolution by a 3D refinement procedure. The particles associated to each 3D class (or a pool or merged similar classes) are refined against the 3D classes volume. Alignment of the particles is initially determined by the 3D class map and then refined iteratively until convergence and until the orientation and translational alignments are stable and very precise.

Post-Processing and Other Corrections

To bring out high-resolution features and enhanced map interpretability, the user should apply a B-factor correction, which uses a filter to boost high frequencies while applying dampening of noise. Further recent developments in software have focused on eliminating fine errors in data collection that have a significant impact when moving toward atomic resolution. These developments include estimations of defocus for each particle in a micrograph [56, 59], performing Ewald sphere correction [56, 60], beam tilt correction [56, 61], optical aberrations, and magnification anisotropy [62].

4 Notes

1. There are many different types of grids and support films available for use in cryo-EM. These can vary in the type of metal mesh, mesh size, support film material, support film hole pattern, and hole size. The choice in grids can be optimized for a specific experiment [32]. Here, we have selected 200 mesh grids because these provide relatively large squares. This makes the setup of data acquisition faster since fewer squares need to be selected although the support film is slightly more prone to damage. We are using R 2/2 carbon support films, as 2 μm holes are large enough to acquire several images per hole,

which increases the speed of data collection, without being so large as to have significant ice gradients.

2. It may be necessary to optimize a protocol depending on the instrument and experiment. If grids still appear to be hydrophobic after glow discharging, it may be necessary to increase the glow discharge current or time. If grids appear to be too hydrophilic or the support film is damaged by the glow discharging process, the discharge current or time could be decreased.

The volume of sample applied to the grid can also be optimized depending on the experiment.

3. We use 95% humidity for our instrument because higher levels lead to condensation in the chamber that may affect the humidity sensor readout.
4. Appropriate values for these parameters often vary between instruments and samples, but once optimized, tend to be consistent for a particular Vitrobot.
5. What to do if you cannot find the beam?
 - I. Make sure that there is high tension and FEG emission and that there are no critical error messages indicating malfunctioning of FEG, lenses, of vacuum. Also check if all electromagnetic compounds are responding to manual changes [M-PC>TUI>System status CP].
 - II. Insert the fluscreen [HP>R1] and make sure that column valves are open and no beam blander or shutter is blocking the beam (e.g., caused by the falcon protector) [M-PC>Microscope Software Launcher>Tools>Blender Shutter Monitor]. If necessary unblank beam [M-PC>TUI>CCD/TV Camera CP>Blank] or follow instructions given by the falcon protector (e.g., do calibrations, increase illumination area, or higher spot size to reach <100% intensity).
 - III. Go to parallel illumination range [HP>Intensity], eucentric focus [HP>Eucentric focus], and reset beam [M-PC>TUI>Beam Settings CP>Reset beam].
 - IV. Go to lowest SA magnification [HP>Magnification] and move stage [HP>Joy stick] to check whether a thick sample or grid bar is blocking the beam. If necessary, go to low LM magnification or even retract the sample [M-PC>TUI>Autoloader-CP>unload].
 - V. Retract the objective aperture as it might be misaligned and block the beam [M-PC>TUI>Aperture-CP>Objective].
 - VI. If the beam is visible on the FluCam but not on the K2, make sure that the beam is aligned with the entrance aperture of the energy filter [HP>Track ball], then retract the fluscreen [HP>R1], insert the K2

[K2-PC>DM>Camera>insert], go on live view [K-2PC>DM>Camera View>Select search>Start view], and retract the slit of the energy filter [K-2PC>DM>untick the slit tick box]. If there is still no beam, contact your system administrator as the system might be misaligned or malfunctioning.

6. Ways to find Eucentric height of the specimen manually (during data acquisition, this task be performed by auto functions).
 - I. Caustic ring method (fast, requires diffracting sample): Insert fluscreen [HP>R1], set eucentric focus [HP>Eucentric focus], and focus the beam [intensity]. If the sample is out of eucentric height, a diffraction pattern will be visible. Change the z-height [HP>Z axis up/down] until the pattern condenses to one central spot.
 - II. Stage wobbler method (slow, more accurate): Insert fluscreen [HP>R1] and move the stage to center a recognizable image feature [HP>Joy stick], activate alpha wobbler [M-PC>TUI>Stage2 CP>Flap-out>Wobbler], and change z-height [HP>Z axis up/down] to minimize sample movement.
 - III. Alternatively, do not wobble the stage but manually rotate alpha by few degrees [HP>alpha tilt +/-] and recenter the feature by adapting the z-height [HP>Z axis up/down]. To refine, repeat in opposite alpha direction and with higher amplitude.
 - IV. By analyzing image defocus (most accurate, often used for fine tuning at higher magnifications): Set Eucentric focus [HP>Eucentric focus], insert K2 camera [K2-PC>DM>Camera>insert], switch to binned continuous read out [K-2PC>DM>Camera View>Select search>-Start view], and calculate a live FFT [K2-PC>DM>Process>live>FFT]. Adapt z-height [HP>Z axis up/down] to increase the size of the Thon rings in the FFT until the rings disappear and the contrast in the image is minimized.

References

1. Bai XC, McMullan G, Scheres SH (2015) How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* 40:49–57
2. Fernandez-Leiro R, Scheres SH (2016) Unravelling biological macromolecules with cryo-electron microscopy. *Nature* 537:339–346
3. Dubochet J, McDowell AW (1981) Vitrification of pure water for electron microscopy. *J Microsc Oxford* 124:Rp3–Rp4
4. Adrian M, Dubochet J, Lepault J, McDowell AW (1984) Cryo-electron microscopy of viruses. *Nature* 308:32–36
5. Dubochet J, Adrian M, Chang JJ, Homo JC et al (1988) Cryo-electron microscopy of vitrified specimens. *Q Rev Biophys* 21:129–228
6. Dobro MJ, Melanson LA, Jensen GJ, McDowell AW (2010) Plunge freezing for electron cryomicroscopy. In: Jensen GJ (ed) *Methods in*

- enzymology, vol 481. Cryo-EM, Part A—Sample Preparation and Data Collection, pp 63–82
7. Russo CJ, Passmore LA (2014) Electron microscopy: ultrastable gold substrates for electron cryomicroscopy. *Science* 346:1377–1380
 8. Pantelic RS, Meyer JC, Kaiser U et al (2010) Graphene oxide: a substrate for optimizing preparations of frozen-hydrated samples. *J Struct Biol* 170:152–156
 9. Russo CJ, Passmore LA (2014) Controlling protein adsorption on graphene for cryo-EM using low-energy hydrogen plasmas. *Nat Methods* 11:649–652
 10. Palovcak E, Wang F, Zheng SQ et al (2018) A simple and robust procedure for preparing graphene-oxide cryo-EM grids. *J Struct Biol* 2018(204):80–84
 11. Cheung M, Adaniya H, Cassidy C et al (2018) Improved sample dispersion in cryo-EM using “perpetually-hydrated” graphene oxide flakes. *J Struct Biol* 204:75–79
 12. Naydenova K, Peet MJ, Russo CJ (2019) Multifunctional graphene supports for electron cryomicroscopy. *Proc Natl Acad Sci USA* 116:11718–11724
 13. Kelly DF, Abeyrathne PD, Dukovski D, Walz T (2008) The affinity grid: a pre-fabricated EM grid for monolayer purification. *J Mol Biol* 382:423–433
 14. Kelly DF, Dukovski D, Walz T (2010) A practical guide to the use of monolayer purification and affinity grids. *Methods Enzymol* 481:83–107
 15. Crucifix C, Uhring M, Schultz P (2004) Immobilization of biotinylated DNA on 2-D streptavidin crystals. *J Struct Biol* 146:441–451
 16. Wang L, Ounjai P, Sigworth FJ (2008) Streptavidin crystals as nanostructured supports and image-calibration references for cryo-EM data collection. *J Struct Biol* 164:190–198
 17. Wang L, Sigworth FJ (2010) Liposomes on a streptavidin crystal: a system to study membrane proteins by cryo-EM. *Methods Enzymol* 481:147–164
 18. Han BG, Walton RW, Song A et al (2012) Electron microscopy of biotinylated protein complexes bound to streptavidin monolayer crystals. *J Struct Biol* 180:249–253
 19. Han BG, Watson Z, Kang H et al (2016) Long shelf-life streptavidin support-films suitable for electron microscopy of biological macromolecules. *J Struct Biol* 195:238–244
 20. Han BG, Watson Z, Cate JHD, Glaeser RM (2017) Monolayer-crystal streptavidin support films provide an internal standard of cryo-EM image quality. *J Struct Biol* 200:307–313
 21. Yu G, Vago F, Zhang D et al (2014) Single-step antibody-based affinity cryo-electron microscopy for imaging and structural analysis of macromolecular assemblies. *J Struct Biol* 187:1–9
 22. Yu G, Li K, Jiang W (2016) Antibody-based affinity cryo-EM grid. *Methods* 100:16–24
 23. Yu G, Li K, Huang P et al (2016) Antibody-based affinity cryo-electron microscopy at 2.6 Å resolution. *Structure* 24:1984–1990
 24. Glaeser RM (2018) Proteins, interfaces, and cryo-EM grids. *Curr Opin Colloid Interface Sci* 34:1–8
 25. Lu Z, Shaikh TR, Barnard D et al (2009) Monolithic microfluidic mixing-spraying devices for time-resolved cryo-electron microscopy. *J Struct Biol* 168:388–395
 26. Jain T, Sheehan P, Crum J et al (2012) Spot-iton: a prototype for an integrated inkjet dispense and vitrification system for cryo-TEM. *J Struct Biol* 179:68–75
 27. Razinkov I, Dandey V, Wei H et al (2016) A new method for vitrifying samples for cryoEM. *J Struct Biol* 195(2):190–198
 28. Arnold SA, Albiez S, Bieri A et al (2017) Blotting-free and lossless cryo-electron microscopy grid preparation from nanoliter-sized protein samples and single-cell extracts. *J Struct Biol* 197:220–226
 29. Dandey VP, Wei H, Zhang Z et al (2018) Spot-iton: new features and applications. *J Struct Biol* 202:161–169
 30. Ravelli RBG, Nijpels FJT, Henderikx RJM et al (2019) Automated cryo-EM sample preparation by pin-printing and jet vitrification. *BioRxiv*:651208. <https://doi.org/10.1101/651208>
 31. Rubinstein JL, Guo H, Ripstein ZA et al (2019) Shake-it-off: a simple ultrasonic cryo-EM specimen-preparation device. *Acta Crystallogr D Struct Biol* 75:1063–1070
 32. Drulyte I, Johnson RM, Hesketh EL et al (2018) Approaches to altering particle distributions in cryo-electron microscopy sample preparation. *Acta Crystallogr D Struct Biol* 74:560–571
 33. Frederik P, Bomans P, Franssen V, Laeven P (2000) A vitrification robot for time resolved cryo-electron microscopy. In: Cech S, Janisch R (eds) *Proceedings of the 12th European Congress on Electron Microscopy*, vol I. Reklamní Atelier Kupa, Brno, pp B383–B384
 34. Thompson RF, Iadanza MG, Hesketh EL, Ranson NA (2019) Collection, pre-processing and on-the-fly analysis of data for high-

- resolution, single-particle cryo-electron microscopy. *Nat Protoc* 14:100–118
35. Noble AJ, Dandey VP, Wei H et al (2018) Routine single particle CryoEM sample and grid characterization by tomography. *elife* 7: e34257
 36. McMullan G, Chen S, Henderson R, Faruqi AR (2009) Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy* 109:1126–1143
 37. Brilot AF, Chen JZ, Cheng A et al (2012) Beam-induced motion of vitrified specimen on holey carbon film. *J Struct Biol* 177:630–637
 38. Li X, Mooney P, Zheng S, Booth CR et al (2013) Electron counting and beam-induced motion correction enables near atomic resolution single particle cryoEM. *Nat Methods* 10:584–590
 39. Zheng SQ, Palovcak E, Armache JP et al (2017) MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* 14:331–332
 40. Campbell MG, Cheng A, Brilot AF et al (2012) Movies of ice-embedded particles enhance resolution in electron cryo microscopy. *Structure* 20:1823–1828
 41. Wade RH (1992) A brief look at imaging and contrast transfer. *Ultramicroscopy* 46:145–156
 42. Rohou A, Grigorieff N (2015) CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J Struct Biol* 192:216–221
 43. Zhang K (2016) Gctf: real-time CTF determination and correction. *J Struct Biol* 193:1–12
 44. Alewijnse B, Ashton AW, Chambers MG et al (2017) Best practices for managing large CryoEM facilities. *J Struct Biol* 199:225–236
 45. Biyani N, Righetto RD, McLeod R et al (2017) Focus: the interface between data collection and data processing in cryo-EM. *J Struct Biol* 198:124–133
 46. Gómez-Blanco J, de la Rosa-Trevín JM, Marabini R et al (2018) Using Scipion for stream image processing at Cryo-EM facilities. *J Struct Biol* 204:457–463
 47. Tegunov D, Cramer P (2019) Real-time cryo-electron microscopy data preprocessing with Warp. *Nat Methods* 16:1146–1152
 48. Sorzano COS, Marabini R, Velazquez-Muriel J et al (2004) XMIPP: a new generation of an open-source image processing package for electron microscopy. *J Struct Biol* 148:194–204
 49. Tang G, Peng L, Baldwin PR, Mann DS et al (2007) EMAN2: an extensible image processing suite for electron microscopy. *J Struct Biol* 157:38–46
 50. Scheres SH (2015) Semi-automated selection of cryo-EM particles in RELION-1.3. *J Struct Biol* 189:114–122
 51. Wagner T, Merino F, Stabrin M et al (2019) SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. 2, 218 doi: <https://doi.org/10.1038/s42003-019-0437-z>
 52. Sigworth FJ (1998) A maximum-likelihood approach to single-particle image refinement. *J Struct Biol* 122:328–339
 53. Sigworth FJ, Doerschuk PC, Carazo JM, Scheres SHW (2010) Chapter ten—an introduction to maximum-likelihood methods in Cryo-EM. *Methods Enzymol* 482:263–294
 54. Scheres SH (2012) RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* 180:519–530
 55. Punjani A, Rubinstein JL, Fleet D, Brubaker MA (2017) cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* 14:290–296
 56. Zivanov J, Nakane T, Forsberg BO et al (2018) New tools for automated high-resolution cryo-EM structure determination in RELION-3. *elife* 7:e42166
 57. Bharat TA, Scheres SH (2016) Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION. *Nat Protoc* 11:2054–2065
 58. Mastronarde DN, Held SR (2017) Automated tilt series alignment and tomographic reconstruction in IMOD. *J Struct Biol* 197:102–113
 59. Grant T, Rohou A, Grigorieff N (2018) cis-TEM, user-friendly software for single-particle image processing. *elife* 7:e35383. 37
 60. Russo CJ, Henderson R (2018) Ewald sphere correction using a single side-band image processing algorithm. *Ultramicroscopy* 187:26–33
 61. Glaeser RM, Typke D, Tiemeijer PC et al (2011) Precise beam-tilt alignment and collimation are required to minimize the phase error associated with coma in high-resolution cryo-EM. *J Struct Biol* 174:1–10
 62. Zivanov J, Nakane T, Scheres SH (2020) Estimation of high-order aberrations and anisotropic magnification from cryo-EM datasets in RELION-3.1. *IUCrJ* 7:253–267



Image Processing in Cryo-Electron Microscopy of Single Particles: The Power of Combining Methods

Carlos Oscar S. Sorzano, Amaya Jiménez-Moreno, David Maluenda, Erney Ramírez-Aportela, Marta Martínez, Ana Cuervo, Robert Melero, Jose Javier Conesa, Ruben Sánchez-García, David Strelak, Jiri Filipovic, Estrella Fernández-Giménez, Federico de Isidro-Gómez, David Herreros, Pablo Conesa, Laura del Caño, Yunior Fonseca, Jorge Jiménez de la Morena, Jose Ramon Macías, Patricia Losana, Roberto Marabini, and Jose-Maria Carazo

Abstract

Cryo-electron microscopy has established as a mature structural biology technique to elucidate the three-dimensional structure of biological macromolecules. The Coulomb potential of the sample is imaged by an electron beam, and fast semi-conductor detectors produce movies of the sample under study. These movies have to be further processed by a whole pipeline of image-processing algorithms that produce the final structure of the macromolecule. In this chapter, we illustrate this whole processing pipeline putting in value the strength of “meta algorithms,” which are the combination of several algorithms, each one with different mathematical rationale, in order to distinguish correctly from incorrectly estimated parameters. We show how this strategy leads to superior performance of the whole pipeline as well as more confident assessments about the reconstructed structures. The “meta algorithms” strategy is common to many fields and, in particular, it has provided excellent results in bioinformatics. We illustrate this combination using the workflow engine, Scipion.

Key words Single particle, Cryo-electron microscopy, Image processing, Scipion

1 Introduction

Cryo-electron microscopy (cryo-EM) is a quickly growing structural technique capable of yielding quasi-atomic models of biological macromolecules [1, 2]. Cryo-EM structures have already found applications in structure-based drug design [3–5]. Additionally, it has the advantage of potentially identifying different conformational states [6, 7]. The recent success of this technique is due to technological advances in sample preparation [8] and microscope,

image acquisition [9], and image-processing technologies [10]. In this chapter, we focus on this latter aspect. At its simplest, a pipeline for image processing enables the parameters for acquiring images from the electron microscope to be determined. These parameters include the gain of the camera, the beam-induced movement, the aberrations of the microscope (most importantly the defocus) for each micrograph/specimen, the orientation of each particle (or class of particles), and possible changes in magnification with respect to the nominal magnification. As in all identifications of parameters in a noisy environment, algorithms will always produce an estimate of those parameters, but they may be correctly or incorrectly identified and our task is to try to discern those parameters that have been incorrectly determined.

For this task, comparing and combining the output of several algorithms is an appropriate approach. The rationale is that the local minima of an algorithm will not be the local minima of another one. In this way, if two different algorithms, with different mathematics underneath, disagree about an estimate, one of the two has to be wrong. On the contrary, if both algorithms agree (within some tolerance), we cannot guarantee that both are right, but at least, it is the best estimate we can have with the tools available. Mathematically, we are interested in unbiased estimates of the parameters. A parameter is considered biased if its expected value (i.e., the average of many repetitions of the estimation process) does not converge to the true (although unknown) parameter. If our parameter estimate is unbiased, we can obtain a better estimate by averaging several estimates. By doing so, we are also reducing the variance of our estimation.

In this chapter, we follow this principle of combining different algorithms as a way of providing a more solid scientific support to structural claims. At present, the strategy of comparing different parameter estimates is not always possible at all the steps along the image-processing pipeline. One of the reasons is that the parametrization of the different processes is not always comparable (e.g., each frame alignment program encodes the beam-induced movement in a different way and the parameter estimates can neither be compared nor averaged). At those places where the comparison and/or averaging is possible, we do it. At those other places in which the comparison is not possible, we simply choose one algorithm that has proven to be robust and to produce good results in experimental cases. We concentrate on those methods currently accessible through Scipion [11], as the integration in a single platform makes comparisons easier.

As an example dataset we have chosen the Brome Mosaic Virus dataset [12] used in the Map Challenge [13] (EMPIAR Entry: 10010, EMDB entry: 6000). This dataset nicely illustrates the difficulties to get good estimates of the underlying parameters (especially those of 3D alignment). The dataset is formed by

424 movies of 37 frames of size 4096×3072 pixels at a pixel size of 0.99 \AA taken at a JEOL 3200FSC with a DE-12 camera. The resolution reported at EMDB is 3.8 \AA , (the highest resolution reported for this dataset was 3.5 \AA , [13], the same as the one reported in this chapter). However, as discussed in Soranzo et al. [14], we should not base our full analysis of the results of a reconstruction on a single number associated to some way to measure resolution or reproducibility since the analysis process is much more complex than that (in fact, and just as an example, resolution is locally and directionally dependent, [15]. Indeed, we show in this chapter (and it is well known in the field) that significantly different maps may report the same resolution when assessed against one global number. What is more, the reported resolution is normally in the lower extreme of the resolution histogram [10, 16] so that the true resolution of the structure is typically lower than the one reported by a single number. Measures based on the ability of the map to accommodate an atomic model should be preferred, understanding that these are only possible for resolutions below 4 \AA , and in any case the reported resolution number should only be taken as a rough estimate of the quality of the map.

2 Methods

2.1 *From Frames to Valid Micrographs*

Frames are acquired at extremely short exposure times resulting in very low-contrast images, often due to the low count of electron hits. Along the acquisition process particles move under the electron beam. This was one of the reasons for the low-resolution maps of cryo-EM before the introduction of direct detector cameras. These devices have greatly improved the point spread function formally due to the scintillators (converters of electrons into photons) and have allowed quick scanning of many images, referred to as frames, with very little exposure. The set of frames corresponding to the same field of view is called a movie and the average of all the aligned frames is called a micrograph. These frames contain the structural information (Fig. 1 top), but before they can be used they have to be corrected for distortions introduced by the camera and the particle movement induced by the electron beam. The most important distortion introduced by the camera is the so-called camera gain. This refers to the fact that a uniform electron illumination is not transformed into a uniform readout of the camera (Fig. 1 middle). The reason is that each electron hit from the beam is amplified by the electronic circuitry of the camera to be translated into an electric potential that is finally read. This amplification depends on electronic currents that may change over time, and the gain correction image needs to be regularly measured. At this point, we can use the algorithm described in Sorzano et al. [17] in order to verify that the experimental images have been properly beam-corrected.

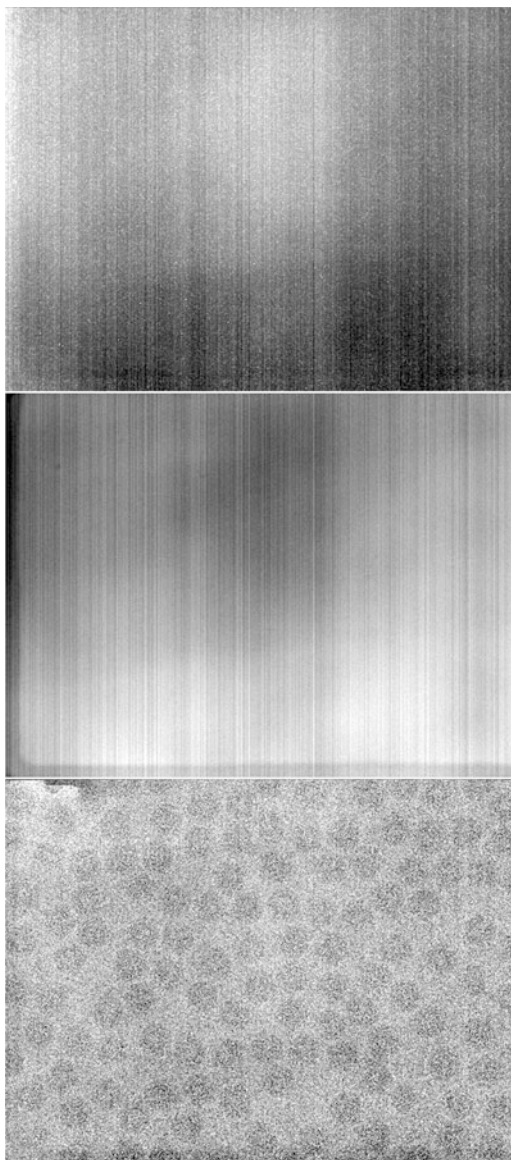


Fig. 1 Example of data acquisition (from EMPIAR 10010). Many frames like the one at the top of the figure are acquired per image field. These frames contain the structural information, but they are distorted by the camera (the middle image shows the correction required for the camera) and by beam-induced movement that has to be corrected after correcting for the gain. Once these two distortions are corrected, and after averaging the aligned frames, particles can be identified on the micrographs (bottom)

There are several algorithms to correct for the beam-induced movement. One can think of the beam-induced movement as the finding of a function that for every pixel in the micrograph tells us where that pixel is coming from at every frame (frames are indexed by i).

$$I_{\text{mic}}(x, y) = I(x + f_x^i(x, y), y + f_y^i(x, y))(1) \quad i$$

With no beam-induced movement, the functions $f_x^i = f_y^i = 0$. In general, we can expand the local beam shift in a Taylor series. The first MotionCorr algorithm [18] can be thought of as a Taylor series of order 0, in which the function is approximated by a constant. MotionCor2 [19] can be thought of as a second-order approximation of the series. Optical flow [20] would be a high order Taylor expansion, thanks to its free-form field. The problem of this latter approach is its computational cost. The solution for trading-off a high order expansion with a low computational cost is to use a deformation field expressed in terms of B-splines [21]. In this chapter, we have used this latter approach as implemented in Xmipp [22]. After estimating the deformation field, we can average the aligned frames to produce a micrograph as the one shown in Fig. 1 bottom.

The parameterization of the beam-induced movement varies among different software packages. Consequently, we cannot estimate this parameter by comparing the output of two different solutions. However, we can make “sanity checks” on the estimated motion field. For instance, the average shift between one frame and the next should not exceed a given threshold (5 Å in our example), and the overall trajectory of the movie should be below another value (15 Å in our example). These values are user defined, and they are meant to prevent incorrectly estimated fields or too quickly moving fields of view to progress along the image pipeline. In this example, only one movie out of the 424 movies failed to meet this requirement.

The next step is to estimate the aberrations of the microscope for each of the micrographs/specimens, the so-called Contrast Transfer Function (CTF) [23]. Again, the parameterization of the different packages is very diverse. However, they all compute the defocus, so that comparisons can be made at this point. In the example developed in this chapter, we calculated the defocus with GCTF [24] and CTFFind4 [25]. We then used the defocus estimated by CTFFind4 as the initialization for the CTF of Xmipp [23]. This latter algorithm has the advantage that it estimates the envelope of the CTF as well as many quality criteria of the CTF regarding its astigmatism, the visibility of the Thon rings, and the quality of the fit, such as the level of ice, correlation of the Thon rings with a 90° version of itself, visibility of the second ring, correlation between the model and the observed data between the first and third zero [26]. 77 out of the 423 micrographs were rejected for some reason related to the CTF (34 because the estimated maximum resolution was above 5 Å, 17 because the correlation between the observed and estimated CTF was below 0.03, 15 because the two CTF estimates did not coincide up to 4 Å [27], 11 due to astigmatism), Fig. 2 shows some of the rejected CTFs. 346 micrographs progressed to the next step. The workflow used for micrograph selection is shown in Fig. 3.

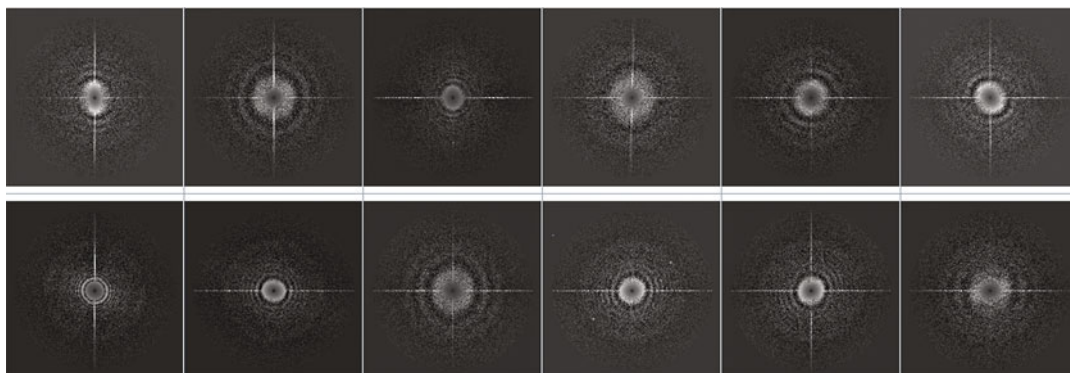


Fig. 2 Examples of rejected micrographs

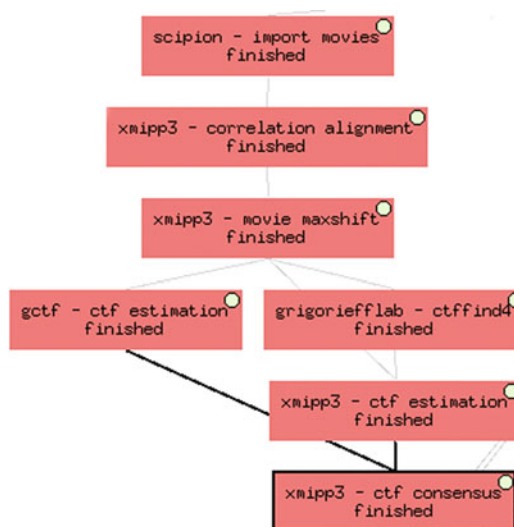


Fig. 3 Workflow used in this example to align the movie frames, estimate the CTF, and select micrographs according to their CTF quality

2.2 Finding Particles in Micrographs

The next step in the data analysis is to find particles (instances of the macromolecular complex we are working with) in the micrographs (see Fig. 1 bottom), where we refer to finding the coordinate of the center of the particle. There are three families of particle pickers, explained here in increasing order with respect to the total information they require. The first family consists of algorithms whose only input information is the particle size. Appion DoG picker [28], Relion LoG picker [29], and Sparx Gaussian picker belong to this family. The second family comprises algorithms that learn from the user or a pre-trained set how a particle looks like. These algorithms normally extract features from the images and employ a classifier to decide whether it seems to be a particle or it does not. Xmipp picker [30] and all the neural network pickers like Topaz [31] or Cryolo [32] belong to this second family. The third family

consists of algorithms that use image templates that are correlated with the micrographs at various orientations to find similar particles. Relion template picker and Gautomatch belong to this latter family. In very broad terms, the performance of the second and third families is similar, and both are better than the first family. As a rough estimate, although it obviously depends on the dataset, most algorithms have a false-positive rate between 10% and 30% (i.e., 10–30% of the found particles are not real particles) and a false-negative rate in a similar range (i.e., 10–30% of the true particles are missed). The current trend is to pick “almost everything” and then sort out the incorrectly selected particles during a 2D cleaning step (as well as other types of cleaning procedures).

In this example, we used the workflow depicted in Fig. 4. We used Relion LoG picker with a box size of 350 pixels (it found 41,695 coordinates). We also picked the particles using Xmipp picker that learns from the user the kind of particles he is interested in (it found 24,386 coordinates). We then identified those coordinates that were found by both pickers (19,173 of them) and used this subset as a positive set for training Cryolo (it found 28,596 coordinates) and Topaz (it found 28,180). Among all pickers they found 49,819 unique coordinates. 16,887 were found by all of them and 19,106 were found by only one of them. We used these two sets as the positive and negative training set, respectively, of a deep consensus approach [33]. The positive set is normally formed by well-centered particles, while the negative set is formed by off-centered particles or contaminants (see Fig. 5). The neural network of the deep consensus learns to distinguish between these two different kinds of particles and assigns a score between 0 (bad) and 1 (good) to each coordinate to assess how good a particle is. We used a threshold of 0.95 resulting in 32,880 particles. We see that this set is twice the size of the set of particles found by all of the algorithms (so that we have a lower false-negative rate). We then submit these coordinates to deep micrograph cleaner [34] that removes those coordinates that fall on for example, contaminations, aggregates, ice crystals, and carbon edges. Only 0.8% of the coordinates were in these kinds of bad regions (showing the power of deep consensus to eliminate these bad particles). The resulting coordinates are shown in Fig. 5 in which we can see that there is a good compromise between having a high positive rate, with low false positives and low false negatives.

Despite the careful particle selection performed by the very sophisticated algorithms above (machine and deep learning algorithms), there could still be images in the dataset that do not really correspond to particles. The screening method described in [35] aimed at identifying large deviations from the dataset. Its application to the particles coming out from the workflow in Fig. 4 identified 230 particles (0.7% of the dataset) that did not follow the general trends of the dataset. Some of them can be seen in

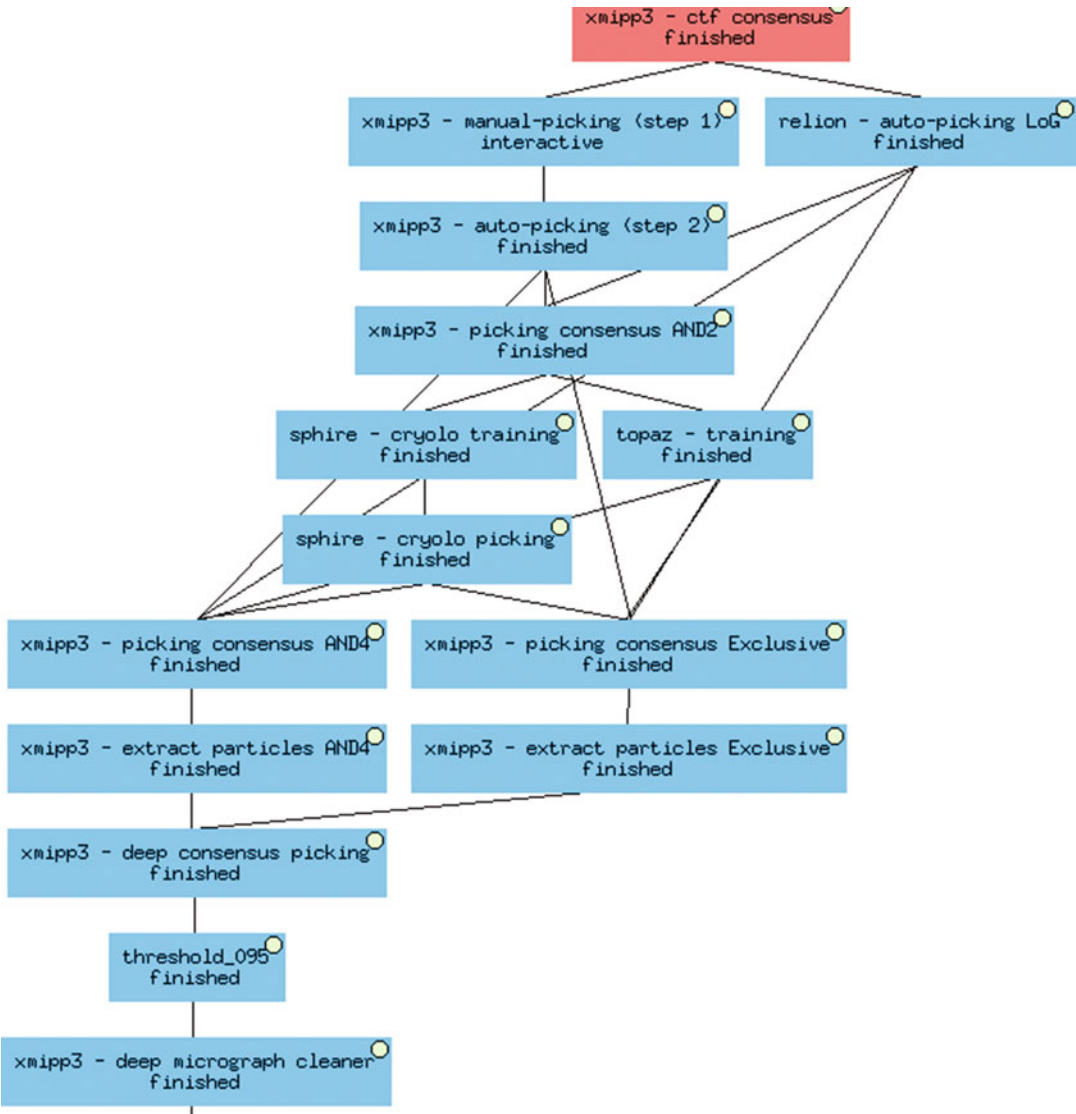


Fig. 4 Workflow used in this example to find particles in the micrographs

Fig. 6. Note that this algorithm is only capable of identifying gross deviations from the main population and that the small amount of particles (0.7%) is an indicator of two things: first, the particle selection described above is pretty accurate; second, all algorithms, no matter how sophisticated they are, have false positives and false negatives and the false positives of one algorithm do not need to be the false positives of another algorithm. In this regard, this is a good example of the need to use several methods to identify possible “pathologies” in the dataset.

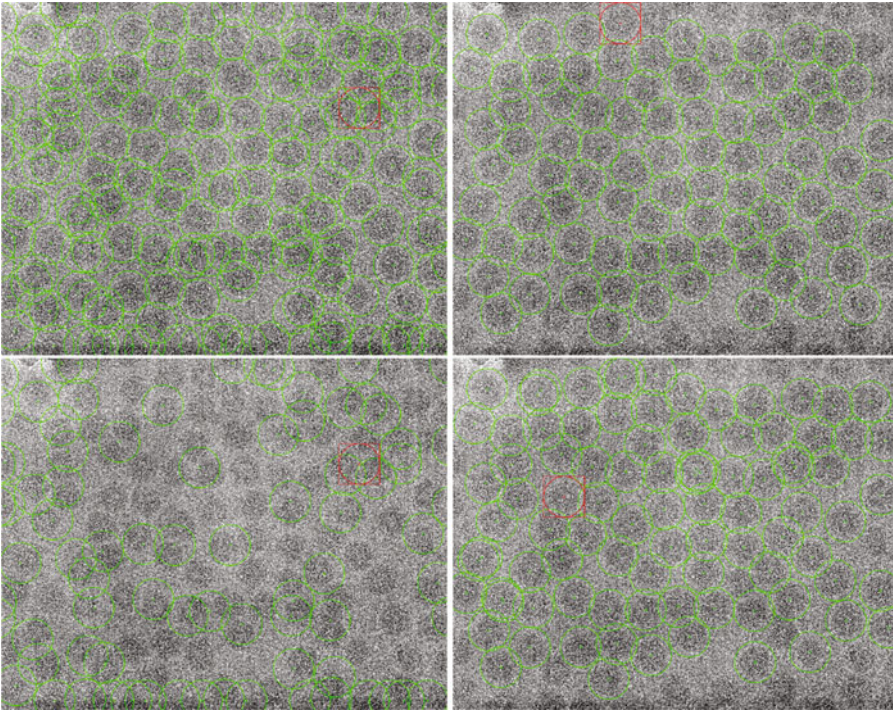


Fig. 5 Example of coordinates found for a particular micrograph by: at least one of the algorithms (top left), all of the algorithms (top right), only one algorithm (bottom left), deep consensus and micrograph cleaner (bottom right)

2.3 2D Classification

The principle of using multiple algorithms to perform the same task is well illustrated by the 2D classification step that we perform next. The goal of 2D classification is to group particles with similar shape, understanding that the shape is determined by the projection direction of the particle and the macromolecule that is imaged (its composition and conformational state). We submitted all the particles surviving the previous steps (a total of 32,299) to 2D classification. We did it using CryoSparc [36] and CL2D ([37]. The 2D classes of each one of the programs are shown in Fig. 7. At the sight of these results, we make two important notes. The first note is that CryoSparc2D classes tend to be either much larger or much smaller than the ones of CL2D. The mathematical reason for this was described in Sorenzo et al. [37] and has to do with the fact that classes with many images assigned (no matter if the assigned images really look like the class average or not) tend to have lower noise and attract even more images. This is a characteristic observed in Relion and CryoSparc, and less so in CL2D. Despite this attraction problem, the less populated classes of CryoSparc still represent spherical particles, which indicate the good quality of the particle selection performed up to this point (the typical Relion or CryoSparc 2D classification result has a few classes with many particles,

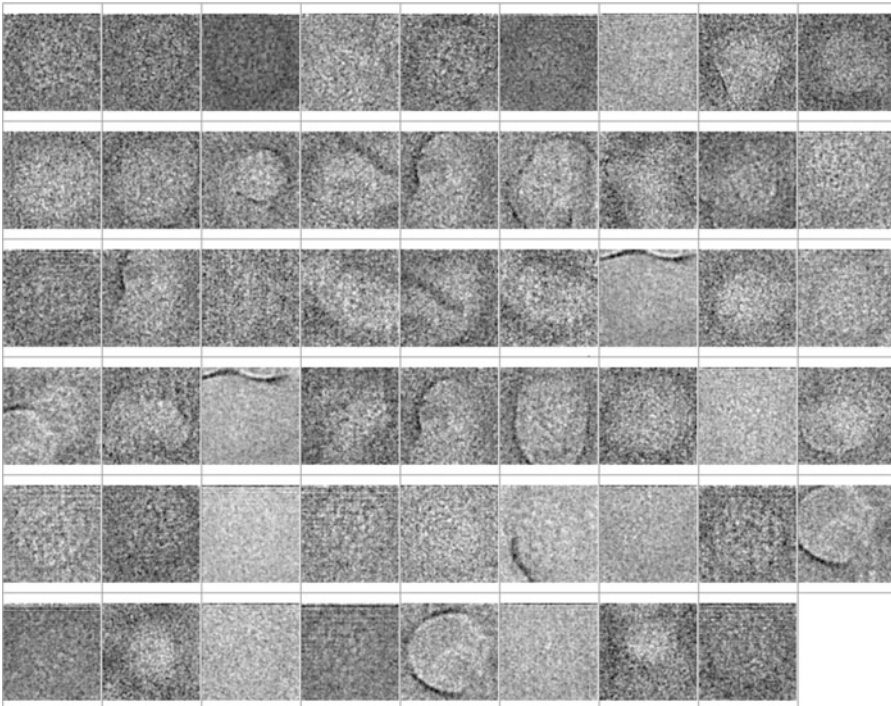


Fig. 6 Example of particles that do not follow the general trends of the dataset (Vargas et al., 2013)

and many empty classes with mostly noise). The second note is that CL2D classes are sharper, and they were capable of identifying sets of particles with a varying illumination background (marked in red in Fig. 7 bottom). This intensity gradient largely drives the alignment of these particles, and they should be excluded from further analysis as they are not well aligned. 3852 particles belonged to these classes with illumination gradients. Additionally, CL2D has the possibility of looking for outliers within the classes [38] Outliers are defined as those particles whose Mahalanobis distance to the centroid of the cluster is larger than a given threshold (typically, 3). Particles with a distance smaller than the threshold are referred to as the core of the class. CL2D is a hierarchical algorithm in which particles are first classified into a small number of classes (typically, 4), and then these classes are split into 2 (resulting in 8, 16, 32, ... classes). Particles are allowed to choose any of the existing classes at any moment. Two particles belong to the stable core of a class if they were always together along the classification process (i.e., when there were 4, 8, 16, ... classes, they were always together). We have observed that empty particles and contaminants tend to be randomly spread over the classes and, consequently, they tend to jump between them. 6.6% of the particles were identified as outliers within its class, resulting in a total of 26,793 particles. At this point, we have a set of particles assigned to 2D classes that are only

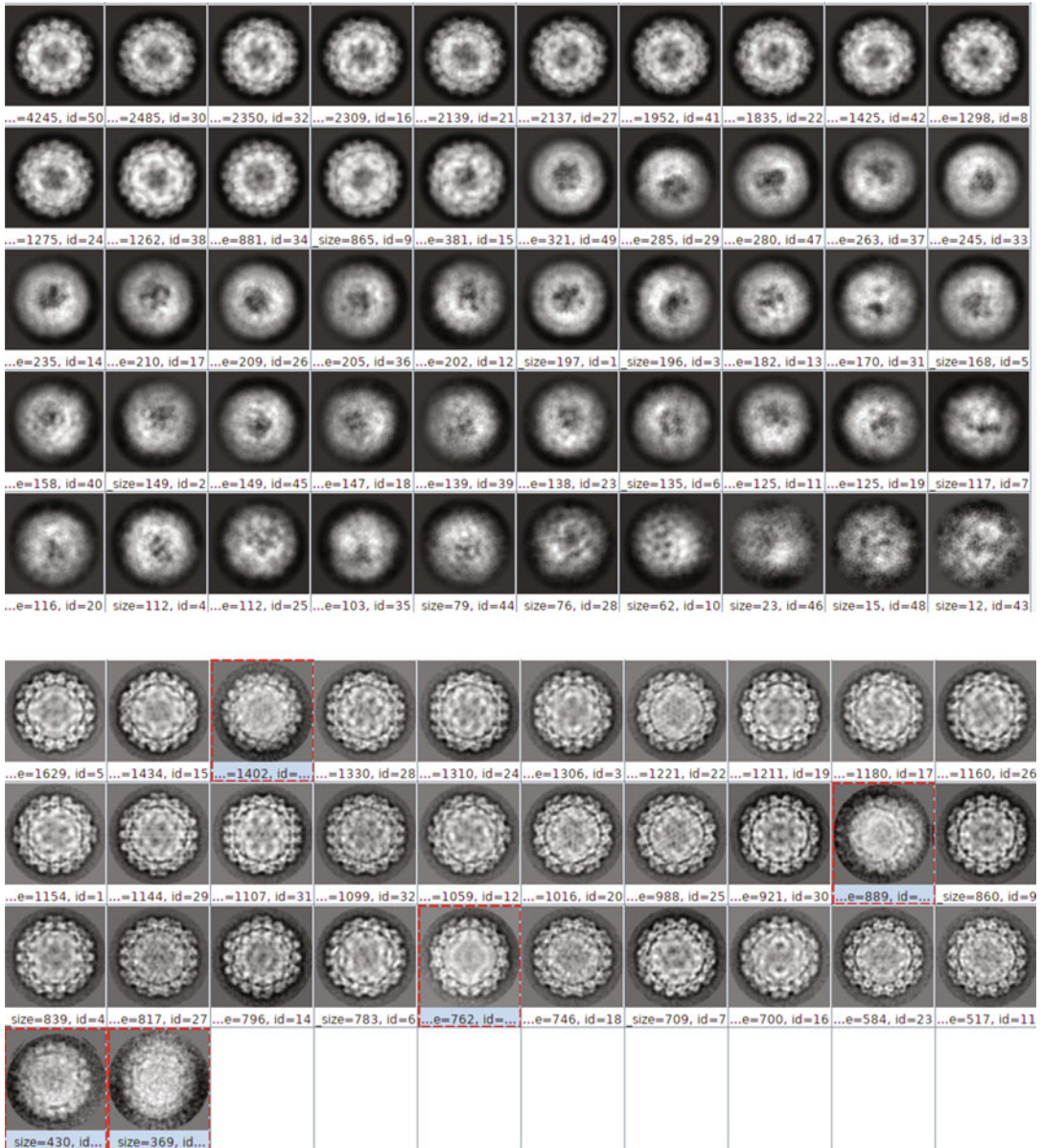


Fig. 7 2D classes calculated by CryoSparc2D (top) and CL2D (bottom)

roughly centered. Therefore, we can refine the coordinates within the micrograph so that they correspond better to the particle center. By doing so, we will be able to identify that some of the allegedly independent coordinates were actually pointing to the same particle (see Fig. 5 bottom right). The protocols within Scipion for this analysis are Xmipp center particles and Xmipp remove duplicates. This step reduced the set of particles by 9%, leaving 24,199 particles available for 3D reconstruction. Another important lesson from this part of the analysis is that among all

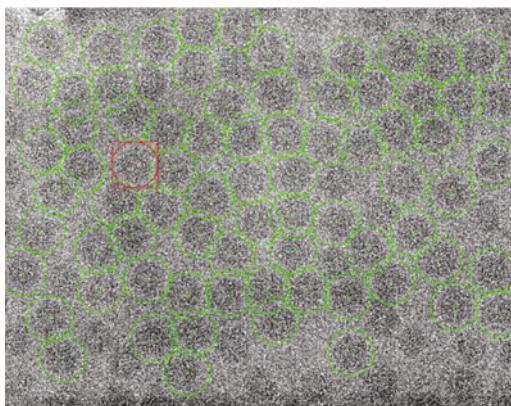


Fig. 8 Example of unique coordinates after 2D classification, centering particles and removal of duplicated particles

identified “particles” at the beginning, selected by at least one sensible and sophisticated algorithm (49,819), less than half of them (24,199) were really independent and supposedly to be good particles (see Fig. 8). This is in agreement with standard observations in the field in which many of the “picked particles” are discarded along the 3D analysis. We should see it as an indication of the fact that most of the picked particles were not really particles, but only suggested coordinates (some of them even pointing to the same particles).

All the 2D classification analysis was performed with particles whose pixel size was 3.46 Å (the box size at this sampling was 100×100). There are two reasons for this. The first one is that they occupy less space and all calculations are faster. For this 2D screening task, we do not need “atomic” resolution. The second one, and maybe more important, is that making the images smaller removes a lot of noise that may cause image misalignment. In this way, we are increasing the signal-to-noise ratio and having a better analysis with a resolution adapted to the needs of the task at hand.

2.4 Constructing an Initial Map

Once we have selected a set of particles, we may proceed to construct an initial volume with them. This is one of the critical steps of the whole procedure because if a bad volume is constructed, it is very likely that the subsequent algorithms cannot escape from this local minimum. Scipion integrates many different algorithms for this task, including Xmipp Ransac [39] and Reconstruct Significant [40], CryoSparc [36], Relion Stochastic Gradient Descent [41], Eman [42], and Simple [43]. All these algorithms employ different flavors of stochastic optimization (an optimization that is not too greedy and allows steps in which the goal function is worse than it was in the previous step), and a simplified optimization landscape (normally by filtering the images, reducing their size, or working with class averages rather than raw images). The goal of these two

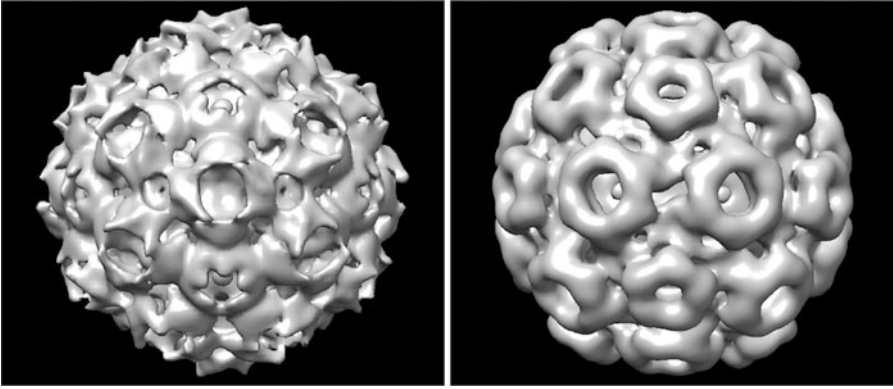


Fig. 9 Example of incorrect (left) and correct (right) initial volume for the Brome Mosaic Virus

strategies is to avoid local minima and try to find the global minimum. However, this is not always achieved for all the proposed volumes. Figure 9 shows an incorrect (local minimum) and correct (close to the global minimum) initial volume proposed by two different algorithms (in this case RANSAC, incorrect, and reconstruct significant, correct; RANSAC proposes a set of candidates and only one of them, shown in the figure, was incorrect). This example was relatively easy, and most initial volume algorithms returned a correct structure. However, this is not always the case and, depending on the dataset, most of the candidates may correspond to local minima. The suggested procedure in this chapter is to run multiple algorithms proposing a diverse population of candidates to initial volume. Traditionally, it was the user who had to choose one of them to continue the processing. At present, there are algorithms that are capable of considering all these candidates and, either by letting them evolve (swarm consensus, [44]) or by comparing them [45] automatically decide (normally correctly) a suitable initial volume.

All this analysis can be performed at relatively low resolution (pixel size 3.6 Å in our example) as a way to speed-up calculations, and more importantly, to smooth the goal function landscape by removing noise. One of the most useful analyses before going further is to compare the re-projections of the initial volume with the 2D class averages. Even the least matching classes should be in good agreement. Figure 10 shows these comparisons for the two initial volumes proposed in Fig. 9. Differences are subtle (sometimes they are much more obvious), showing the difficulty encountered by the initial volume algorithms, but an analysis of the similarity between these re-projections (see Fig. 10) indicates which one of the two is better (the bottom one, since the histogram of correlations is shifted towards higher values). If experimental SAXS data is available, then the SAXS curve of the different

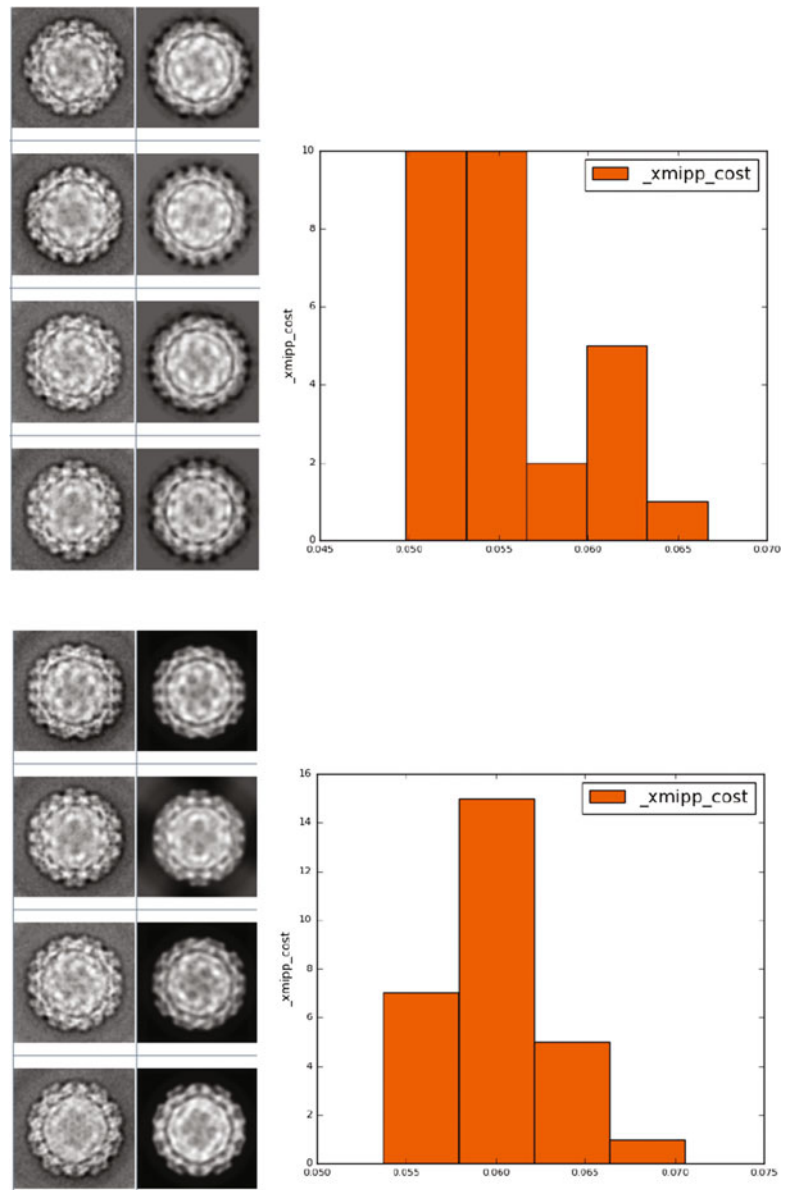


Fig. 10 Comparison of the re-projections of the initial volume and the 2D classes for an incorrect initial volume (top) and a correct one (bottom). For each of the volumes, the least correlating 2D class averages are shown (left) along with the corresponding re-projection (right) and the histogram of all correlations within the comparison

proposals to initial candidate can be simulated and compared to the experimental SAXS curve. Jimenez et al. (2019) [46] shows that this strategy can indeed distinguish between incorrect and correct initial volumes.

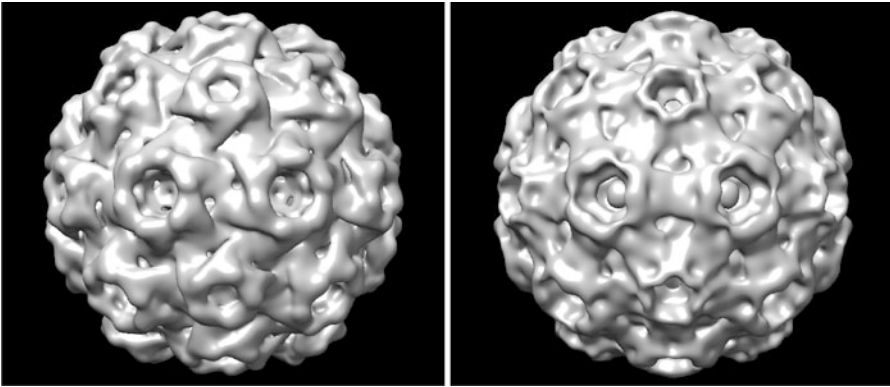


Fig. 11 The 3D Classification of 24,199 particles of the Brome Mosaic Virus into two classes by Relion. The first class (left) received 68.4% of the particles, while the second class (right) received 31.6%

2.5 Achieving an Homogeneous Population

The next step in the pipeline is to separate the selected particles into structurally homogeneous datasets. This is done through 3D classification steps. Using the same rationale as in the previous steps, we may perform this analysis at an intermediate pixel size (1.73 \AA in our example, resulting in a box size of 200 pixels). The two algorithms for 3D classification in Scipion are Relion and CryoSparc. We used Relion to separate into two classes the 24,199 particles selected in previous steps. 68.4% of the particles went to a class similar to the initial volume and 31.6% to another class (see Fig. 11).

In most publications, the classification step is executed only once. However, we were interested in the reproducibility of this class assignment, and we ran this process three times. The number of particles assigned to the first class ranged from 68.4% to 82.8%, but the two classes were invariably the same. However, and very interestingly, only 55.8% of the images were always assigned to the first class in all three classifications (this set can be calculated using the consensus classes 3D protocol of Scipion). This is an indication of the instability of the classification process. The belonging or not to a class is another parameter that needs to be estimated in a very noisy environment, and consequently its estimation is prone to errors. This statistical principle is not generally well rooted in the 3DEM community, and most papers perform a single classification step taking its result as the “ground true” classification of the dataset at hand. The take-home message should be to trust only those subsets of images that consistently have been assigned to the same class because the class assignment is another parameter to estimate (and one of the most important ones), and there could be errors in its estimation. Naturally, the key point is to know how this clear instability associated to the angular estimation translates into the map itself.

In very general terms, let us assume that among all the particles assigned to class 1, there is a fraction α (between 0 and 1) of particles that truly belong to class 1 and $1 - \alpha$ that truly belong to class 2. Then, by miss-estimating this $1 - \alpha$ fraction, the estimated class 1 volume would be (in a simplified manner, because the true dependence also depends on the angular assignment).

$$\hat{V}_1 = \alpha V_1 + (1 - \alpha) V_2$$

Still, the way the additional term $(1 - \alpha) V_2$ translates into the map is virtually impossible to know. If errors in the alignment follow a more or less random pattern, then the result will be the introduction of a blur-like in the map, which may not affect its interpretation after sharpening and applying a threshold. Indeed, this is the case for main class 1 since the map does not virtually change. However, this is not the case for class 2, where the lack of analysis of alignment instability would have led to totally wrong conclusions, as we present in the following.

The fact that between 31.6% and 17.2% of the particles are assigned to class 2 would lead us to think that class 2 is a truly existing class in the dataset (we may even try to find its biological role). However, we suspected that this was an image-processing artifact. For testing this hypothesis, we selected the particles assigned to class 2 and ran an initial volume protocol (Relion stochastic gradient descent, in this example) obtaining the volume in class 1! Actually, in this dataset there is only one distinguishable conformation. In this example, class 2 comprises a mixture of particles with a correct angular assignment and particles with a mirrored angular assignment. This exercise brings two very important take home messages: (1) class separation could be partially or totally artifactual due to the image processing; (2) even if the parameter of class belonging is correctly estimated, the angular orientation of a particle with respect to the class may still be incorrect. The only way of verifying if this is our case is by estimating the class parameter and angular assignment multiple times, preferably with different algorithms based on different mathematics and only trust those classes and angular assignments that are consistent between algorithms and runs. Still, being consistent is not a guarantee of being correct. But, being inconsistent is a guarantee of being incorrect.

2.6 Refining an Homogeneous Population

Once we have divided the set of particles into structurally homogeneous datasets (in our example, the 24,199 particles belonged to a single structural class) the final step is to align the particles with respect to an initial volume that must be refined. Xmipp highres [38], Relion [47], and CryoSparc [36] can be used for this task within Scipion. From the previous steps, we now understand that we are estimating the pose parameters of the particles and that these

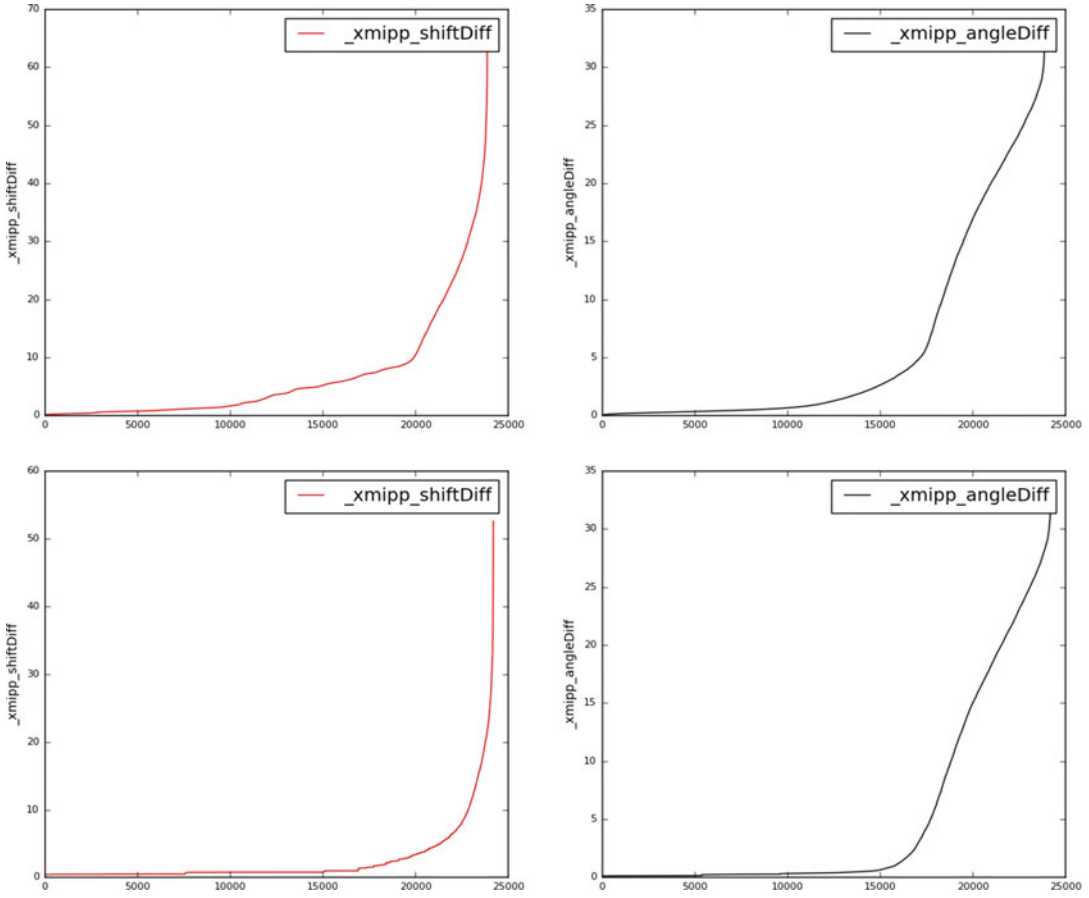


Fig. 12 Comparison of the shifts (left, in pixels) and angles (right, in degrees) of Relion and Highres (top), and two independent runs of Relion (bottom)

parameters may be estimated with noise. For this reason, we ran Relion and Xmipp Highres. Both algorithms agreed in the angular assignment of the image for about 50% of the particles (see Fig. 12, top). For illustrative purposes, we show in the same figure (bottom) the shift and angular comparison of two independent runs of Relion (they agree for about 75% of the particles).

This disagreement between two reasonable angular assignment methods (or even a method with itself) speaks about the difficulty of estimating parameters in noisy environments and the need to confirm the validity of those parameters. As with the incorrect 3D classification, if a fraction α of the particles are correctly aligned, while $1 - \alpha$ are incorrectly aligned, our reconstruction will be

$$\hat{V} = \alpha V_{\text{correct}} + (1 - \alpha) V_{\text{incorrect}}$$

Identifying the correctly assigned particles is a difficult task with experimental data as the ground truth is never known. At this point, we have found useful performing a 3D classification of the aligned

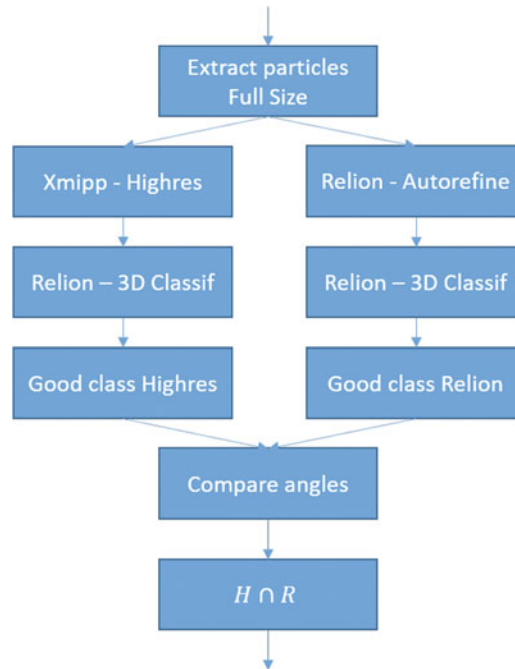


Fig. 13 Starting from a common set of particles at full size, we perform two independent angular assignments with Xmipp Highres and Relion Autorefine. The outputs of both programs are classified into 3D without re-estimating the angular orientation, and the images assigned to a good class are identified. We then compare the angular assignment of these good images between Highres and Relion. We label as $H \cap R$ the set of images with similar angular assignment in Highres (H) and Relion (R)

particles in two classes without re-estimating the angular parameters, but just the class parameters. We have observed that normally the set of particles is separated into a class that gives raise to a high-quality reconstruction (and we assume that most of the particles in that class are reasonably aligned) and another class with lower resolution (and we assume that most of the particles in that class are not well aligned). We may then compute the intersection of the two good classes (one from Highres and one from Relion). 10,471 particles (47.2% of the original subset) were classified into both classes as good, and the corresponding angular assignments differed in less than 2.5° and 4 pixels (1% of the image size). See Fig. 13 for a graphical summary of this strategy. Although we can never guarantee that the angular assignment is correct with experimental data, we know that this set of 10,471 particles were assigned very similar 3D pose by two independent 3D reconstruction algorithms and that they were assigned to a high-quality 3D class by two independent executions of a 3D classification algorithm. At this point, and to the best of our ability to estimate parameter errors, we would say that these images correspond to well-assigned particles

within a relatively homogeneous class. A drawback of this approach is that about half the particles have not made their way to the final reconstruction, and indeed this is a problem for an example in which the resolution is known to be limited by the number of particles (see Fig. 21 and related comments). Perhaps better strategies can be derived in the future to reduce this loss, possibly following ways conceptually comparable with some of the methods for “rescuing” particles while picking that were presented in past section, but they will always be rooted in the key realization that some way to estimate parameter instability has to be incorporated into the whole approach.

2.7 Final 3D Reconstruction with Consensus Geometry

Scipion has the possibility of merging different sets of particles while keeping their angular assignment. Additionally, the angular assignments by any of the two methods (Xmipp Highres or Relion Autorefine) are both expressed in a common geometrical framework [48] so that we may continue the 3D reconstruction process benefiting from our best estimates disregarding their origin (although CryoSparc has not been used in this experiment, the internal angular consistency within Scipion would be the same). We used Xmipp Highres local iterations for this task. This algorithm allows local refinement of the 3D pose parameters (Euler angles and in-plane shifts), as well as possible anisotropic magnification errors, local defocus values, and gray normalization parameters. In Fig. 14, we show the FSC of Highres and Relion Autorefine as well as representative slices of the two reconstructions. The calculated resolution by Highres is 3.5 Å. However, this number should not be used as the sole measure of map goodness. More interesting than the low end of the FSC curve is its behavior before it starts to fall down (before the 0.5 threshold). We observe that the Highres reconstruction is much more consistent (closer to 1) in a wider range of frequencies. This behavior of the Highres reconstruction is also observed in the slices, as shown in Fig. 14, where the Relion reconstruction seems to be a low-pass filtered version of the reconstruction of Highres. It should be noted, however, that Highres introduces some nonlinear constraints in the reconstruction process for noise suppression; from this point of view, it can be considered that it incorporates some form of masking, while in Relion this is a separated process. The introduction of these nonlinear constraints also explains the fact that the FSC does not always fall to zero.

Local resolution can be measured in Scipion by any of the standard tools in the field: Blocres [49], Resmap [50], MonoRes [10], or DeepRes [51]. We note that this latter method is mask-invariant, which will be an important feature later on in this work. We have found that the resolution reported by the FSC is typically in the lower extreme of the resolution histogram reported by the local resolution tools [51]. In this case, most voxels are in the range

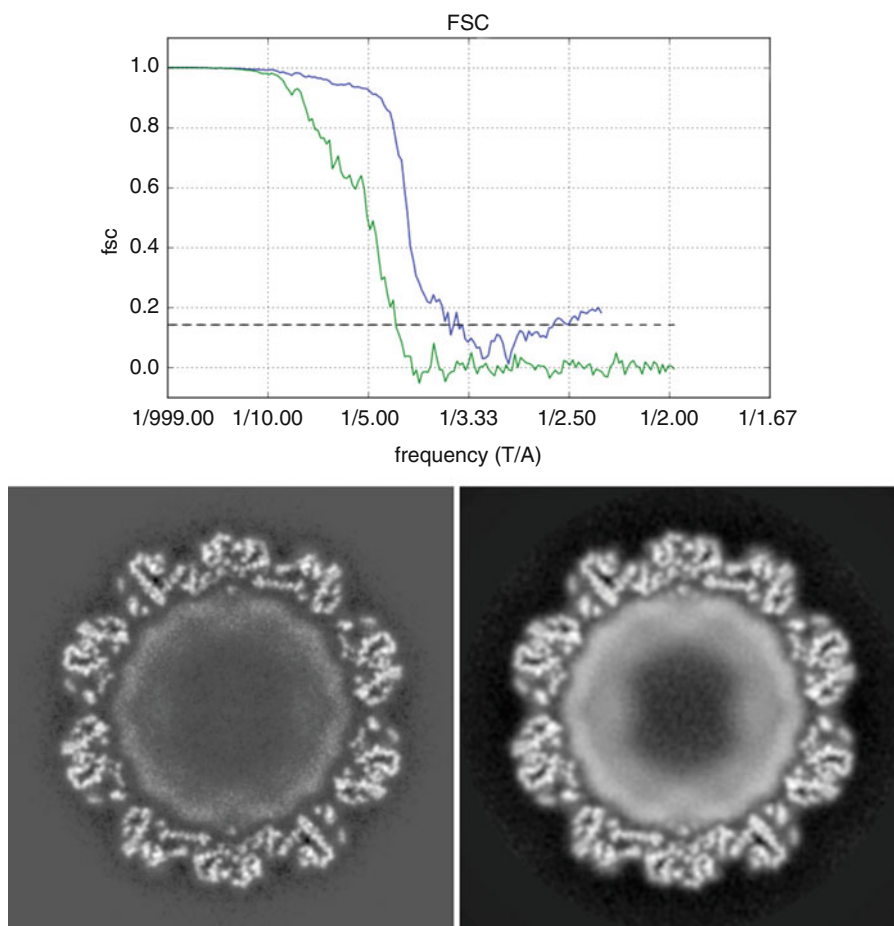


Fig. 14 Top: FSC of Relion Autorefine (green) and Xmipp Highres local refinement (blue) after the particle selection and angular assignment strategy described along the chapter. Bottom: Representative slice of the reconstruction with Highres (left) and Relion Autorefine (right)

between 4 and 6 Å as reported by DeepRes, and this estimation is consistent with the visual appearance of a close-up of the capsid structure (see Fig. 15).

2.8 Post-Processing

At present, the two most widely post-processing tools are masking and B-factor correction [52]. The FSC is insensitive to the B-factor correction because it is an isotropic filter, but it is affected by the mask [14]. Actually, the improvement in FSC that we see in Relion post-processing (see Fig. 16) is purely due to the change of mask between the FSC measured during the reconstruction and one resulting from the mask used in the post-processing. The dependence on the mask poses a real problem to the FSC as an objective evaluator of the quality of the reconstructed map, as different reasonable masks (or even over-smooth masks) result in different resolution measurements (see Fig. 17).

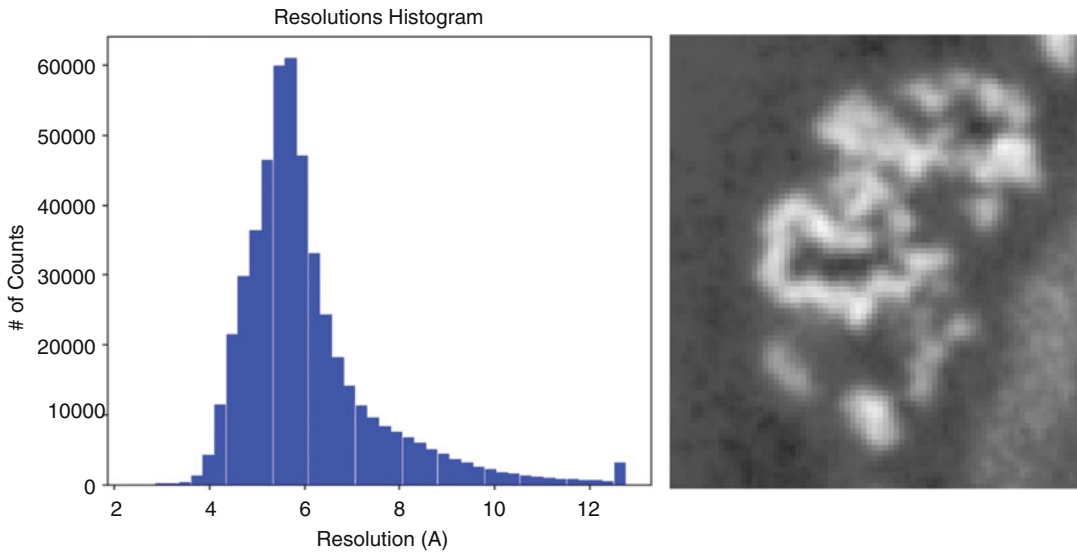


Fig. 15 Left: Local resolution histogram calculated by DeepRes. Right: Zoomed version of one of the capsomers of the Highres slice shown in Fig. 14

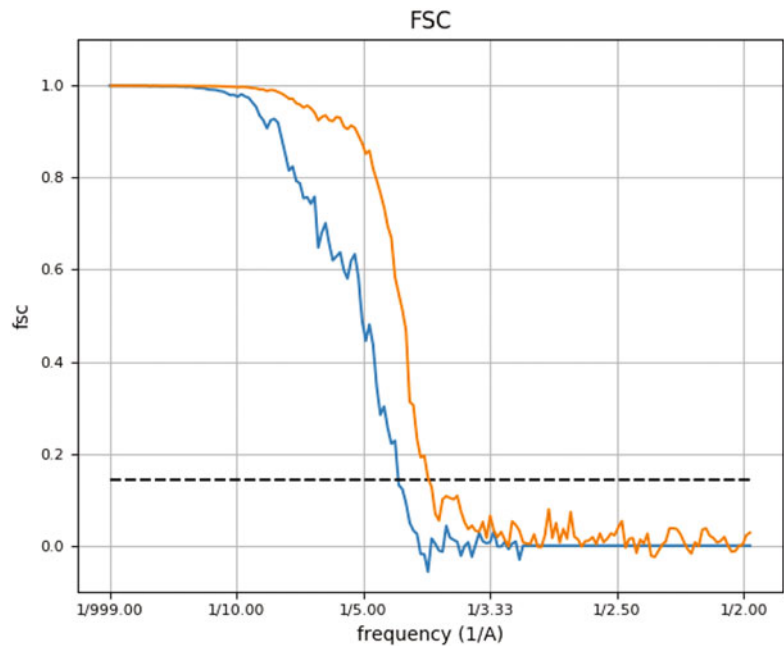


Fig. 16 Difference observed in Relion post-processing between the unmasked (blue) and masked reconstructions (orange)

Beside the dependence on the mask, the current post-processing practice of boosting high frequencies by applying a B-factor has the problem that it does not take into account local

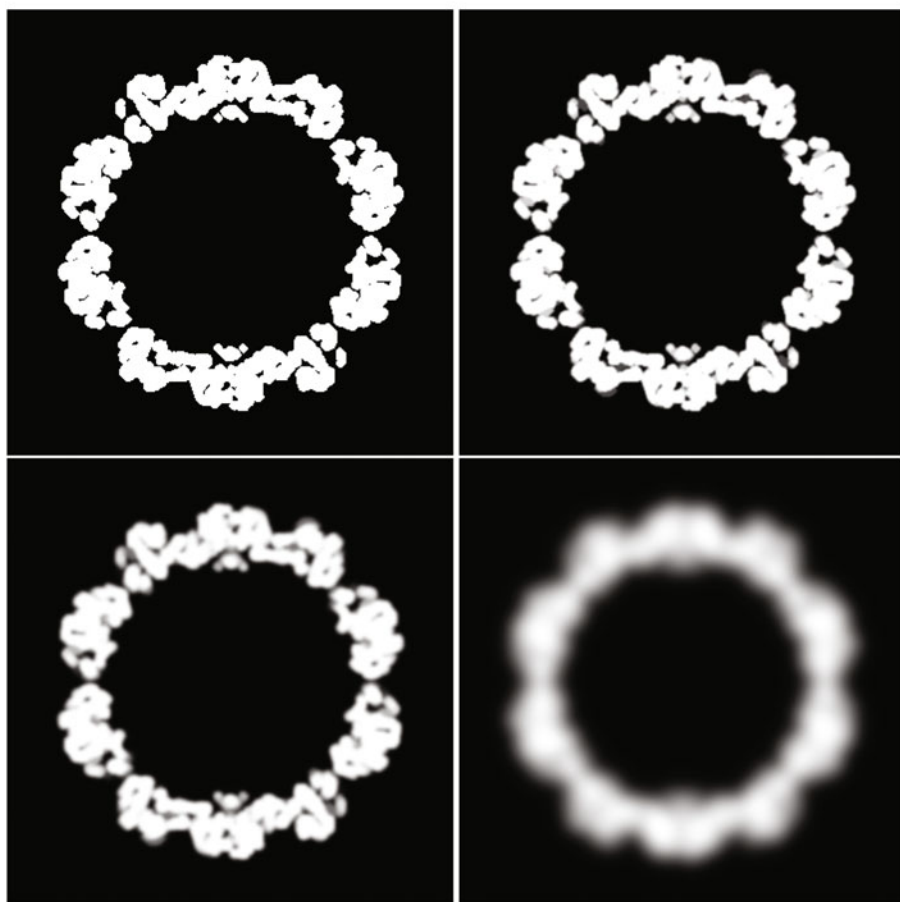
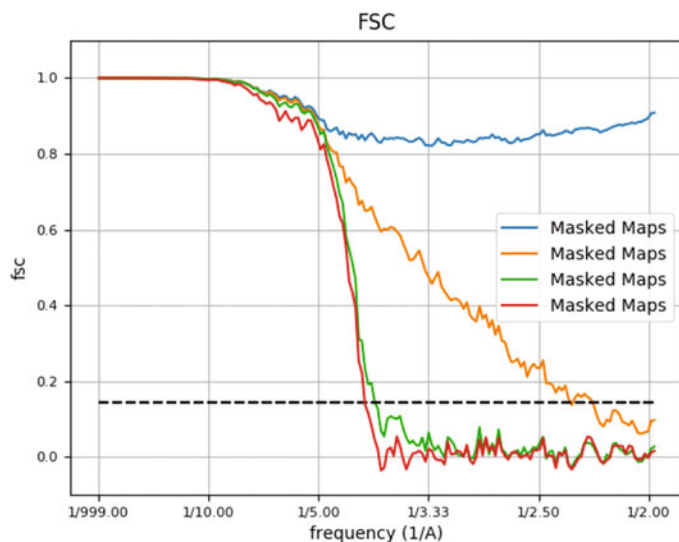


Fig. 17 Top: Different FSCs resulting from different masks. A tight, sharp mask (top left) results in an FSC that does not drop from around 1. If this mask is smoothed with a Gaussian of $\sigma = 1$ (almost undistinguishable by eye with respect to the sharp mask), 2 or 10 (clearly oversmoothed), the FSC drops from almost 1 to a different curve. Interestingly, the FSC does not distinguish the maps masked with masks as different as the ones convolved with $\sigma = 2$ and $\sigma = 10$

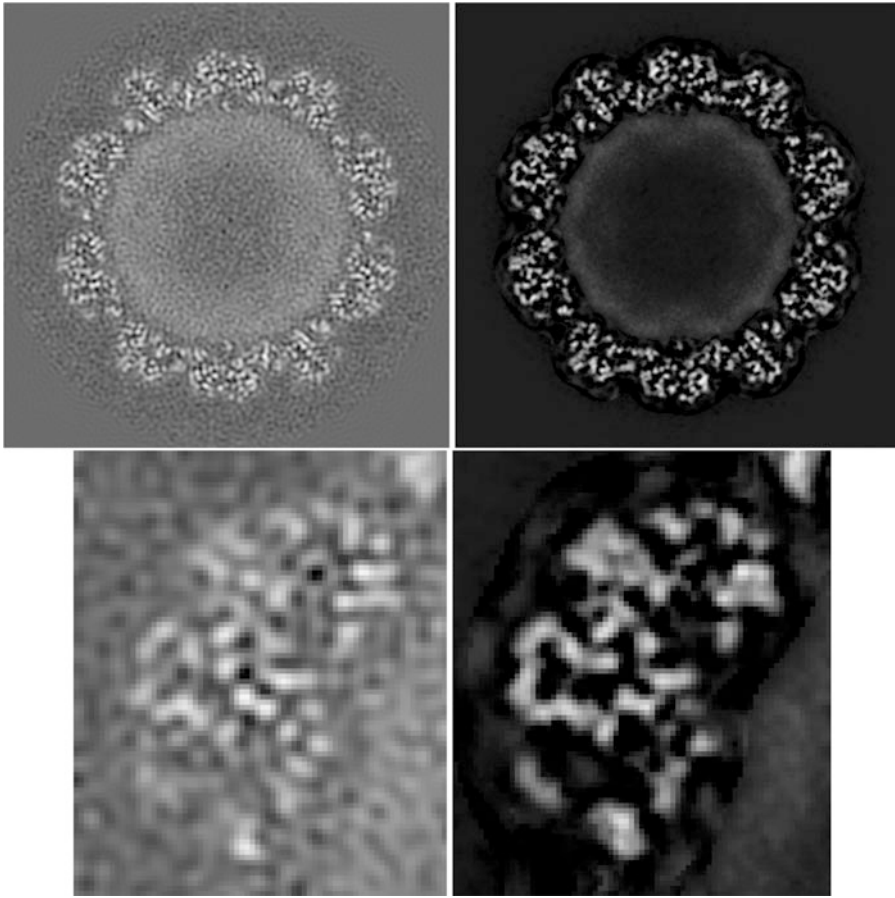


Fig. 18 Top: Representative slice of the B-factor corrected map by Relion post-processing (left) and Local Deblur (right). Bottom: Zoomed regions of the maps above

differences in cryo-EM maps (usually studied by the local resolution) and produces maps whose Fourier spectrum does not match the expected decay predicted by the diffraction theory [53]. Alternative methods have been proposed based on the matching of the spectrum falloff of the reconstructed map to the falloff of a fitted atomic model [54] or the use of the local resolution to locally deblur the reconstructed map [51]. The drawback of the first approach is that it requires fitting an atomic model to the reconstructed map, which is a whole job in itself, but it works very well otherwise. In Fig. 18, we show the maps produced by B-factor correction of the Relion autorefine map and by local deblur of the Xmipp Highres. The appearance of both reconstructions is totally different at the level of noise (B-factor is expected to boost noise, while local deblur to suppress it) and, consequently, map interpretability is directly affected by the appearance of the map we are looking at. Moreover, in Ramírez-Aportela et al. [51] we showed that the decay of the deblurred maps in Fourier space corresponds

to the expected behavior of biological macromolecules as predicted by the diffraction theory. The local resolution histogram of the deblurred map, as measured by DeepRes (which is not affected by masks), shifted from the 4–6 Å region to the 3.5–4.5 Å.

2.9 Validation

An important part of the image-processing pipeline is the checking of some necessary conditions that cannot guarantee the correctness of the map, but failing to meet them guarantees its failure. First, we must visually inspect the slices of the map: stripes, especially radial, or artifacts outside the map should not be present. This visual inspection is most useful in the raw result of the 3D reconstruction process (Fig. 14). The angular distribution of the map should also be inspected. For symmetric maps (especially highly symmetrical like the one used in this work), it is difficult to evaluate the homogeneity of the angular distribution as angles are only estimated within the asymmetric unit. We may break the symmetry by randomly assigning to each particle one of the equivalent positions in the projection sphere and, then, evaluate whether or not the angular distribution is uniformly covered. In Fig. 19, we show the angular distribution (with broken symmetry) of the example developed along the chapter. In spite of not being perfectly uniform, it is sufficiently varied as to prevent large regions of the Fourier space

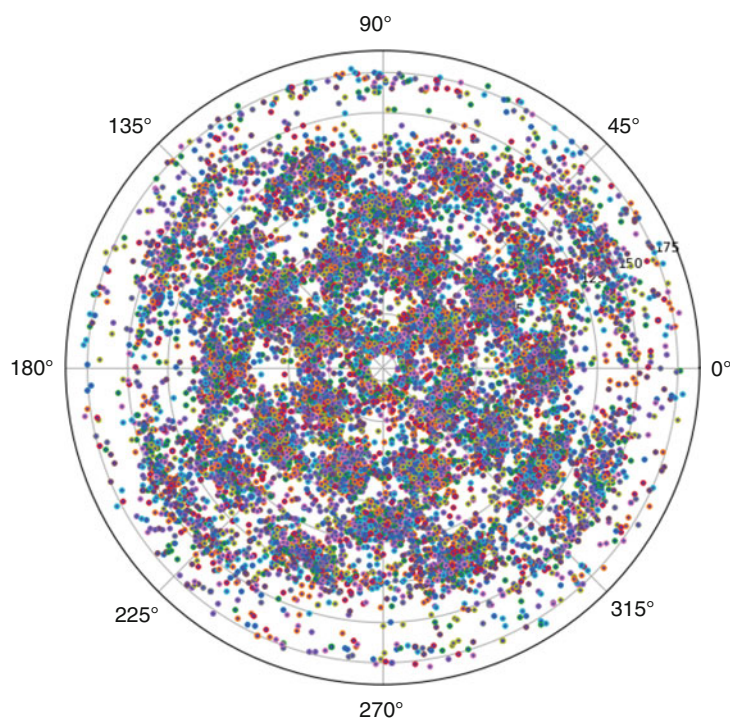


Fig. 19 Projection sphere in which the projection direction of all particles is represented

not being measured. Although there are algorithms to quantify the evenness of the angular distribution [55], these are not generally applied and thresholds to decide whether the angular distribution is sufficiently good or not are not available.

Vargas et al. [56, 57] proposed a couple of metrics to measure the “alignability” of the particle set (they referred to them as soft validation indicators). The first measure (angular precision) analyzed the distance between the top N best angular assignments (particles whose best N angular assignments are scattered over the projection sphere are less reliable than particles whose best N angular assignments are all clustered around the same projection direction; by default $N = 7$). The second measure (angular accuracy) analyzes the consistency between the final angular assignment, which followed a global to local optimization with the risk of getting trapped into a local minimum, and a global angular assignment performed de novo when the final structure is known (those particles for which the two angular assignments coincide are more reliable than those for which they do not). These measures are translated into a score per particle between -1 and 1 . Ideally, good particles should have a score above 0.5 in both metrics. These two metrics are accessible in Scipion under the protocol multi-reference alignability and the result for this example is shown in Fig. 20. The average precision and accuracy in the dataset are both 0.89 that is well above the 0.5 threshold that is required to have a reliable angular assignment.

Finally, we may wonder whether by adding more particles we would gain more resolution or, otherwise, the resolution is limited by other factors (heterogeneity, misalignment, low-resolution images, microscope aberrations, etc.). This can be determined by a ResLog plot [58] that shows the increase in resolution as new particles are added. Heymann (2015) [59] showed that an increasing number of pure noise particles also increased the resolution of the reconstructed volume. The idea is that the resolution obtained at a given number of noise images (i.e., no particles inside that can also be referred to as noise particles) should always be smaller than the resolution obtained with the same number of true particles (see Fig. 21). However, if the map is overfitted, the resolution obtained with noise particles reaches a similar level to the one of supposedly true particles. This plot also answers the question of whether the resolution is limited by the number of particles or by other factors. If the plot saturates (reaches a plateau), then the resolution is not limited by the number of particles since adding more particles does not seem to increase the resolution. In our example, the resolution is limited by the number of particles.

2.10 Interpretation and Model Building

Once we have validated the map, we may proceed to the final step: visualizing the map and interpreting it. Isosurfaces at a given threshold or any other more sophisticated segmentation tool (like

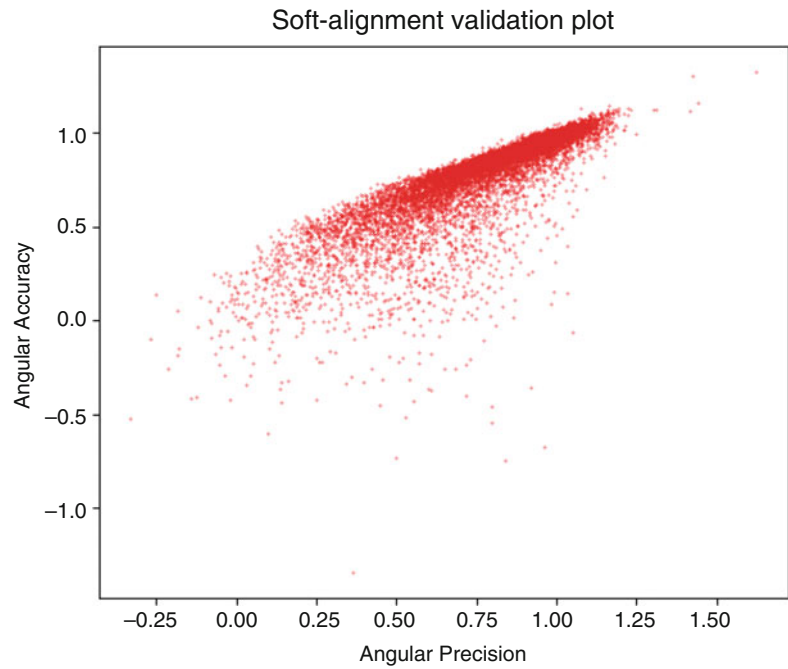


Fig. 20 The angular precision and accuracy for the set of particles used in the 3D reconstruction of the virus shown in this chapter

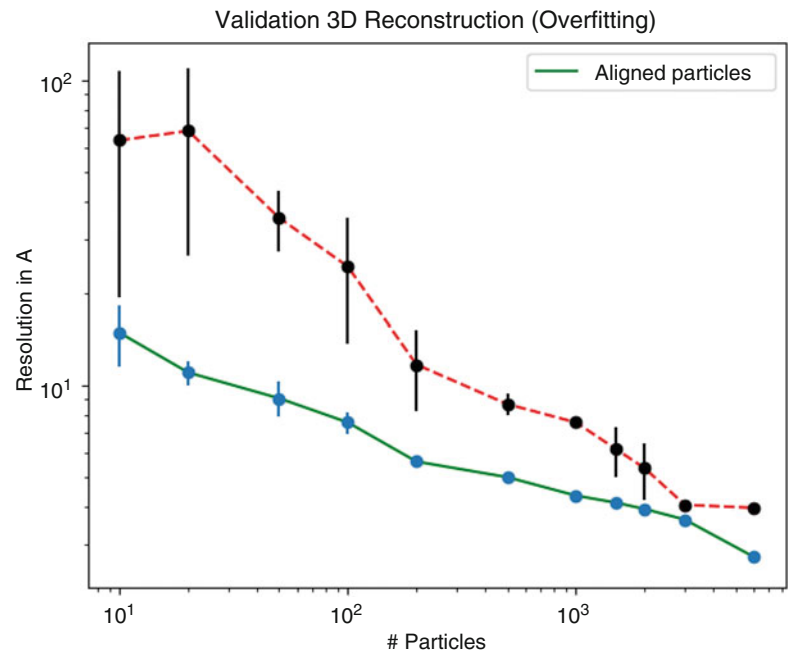


Fig. 21 Plot of the resolution vs the number of particles used for the reconstruction. The solid line at the bottom part refers to true particles, while the one above corresponds to noise particles

the one based on the False Discovery Rate, [60]) play an important role in this regard as they help to provide a continuous surface that contains the macromolecule (see Fig. 22).

Biological conclusions can be derived at this point. However, the most common next step is to build an atomic model with the help of the constraints given by the EM measurement. The map and atomic model together are usually referred to as a hybrid model (see Fig. 23). This model building is rather time consuming and can be regarded in itself as full of decisions and workflow branches as the process leading to the map that has been described in this chapter. The parameters sought in this case are the spatial location of each one of the atoms in the macromolecule. Scipion is expanding in

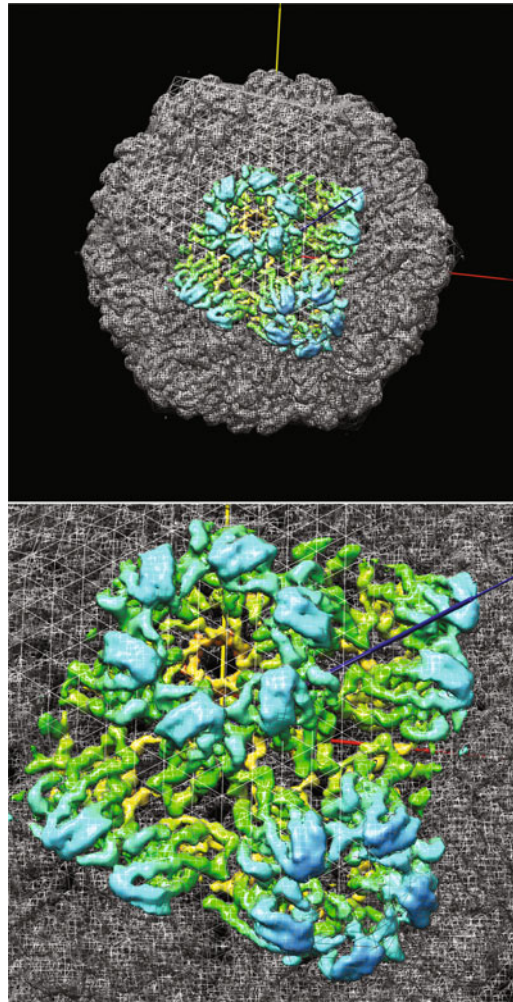


Fig. 22 Isosurface representation of the virus capsid (top) and a zoomed version of it (bottom). The whole capsid has been represented by a gray mesh, while the unit cell of the virus (and a small region around it) has been highlighted in colors according to its distance to the virus center

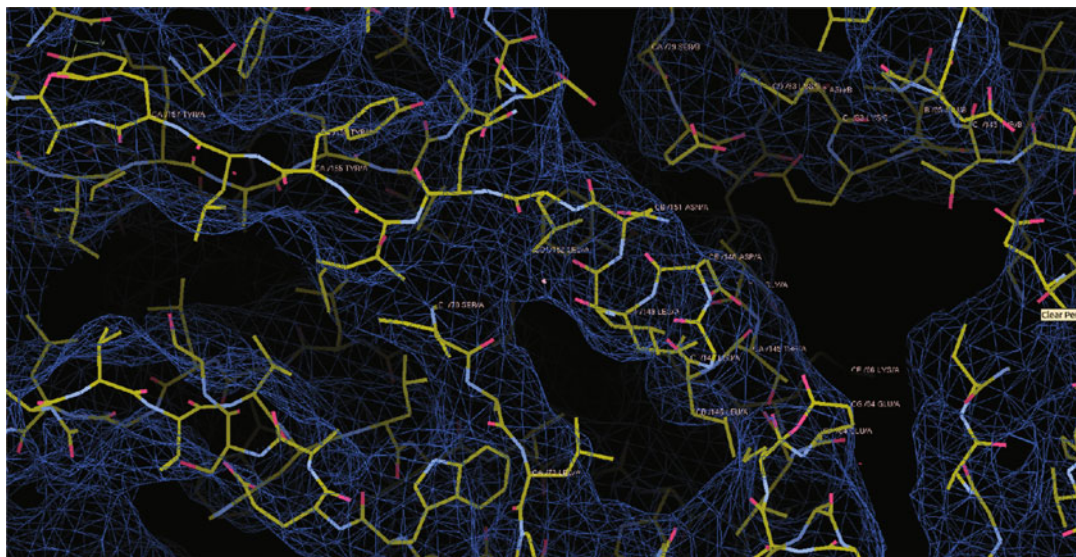


Fig. 23 Detail of the fitting process in which the EM map is used as a spatial constraint to construct an atomic model of the macromolecule

that domain, and it has incorporated some of the most common algorithmic tools used for this task [61]. The interested reader is referred to this latter publication and to the extensive tutorial about model building at <https://scipion-em.github.io/docs/docs/user/user-documentation.html#tutorials> for expanding this information.

3 Conclusions

There is an old saying in Electron Microscopy that goes like “the good thing about 3D Electron Microscopy is that you always get a volume; the bad thing about 3D Electron Microscopy is that you always get a volume.” It refers to the fact that by acquiring some images at the microscope and by processing them we will always get a map that pretends to be a faithful model compatible with our measurements. Our success in effectively achieving this goal depends on our ability to choose good micrographs preserving structural information to high resolution, particles that really correspond to a single structure, and being able to find the relative orientations of these particles in space. Making incorrect decisions in any of these steps will necessarily degrade the quality of the reconstructed map. In the limit, bad data quality, population mixtures, or incorrect alignments may yield the reconstructed map useless. Validation tools can be regarded as sanity checks that good maps satisfy. However, meeting the requirements of the

validation tools do not automatically make our reconstruction good. But, on the contrary, failing to meet these requirements, almost surely raises a warning on the quality of our reconstruction.

The EM old saying can be actually generalized to a wider problem: parameter estimation (“the good thing about parameter estimation is that you always get a value; the bad thing about parameter estimation is that you always get a value”). The whole process of image processing in EM is about finding parameters (whether a micrograph preserves information at high resolution or not, whether this small image is a centered particle or not, whether this particle belongs to this population or not, which is the 3D pose of the particle with respect to this volume. The old EM saying is just a particularization of this more general principle to the parameters describing the Coulomb potential in space of the macromolecule under study. All these parameters are estimated by computational algorithms that, by definition, will have false positives, false negatives, biases, and variances associated to the parameter estimation. This task is further complicated by the fact that the Signal-to-Noise Ratio (SNR) of the measurements is between $1/10$ and $1/100$ (i.e., there is 10–100 times more noise than signal) if we talk about micrographs, or several hundredths of this if we talk about movie frames. In an experimental study, we can never be sure that we have correctly identified all the parameters as the true values are unknown. The most we can do is to estimate the same parameter multiple times (preferably with different algorithms based on completely “orthogonal” mathematical principles), and only trust those estimates that are consistently estimated to similar values. This principle does not preclude bias (the different estimates may consistently point to a wrong answer), but the probability of this event is smaller if multiple estimates point to the same value, especially in the case of the use of several algorithms rather than multiple runs of the same algorithm. In this task, having a platform like Scipion that integrates over 30 packages with more than 300 protocols solving high-level image-processing tasks is not only convenient, but an imperative requirement if we are to combine the results from multiple runs and algorithms without having to deal with all the internal details and conventions of the different software packages.

At present, there seems to be an obsession in reporting structures with a resolution number as low as possible. However, it has been already shown [10, 51] and illustrated in this chapter that the resolution reported by the FSC (even if it is coming from two independent halves of the data) is at the low end of the local resolution range, and consequently, it is not the resolution of the map, but the resolution of the best resolved voxels in the map. Moreover, the FSC can be inadvertently distorted by the choice of the mask used for its calculation. Map-model comparisons [62] are alternative, objective measurements of the quality of the map.

Although they have the drawback that they also depend on the quality of the fitting, so that a good map can be poorly evaluated due to a poor quality fitting. Probably, there does not exist a single measure that can unambiguously determine the quality of a map, and all figures of merit can be fooled by pathological features. Our opinion is that reporting measures like the ones shown along the chapter, specifically targeting the identification of incorrectly estimated parameters, should be promoted as a good practice within the structural biology community. Relying on the algorithm to correctly deal with the miss-estimates is a dangerous practice that should be avoided in a solid scientific work.

Single Particle Analysis by EM has witnessed in the last years an incredible boost in its throughput and in the range of interesting biological problems addressed. In a way, this boost has been promoted by hardware improvements as well as faster, more robust, and easy-to-use algorithms. However, no algorithm is free from errors, particularly so because at this low level of SNR currently the correctness of many estimated parameters is very difficult to assess. As a field, we have given a huge step forward in reproducibility and open science by making public first the reconstructed maps (EMDB, [63]) and more recently the raw data (EMPIAR, [64]). We foresee a near future in which the intermediate steps connecting the raw data and the final map are also made public in such a way that the correctness of the estimated parameters can be verified, and even the data could be reprocessed at those points where the first data analysis is suspected to produce suboptimal parameter estimates.

Acknowledgments

The authors would like to acknowledge economical support from: The Spanish Ministry of Economy and Competitiveness through Grants BIO2016-76400-R(AEI/FEDER, UE), the “Comunidad Autónoma de Madrid” through Grant: S2017/BMD-3817. Instituto de Salud Carlos III through Grant: PT17/0009/0010 (ISCIII-SGEFI / ERDF). European Union (EU) and Horizon 2020 through grants: CORBEL (INFRADEV-1-2014-1, Proposal: 654248) Instruct ULTRA (Proposal: 731005), EOSC Life (Proposal: 824087), HighResCells (Proposal: 810057), IMPaCT (Proposal: 857203), EOSC—Synergy (Proposal: 857647), iNEXT-Discovery (Proposal: 871037), and European Regional Development Fund-Project “CERIT Scientific Cloud” (No. CZ.02.1.01/0.0/0.0/16_013/0001802). The authors acknowledge the support and the use of resources of Instruct, a Landmark ESFRI project.

References

- Benjin X, Ling L (2020) Developments, applications, and prospects of cryo-electron microscopy. *Protein Sci* 29:872–882
- Lyumkis D (2019) Challenges and opportunities in cryo-em single-particle analysis. *J Biol Chem* 294:5181–5197
- Eisenstein M (2018) Drug designers embrace cryo-EM. *Nat Biotechnol* 36:557–558
- Scapin G, Potter CS, Carragher B (2018) Cryo-em for small molecules discovery, design, understanding, and application. *Cell Chem Biol* 25:1318–1325
- Saur M, Hartshorn MJ, Dong J, Reeks J et al (2019) Fragment-based drug discovery using cryo-em Drug discovery today. doi: <https://doi.org/10.1016/j.drudis.2019.12.006>
- Jonic S (2017) Computational methods for analyzing conformational variability of macromolecular complexes from cryo-electron microscopy images. *Curr Opin Struct Biol* 43:114–121
- Sorzano COS, Jiménez A, Mota J, Vilas JL et al (2019) Survey of the analysis of continuous conformational variability of biological macromolecules by electron microscopy. *Acta Crystallogr Sect F, Struct Biol Commun* 75:19–32
- Arnold SA, Müller SA, Schmidli C et al (2018) Miniaturizing EM sample preparation: opportunities, challenges, and “visual proteomics”. *Proteomics* 18:e1700176
- Faruqi AR, McMullan G (2018) Direct imaging detectors for electron microscopy. *Nucl Instrum Methods Phys Res, Sect A* 878:180–190
- Vilas JL, Gómez-Blanco J, Conesa P et al (2018) MonoRes: automatic and unbiased estimation of local resolution for electron microscopy maps. *Structure* 26:337–344
- de la Rosa-Trevín JM, Quintana A, Del Cano L et al (2016) Scipion: a software framework toward integration, reproducibility and validation in 3D electron microscopy. *J Struct Biol* 195:93–99
- Wang Z, Hryc CF, Bammer B et al (2014) An atomic model of brome mosaic virus using direct electron detection and real-space optimization. *Nat Commun* 5:4808
- Heymann JB, Marabini R, Kazemi M et al (2018) The first single particle analysis map challenge: a summary of the assessments. *J Struct Biol* 204:291–300
- Sorzano COS, Vargas J, Oton J et al (2017) A review of resolution measures and related aspects in 3D electron microscopy. *Prog Biophys Mol Biol* 124:1–30
- Vilas JL, Tagare HD, Vargas J et al (2020) Measuring local-directional resolution and local anisotropy in cryo-EM maps. *Nat Commun* 11:55
- Ramírez-Aportela E, Mota J, Conesa P et al (2019) Deep-res: a new deep-learning- and aspect-based local resolution method for electron-microscopy maps. *IUCrj* 6:1054–1063
- Sorzano COS, Fernández-Giménez E, Peredo-Robinson V et al (2018) Blind estimation of DED camera gain in electron microscopy. *J Struct Biol* 203:90–93
- Li X, Mooney P, Zheng S, Booth CR et al (2013) Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* 10:584–590
- Zheng SQ, Palovcak E, Armache JP et al (2017) Motion-cor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* 14:331–332
- Abrishami V, Vargas J, Li X, Cheng Y et al (2015) Alignment of direct detection device micrographs using a robust optical flow approach. *J Struct Biol* 189:163–176
- Tegunov D, Cramer P (2019) Real-time cryo-electron microscopy data preprocessing with warp. *Nat Methods* 16:1146–1152
- de la Rosa-Trevín JM, Otón J, Marabini R et al (2013) Xmipp 3.0: an improved software suite for image processing in electron microscopy. *J Struct Biol* 184(2):321–328
- Sorzano COS, Jonic S, Núñez Ramírez R et al (2007) Fast, robust and accurate determination of transmission electron microscopy contrast transfer function. *J Struct Biol* 160:249–262
- Zhang K (2016) Gctf: real-time ctf determination and correction. *J Struct Biol* 193:1–12
- Rohou A, Grigorieff N (2015) Ctfind4: fast and accurate defocus estimation from electron micrographs. *J Struct Biol* 192:216–221
- Maluenda D, Majtner T, Horvath P et al (2019) Flexible workflows for on-the-fly electron-microscopy single-particle image processing using scipion. *Acta Crystallogr Sect D, Struct Biol* 75:882–894
- Marabini R, Carragher B, Chen S et al (2015) Ctf challenge: result summary. *J Struct Biol* 190:348–359
- Voss NR, Yoshioka CK, Radermacher M et al (2009) Dog picker and tiltpicker: software tools to facilitate particle selection in single

- particle electron microscopy. *J Struct Biol* 166 (2):205–213
29. Scheres SHW (2015) Semi-automated selection of cryo-em particles in relion-1.3. *J Struct Biol* 189:114–122
 30. Abrishami V, Zaldívar-Peraza A, de la Rosa-Trevín JM et al (2013) A pattern matching approach to the automatic selection of particles from low-contrast electron micrographs. *Bioinformatics* 29:2460–2468
 31. Bepler T, Morin A, Rapp M, Brasch J et al (2019) Topaz: a positive-unlabeled convolutional neural network cryoem particle picker that can pick any size and shape particle. *Microsc Microanal* 25:986–987
 32. Wagner T, Merino F, Stabrin M et al (2019) Sphire-cryolo is a fast and accurate fully automated particle picker for cryo-EM. *Commun Biol* 2:218
 33. Sanchez-Garcia R, Segura J, Maluenda D et al (2018) Deep consensus, a deep learning-based approach for particle pruning in cryo-electron microscopy. *IUCrJ* 5:854–865
 34. Sánchez-García R, Segura J, Maluenda D et al (2020) Micrograph cleaner: a python package for cryo-EM micrograph cleaning using deep learning. *bioRxiv*. <https://doi.org/10.1101/677542>
 35. Vargas J, Abrishami V, Marabini R et al (2013) Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. *J Struct Biol* 183:342–353
 36. Punjani A, Brubaker MA, Fleet DJ (2017) Building proteins in a day: efficient 3D molecular structure estimation with electron cryomicroscopy. *IEEE Trans Pattern Anal Mach Intell* 39:706–718
 37. Sorzano COS, Bilbao-Castro JR, Shkolnisky Y et al (2010) A clustering approach to multi-reference alignment of single-particle projections in electron microscopy. *J Struct Biol* 171:197–206
 38. Sorzano COS, Vargas J, de la Rosa-Trevín JM et al (2014) Outlier detection for single particle analysis in electron microscopy. In: *Proc. Intl. Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO*, p 950
 39. Vargas J, Álvarez-Cabrera AL, Marabini R et al (2014) Efficient initial volume determination from electron microscopy images of single particles. *Bioinformatics* 30:2891–2898
 40. Sorzano COS, Vargas J, de la Rosa-Trevín JM et al (2015) A statistical approach to the initial volume problem in single particle analysis by electron microscopy. *J Struct Biol* 189:213–219
 41. Scheres SHW (2012) Relion: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* 180:519–530
 42. Tang G, Peng L, Baldwin PR, Mann DS et al (2007) Eman2: an extensible image processing suite for electron microscopy. *J Struct Biol* 157:38–46
 43. Reboul CF, Eager M, Elmlund D, Elmlund H (2018) Single-particle cryo-EM- improved *ab initio* 3D reconstruction with simple/prime. *Protein Sci* 27:51–61
 44. Sorzano COS, Vargas J, Vilas JL et al (2018) Swarm optimization as a consensus technique for electron microscopy initial volume. *Appl Anal Optim* 2:299–313
 45. Gomez-Blanco J, Kaur S, Ortega J, Vargas J (2019) A robust approach to *ab initio* cryo-electron microscopy initial volume determination. *J Struct Biol* 208:107397
 46. Jimenez A, Jonic S, Majtner T et al (2019) Validation of electron microscopy initial models via small angle x-ray scattering curves. *Bioinformatics* 35:2427–2433
 47. Kimanius D, Forsberg BO, Scheres SH, Lindahl E (2016) Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *elife* 5:e18722
 48. Sorzano COS, Marabini R, Vargas J et al (2014) Computational Methods for Three-Dimensional Microscopy Reconstruction, Springer, chap Interchanging geometry information in electron microscopy single particle analysis: mathematical context for the development of a standard, pp 7–42
 49. Cardone G, Heymann JB, Steven AC (2013) One number does not fit all: mapping local variations in resolution in cryo-em reconstructions. *J Struct Biol* 184:226–236
 50. Kucukelbir A, Sigworth FJ, Tagare HD (2014) Quantifying the local resolution of cryo-EM density maps. *Nat Methods* 11:63–65
 51. Ramírez-Aportela E, Vilas JL, Glukhova A et al (2019) Automatic local resolution-based sharpening of cryo-EM maps. *Bioinformatics* 36:765–772
 52. Fernández JJ, Luque D, Castón JR, Carrascosa JL (2008) Sharpening high resolution information in single particle electron cryomicroscopy. *J Struct Biol* 164(1):170–175
 53. Vilas JL, Vargas J, Martínez M et al (2020b) Re-examining the spectra of macromolecules: current practice of spectral quasi b-factor flattening. *J Struct Biol* 209:107447
 54. Jakobi AJ, Wilmanns M, Sachse C (2017) Model-based local density sharpening of cryo-EM maps. *elife* 6:e27131

55. Naydenova K, Russo CJ (2017) Measuring the effects of particle orientation to improve the efficiency of electron cryomicroscopy. *Nat Commun* 8:629
56. Vargas J, Otón J, Marabini R et al (2016) Particle alignment reliability in single particle electron cryomicroscopy: a general approach. *Sci Rep* 6:21626
57. Vargas J, Melero R, Gómez-Blanco J et al (2017) Quantitative analysis of 3D alignment quality: its impact on soft-validation, particle pruning and homogeneity analysis. *Sci Rep* 7:6307
58. Stagg SM, Noble AJ, Spilman M, Chapman MS (2014) Reslog plots as an empirical metric of the quality of cryo-em reconstructions. *J Struct Biol* 185:418–426
59. Heymann B (2015) Validation of 3dem reconstructions: the phantom in the noise. *AIMS Biophys* 2:21–35
60. Beckers M, Jakobi AJ, Sachse C (2019) Thresholding of cryo-em density maps by false discovery rate control. *IUCrJ* 6(1):18–33
61. Martínez M, Jiménez-Moreno A, Maluenda D et al (2020) Integration of cryo-EM model building software in Scipion. *J Chem Inf Model* 26:2533–2540
62. Afonine PV, Klaholz BP, Moriarty NW et al (2018) New tools for the analysis and validation of cryo-em maps and atomic models. *Acta Crystallogr Sect D, Struct Biol* 74:814–840
63. Patwardhan A (2017) Trends in the electron microscopy data bank (emdb). *Acta Crystallogr Sect D: Struct Biol* 73:503–508
64. Iudin A, Korir PK, Salavert-Torres J et al (2016) Empiar: a public archive for raw electron microscopy image data. *Nat Methods* 13:387–388



Setup and Troubleshooting of Volta Phase Plate Cryo-EM Data Collection

Otilie von Loeffelholz and Bruno P. Klaholz

Abstract

Cryo electron microscopy (cryo-EM) has become a method of choice in structural biology to analyze isolated complexes and cellular structures. This implies adequate imaging of the specimen and advanced image-processing methods to obtain high-resolution 3D reconstructions. The use of a Volta phase plate in cryo-EM drastically increases the image contrast while being able to record images at high acceleration voltage and close to focus, i.e., at conditions where high-resolution information is best preserved. During image processing, higher contrast images can be aligned and classified better than lower quality ones resulting in increased data quality and the need for less data. Here, we give step-by-step guidelines on how to set up high-quality VPP cryo-EM single particle data collections, as exemplified by human ribosome data acquired during a one-day data collection session. Further, we describe specific technical details in image processing that differ from conventional single particle cryo-EM data analysis.

Key words Volta phase plate, Phase contrast, Cryo electron microscopy, Human ribosome, Structural biology

1 Introduction

Visualization of biomolecules relies on good image contrast, which in cryo electron microscopy (cryo-EM) is almost entirely achieved by phase contrast. Macromolecular complexes are weak phase objects resulting in very low contrasted images when they are imaged in focus [1]. Changes in imaging, such as the use of lower voltage [2, 3] and defocusing [4] are traditionally implemented to generate higher contrast. Insertion of a phase plate in the back-focal plane instead of an objective aperture (Fig. 1) also allows imaging at high contrast. It also has the advantage that imaging can be done at high voltage and very close to focus, typically two parameters that help acquiring strong high-frequency components, i.e., the high-resolution data. This can be achieved by an additional negative phase shift between the sample-scattered and the

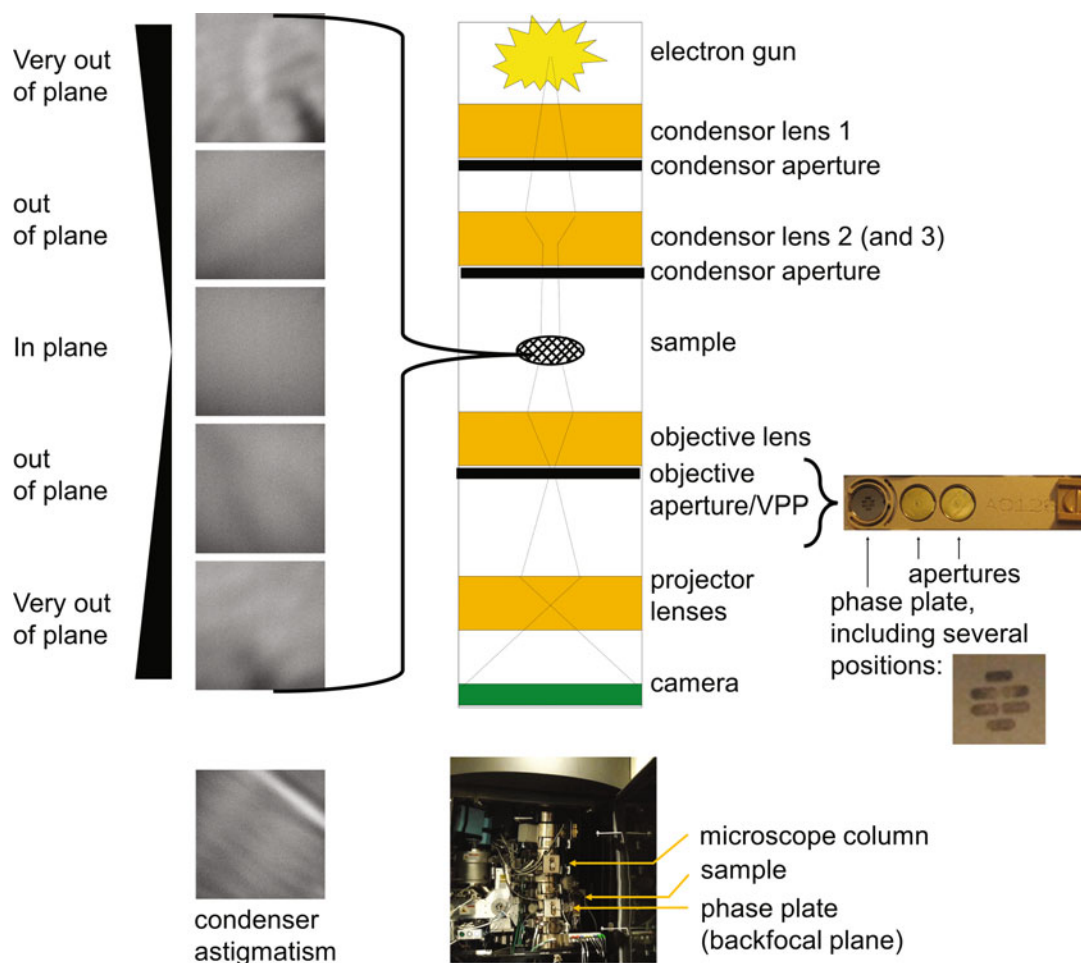


Fig. 1 Placement of the VPP into the back-focal plane. Scheme of an electron microscope including the position of the VPP at the height of the objective apertures. Images show the artifacts visible when the VPP is not in the back-focal plane or when condenser astigmatism is present

unscattered electrons that result in a drastic increase of the sample-scattered image intensities. Various phase plate approaches [5–9] have been tried in the past, but the one that now allows routine usage for data collection both for single particle cryo-EM and cryo electron tomography (cryo-ET) is the Volta phase plate (VPP) [10].

The VPP consists of a thin carbon film placed in the back-focal plane of the optical system; exposure to the electron beam induces a phase shift which increases with the exposure time, i.e., it is an effect building up under cumulative dose [10]. Phase shifts from $\sim 45^\circ$ to 135° have been shown to be optimal for image processing [11], implying that the phase plate position needs to be changed on a regular basis when the phase shift becomes too high (e.g., after

1–2 h or every 100–150 images, depending on imaging settings). The relatively easy handling, reproducibility, and low level of artifacts makes the VPP currently superior to other phase plate developments, such as the Zernike phase plate [6, 12, 13] and electrostatic phase plates [5, 7, 9], but there are nevertheless specific aspects to be taking care of on the electron microscope before starting a data collection (e.g., constant heating temperature, check for low contamination, VPP centering, etc. as detailed below). When a VPP is used, the increase in contrast correlates with a gradually increasing phase shift. After a phase shift of $\sim 90^\circ$ the low-resolution information takes overhand leading to increased image blurring [11]. It needs to be mentioned that the estimation of the contrast transfer function (CTF) is not yet well understood at phase shifts beyond 180° . Therefore, even though it was suggested that high-resolution structures can be obtained from data including high phase shift images [14], it is usually not advised to use images with a phase shift higher than 180° . Additionally, it is advisable to collect VPP cryo-EM data with a small defocus [11] even though in-focus [11] and even over-focus data [15] collections were shown to be usable in image processing leading to structures in the 3 Å range. Defocus VPP data collection is beneficial in terms of (a) automatically setting up the defocus target in the microscope, (b) estimation of the correct CTF, and (c) speed of data collection [11, 16].

The huge advantage of using the VPP for single particle data collection is the high contrast obtained that helps aligning the particles [10, 11, 16] so that structures of proteins down to a size of 50–80 kDa [17–19] can be solved to 3–4 Å resolution by single particle image-processing approaches (provided sample homogeneity and particle distribution are good). It also facilitates 2D particle classifications and 3D classifications of the reconstructions using various structure sorting approaches [20–23], structure refinement with focused classifications and refinements [24, 25], and it facilitates cryo-EM map interpretation and atomic model building [26]. The VPP is a promising tool also for acquiring cryo electron tomograms to increase the visibility of otherwise too noisy samples, which allows better segmentation and interpretation of the data [27, 28].

In this chapter, we describe a hands-on protocol for setting up a successful VPP cryo-EM data collection session including practical hints on how to overcome difficulties that can occur during alignment of the VPP for data collection and for CTF estimation during image data processing.

2 Materials

1. Cryo-EM grids suitable for single particles data collection. These can be, for example, Quantifoil or C-flat grids with regular hole sizes and regular spacing in between them. Irregular grids (e.g., Lacy) are also possible but will reduce throughput and ease of setting up automated data collection.
2. A cryo electron microscope equipped with a VPP, which is inserted instead of an objective aperture. In our case, a Titan Krios electron microscope (Thermo Fisher Scientific) is used.
3. Software for automated data collection such as EPU (Thermo Fisher Scientific) or SerialEM [29].
4. A GPU machine with programs for data processing installed (here: GCTF, Motioncor2, Excel, or OpenOffice).

3 Methods

The VPP is very sensitive mechanically and can easily be contaminated. Therefore, it needs to be removed every time the grid in the microscope is changed. Additionally, to avoid contamination of the VPP during data collection the heating current of the VPP should be set at around 25 mA to achieve a stable temperature of $\sim 200^\circ\text{C}$. The heating device should remain on to maintain a stable temperature; in case of vacuum failure of the microscope, the heating should be switched off to avoid oxidation of the VPP carbon foil.

1. Go to Eucentric height and desired recording magnification and set Eucentric focus. Choose the 50 μm condenser aperture.
2. Perform direct alignments procedure including beam pivot points, beam shift, rotation center on a Cross Grating grid (2160 lines / mm).
3. Move to an empty area on the grid and decide about the spot size used for data collection. Insert the phase plate (instead of an objective aperture). Insert the fluorescent screen.
4. Bring the phase plate in the back-focal plane. Therefore, tick the box “MF-Y fine focus back-focal plane” in the phase plate tab and turn the MF-Y button until a minimum of artifacts are seen, e.g., stripy or cloud-like shades (close to the back-focal plane) or white or black contamination spots (far from the back-focal plane; Fig. 1, see **Note 1**).
5. Additional condenser astigmatism (Fig. 1) can be removed by ticking the condenser box in the apertures tab additionally to the “MF-Y fine focus back-focal plane” box and turning MF-X and MF-Y buttons.

6. Go to the Alignments tab and choose “Align PhasePlate” and then “Phase Plate μ P: microprobe mode”. Perform all steps by following all the instructions (*see* **Note 2**). Then move on to “Phase Plate nP: nanoprobe mode” and perform all steps following the instructions. This procedure sets the diffraction lens value that focuses on the phase plate and also sets accurate beam shift pivot points on the phase plate (*see* **Note 3**), so that the same position of the VPP is exposed in focus and exposure settings, which is important to keep the phase shift relatively similar. The alignments should be repeated until the settings do not need to be changed anymore.
7. Retract the phase plate.
8. If present, perform a full alignment of the energy filter in suitable software.
9. In case the microscope contains a Cs corrector, go to Eucentric height on a flat area of the grid and use the image corrector software to align the corrector.
10. If necessary, realign the rotation center again.
11. Acquire Gain and Dark references on an empty grid square.
12. Change the grid to the specimen grid and acquire a grid map in low magnification.
13. Choose suitable areas for data collection in the grid map, ensure the image shift is minimal, and acquire medium magnification images of the grid squares.
14. Select suitable areas for data collection, decide about the maximum number of images to be acquired per hole, and define focusing area.
15. Go to an empty area on the grid; choose the spot size for imaging. Decide about the target dose and fractionation scheme, insert the phase plate, and repeat **steps 3** and **4** immediately before starting the data collection.
16. Take an image in record settings to ensure that there are no artifacts created by the VPP (these can be recognized from shades over the image of an empty grid square (Fig. 1) or uneven contrast of particles inside the image (Fig. 2d, e).
17. Move to an area with carbon and correct the astigmatism induced by the phase plate with the objective stigmator.
18. Move to the next position of the VPP.
19. Start automated data collection with only one fixed defocus (e.g., -500 nm) or a small target defocus range (differing only by -200 - -600 nm) and focusing at every stage position (*see* **Notes 4** and **5**). Here, the difference between the set and the actual defocus taken as well as the point spread function at higher defocus values should be taken into account when setting the target defocus.

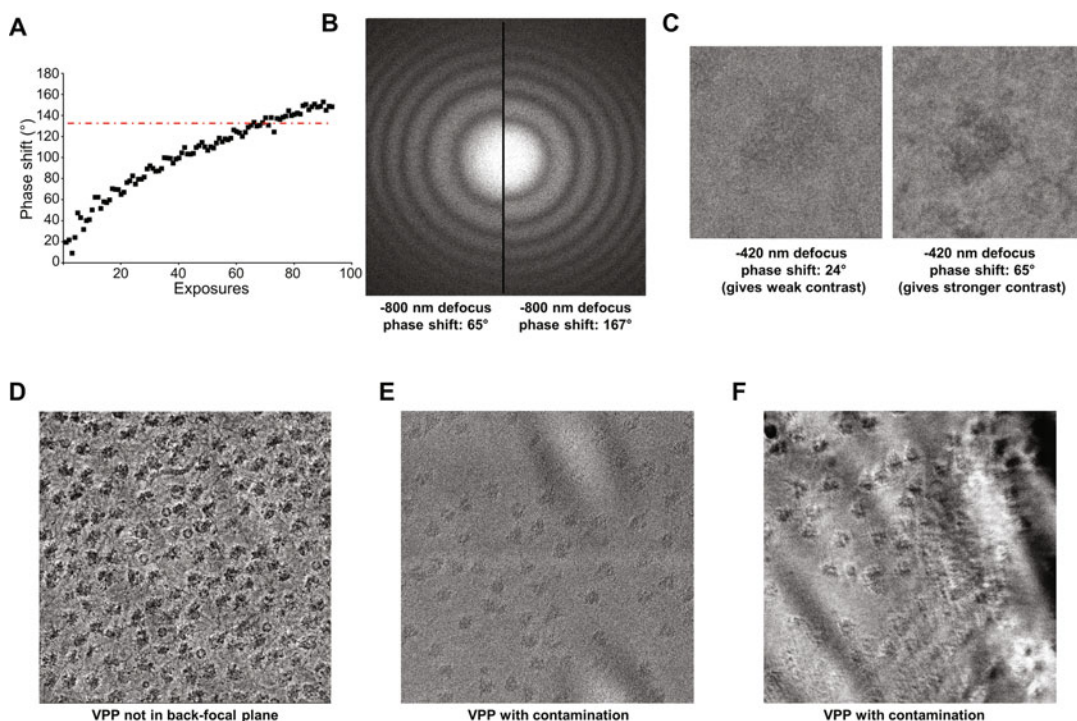


Fig. 2 Monitoring of the phase shift during VPP cryo-EM data acquisition. **(a)** Phase shift development over time/exposure dose on a single phase plate position. **(b)** Zoomed power spectrum halves of images with a low phase shift (left) and a high phase shift (right) at the same defocus. **(c)** Ribosome images acquired at a similar defocus with the same dose of $30 \text{ e}/\text{\AA}^2$ containing a low phase shift (left) and a high phase shift (right). **(d)** Shadows appearing in a cryo-EM image of human ribosomes collected with a VPP that was not inserted well in the back-focal plane **(e, f)** Cryo-EM image of human ribosomes collected with a contaminated VPP

20. Align movie frames, e.g., with MotionCor2 program [30, 31], WARP [32], or other software.
21. Estimate the CTF of the micrographs (e.g., Gctf [33], CTFind [34], WARP [32]). Therefore, the approximate target defocus range (e.g., for -500 nm set target defocus, search between -200 and -1200 nm) and phase shift (e.g., 10° – 180°) should be set in the search to restrict the program search, which will avoid misinterpretations of the program. Images taken with high phase shift can easily be interpreted as higher defocus images instead (Fig. 2b, c).
22. Plot the estimated phase shifts of the aligned movies (e.g., with Microsoft Excel) in the order that the movies were taken. For each phase plate position, the phase shift should build up gradually until it reaches a plateau as shown in Fig. 2a (see **Note 6**). This can be done after the collection of ~ 100 images. The change of the phase plate position should be set to a time/number of exposures that corresponds to the time needed for the phase shift to reach a plateau at around 140° (so that the

phase shift range of the data lies between 90° and 135°). Here, the exact number of degrees is less important than the quality of the overall aligned images and the plateau-behavior of the phase plate (see **Notes 7–9**). The change of the phase plate position can be done automatically by an independent auto-clicking program (e.g., ClickWhen) and/or inside the data collection program (e.g., EPU (Thermo Fisher Scientific)). See also **Notes 10 and 11**.

4 Notes

1. The condenser aperture can be enlarged to $100\ \mu\text{m}$, if it is hard to see any features on the fluorescent screen. Additionally, the magnification can be increased momentarily.
2. During the phase plate alignment both modes, microprobe and nanoprobe, need to be aligned even if only one of the modes is used for data collection.
3. It may happen that this procedure changes the Diffraction Astigmatism. Therefore, one can reset the diffraction astigmatism in Diffraction mode (VPP retracted because of the strong crossover beam to protect the VPP) using the “Diffraction” box in the “Stigmator” tab.
4. Since data collection is rather close to focus, it is important to focus precisely to avoid significant data loss due to in-focus or over-focused images.
5. It should be ensured that Thon rings are clearly visible in the Power spectra. If the signal of the images is so low that at $-500\ \text{nm}$ defocus only one or no Thon rings are visible, the target defocus should be increased, provided the ice thickness is suitable for high resolution work.
6. If the phase shift plot does not show a gradual increase of phase shift over time/exposure dose, it may indicate that either the phase shift estimation is not correct, the beam shift pivot points on the VPP are not correctly set or the VPP is not in the back-focal plane; see steps described above.
7. Phase shifts around 135° – 180° are more difficult to process because of image blurring. Phase shifts beyond 180° can currently not be used for CTF correction. Also certain error proneness in the correctness of the programs reporting the phase shift needs to be considered. Therefore, it is advisable to consider the behavior of the phase shift over time (Fig. 2a) rather than the phase shift degree values.
8. The accumulative dose for phase shift activation can change during aging of the VPP. Therefore, it is advisable to monitor

the phase shift behavior during each data collection and over time.

9. For tomography, data acquisition with a gradual phase shift increase is inconvenient. Therefore, the phase plate should be activated to a phase shift of $\sim \frac{1}{2} \pi$ (90°) before the tilt series data collection is started (i.e., close to the above described plateau; this helps to have similar phase shift values within a tilt series).
10. Occasionally, small contaminations on the VPP induce a non-equal phase shift inside an image. These contaminations can be permanent or not (some may disappear during exposure to the electron beam).
11. The astigmatism induced by the VPP could vary in some pre-set positions on the VPP. It may be possible to exclude VPP positions with high astigmatism by scripting (Thermo Fisher Scientific or SerialEM [29] before data collection.

Acknowledgments

We thank Jonathan Michalon, Mathieu Schaeffer, Remy Fritz, and Romaric David for IT support. This work was supported by CNRS, Association pour la Recherche sur le Cancer (ARC), Institut National du Cancer (INCa), the Fondation pour la Recherche Médicale (FRM), Ligue nationale contre le cancer (Ligue), Agence National pour la Recherche (ANR), and USIAS (USIAS-2018-012). The electron microscope facility was supported by the Alsace Region, FRM, Inserm, CNRS and ARC, the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05-01, by Instruct-ERIC and Instruct-ULTRA (Coordination and Support Action Number ID 731005) funded by the EU H2020 framework to further develop the services of Instruct-ERIC.

References

1. Zernike F (1955) How I discovered phase contrast. *Science* 121:345–349
2. Herzik MA, Wu M, Lander GC (2019) High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nat Commun* 10:1032
3. Orlov I, Rochel N, Moras D et al (2012) Structure of the full human RXR/VDR nuclear receptor heterodimer complex with its DR3 target DNA. *EMBO J* 31:291–300
4. Dubochet J, Adrian M, Chang JJ et al (1988) Cryo-electron microscopy of vitrified specimens. *Q Rev Biophys* 21:129–228
5. Cambie R, Downing KH, Typke D et al (2007) Design of a microfabricated, two-electrode phase-contrast element suitable for electron microscopy. *Ultramicroscopy* 107:329–339
6. Danev R, Nagayama K (2001) Transmission electron microscopy with Zernike phase plate. *Ultramicroscopy* 88:243–252
7. Frindt N, Oster M, Hettler S et al (2014) In-focus electrostatic Zach phase plate imaging for transmission electron microscopy with tunable phase contrast of frozen hydrated biological samples. *Microsc Microanal* 20:175–183

8. Majorovits E, Barton B, Schultheiss K et al (2007) Optimizing phase contrast in transmission electron microscopy with an electrostatic (Boersch) phase plate. *Ultramicroscopy* 107:213–226
9. Walter A, Steltenkamp S, Schmitz S et al (2015) Towards an optimum design for electrostatic phase plates. *Ultramicroscopy* 153:22–31
10. Danev R, Buijsse B, Khoshouei M et al (2014) Volta potential phase plate for in-focus phase contrast transmission electron microscopy. *Proc Natl Acad Sci U S A* 111:15635–15640
11. Danev R, Tegunov D, Baumeister W (2017) Using the Volta phase plate with defocus for cryo-EM single particle analysis. *eLife* 6: e23006
12. Dai W, Fu C, Raytcheva D et al (2013) Visualizing virus assembly intermediates inside marine cyanobacteria. *Nature* 502:707–710
13. Murata K, Liu X, Danev R et al (2010) Zernike phase contrast cryo-electron microscopy and tomography for structure determination at nanometer and sub-nanometer resolutions. *Structure* 1993(18):903–912
14. Li K, Sun C, Klose T et al (2019) Sub-3 Å apoferritin structure determined with full range of phase shifts using a single position of Volta phase plate. *J Struct Biol* 206:225–232
15. Fan X, Zhao L, Liu C et al (2017) Near-atomic resolution structure determination in over-focus with Volta phase plate by Cs-corrected Cryo-EM. *Structure* 25:1623–1630. e3
16. von Loeffelholz O, Papai G, Danev R et al (2018) Volta phase plate data collection facilitates image processing and cryo-EM structure determination. *J Struct Biol* 202:191–199
17. Fan X, Wang J, Zhang X et al (2019) Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Å resolution. *Nat Commun* 10:2386
18. Hill CH, Boreikaitė V, Kumar A et al (2019) Activation of the endonuclease that defines mRNA 3' ends requires incorporation into an 8-subunit Core cleavage and polyadenylation factor complex. *Mol Cell* 73:1217–1231
19. Khoshouei M, Radjainia M, Baumeister W et al (2017) Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nat Commun* 8:16099
20. Abeyrathne PD, Koh CS, Grant T et al (2016) Ensemble cryo-EM uncovers inchworm-like translocation of a viral IRES through the ribosome. *eLife* 5:e14874
21. Klaholz BP, Myasnikov AG, van Heel M (2004) Visualization of release factor 3 on the ribosome during termination of protein synthesis. *Nature* 427:862–865
22. Penczek PA, Frank J, Spahn CM (2006) A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. *J Struct Biol* 154:184–194
23. Scheres SH (2010) Classification of structural heterogeneity by maximum-likelihood methods. *Methods Enzymol* 482:295–320
24. von Loeffelholz O, Natchiar SK, Djabeur N et al (2017) Focused classification and refinement in high-resolution cryo-EM structural analysis of ribosome complexes. *Curr Opin Struct Biol* 46:140–148
25. Zivanov J, Nakane T, Forsberg BO et al (2018) New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* 7:e42166
26. Afonine PV, Klaholz BP, Moriarty NW et al (2018) New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallogr Sect Struct Biol* 74:814–840
27. Khoshouei M, Pfeffer S, Baumeister W et al (2017) Subtomogram analysis using the Volta phase plate. *J Struct Biol* 197:94–101
28. Rast A, Schaffer M, Albert S et al (2019) Biogenic regions of cyanobacterial thylakoids form contact sites with the plasma membrane. *Nat Plants* 5:436
29. Mastronarde DN (2005) Automated electron microscope tomography using robust prediction of specimen movements. *J Struct Biol* 152:36–51
30. Li X, Mooney P, Zheng S et al (2013) Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* 10:584–590
31. Zheng SQ, Palovcak E, Armache JP et al (2017) MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* 14:331–332
32. Tegunov D, Cramer P (2019) Real-time cryo-EM data pre-processing with Warp. *Nat Methods* 16:1146–1152
33. Zhang K (2016) Gctf: real-time CTF determination and correction. *J Struct Biol* 193:1–12
34. Rohou A, Grigorieff N (2015) CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J Struct Biol* 192:216–221



Cryo-Focused Ion Beam Lamella Preparation Protocol for in Situ Structural Biology

Jana Moravcová, Radka Dopitová, Matyáš Pinkas, and Jiří Nováček

Abstract

The advances in electron cryo-microscopy have enabled high-resolution structural studies of vitrified macromolecular complexes in situ by cryo-electron tomography (cryo-ET). Since utilization of cryo-ET is generally limited to the specimens with thickness < 500 nm, a complex sample preparation protocol to study larger samples such as single eukaryotic cells by cryo-ET was developed and optimized over the last decade. The workflow is based on the preparation of a thin cellular lamella by cryo-focused ion beam milling (cryo-FIBM) from the vitrified cells. The sample preparation protocol is a multi-step process which includes utilization of several high-end instruments and comprises sample manipulation prone to sample deterioration. Here, we present a workflow for preparation of three different model specimens that was optimized to provide high-quality lamellae for cryo-ET or electron diffraction tomography with high reproducibility. Preparation of lamellae from large adherent mammalian cells, small suspension eukaryotic cell line, and protein crystals of intermediate size is described which represents examples of the most frequently studied samples used for cryo-FIBM in life sciences.

Key words Cryo-focused ion beam milling, Cryo-electron microscopy, Lamella, Adherent cells, Protein crystal, *Saccharomyces cerevisiae*

1 Introduction

Electron cryo-microscopy (cryo-EM) has been experiencing significant expansion in Structural and Cellular Biology research. This is mainly due to the technological developments in the field of electron detection [1], the development of the electron microscopes capable of unsupervised automated data collection, and a significant progress in the development of the software for the data analysis [2]. Apart from the single particle cryo-EM that is now a well-established technique in Structural Biology, the cryo-electron tomography (cryo-ET) is developing as a method for studies of pleomorphic samples or structural studies of macromolecular complexes in situ [3]. Cryo-ET can reach single nanometer resolution data when applied to pleomorphic objects due to low sensitivity

obtained from a single tomogram, while near-atomic resolution can be obtained through sub-volume averaging of multiple objects with the same structure [4, 5]. One of the major applications of cryo-ET is the structural characterization of the macromolecular complexes in the context of other proteins and nucleic acids inside the cell. Utilization of electron tomography with the microscope set to diffraction mode opens a new application area for the structure determination from single crystals [6, 7]. When applied to protein or small organic compound crystals with thickness of ~100 nm, the technique is termed micro-ED and is nowadays rapidly developing method with significant potential in Structural Biology research.

All transmission electron microscopy (TEM) methods are limited by the sample thickness. In general, this is not a problem for single particle cryo-EM where the sample preparation can eventually be optimized to vitrify the molecules of interest into thin layers. However, more complex biological samples, such as the whole cells, are in fact impassable to 300 keV electron beams prohibiting the acquisition of cryo-ET data. Similarly, micrometer-sized crystals are non-transparent for 200 or 300 keV electrons which significantly limits the utilization potential of micro-ED. In such cases, the sample must be first thinned down to render it electron transparent for the TEM imaging while maintaining it in the vitreous state. Over the last decade, two approaches have been developed to manage TEM imaging of thick vitrified specimens. Firstly, cryo-ultramicrotomy, which utilizes mechanical slicing of the sample with a diamond knife to prepare 60–80 nm cross-sections from cells and tissues [8–10]. Unfortunately, the technique can result in a number of artifacts, including curved sections, crevasses, and sample compression [9, 11]. Furthermore, the sections prepared by using a cryo-ultramicrotome poorly attach to TEM grids [12]. Even though the sample is maintained in the vitreous state, the presence of the artifacts from the sample preparation may significantly limit the interpretability of the cryo-ET data. The second approach uses a cryo-focused ion beam to mill thin cross-sections from vitrified cells or protein crystals (cryo-FIBM) and is now the method of choice [13, 14]. The frozen hydrated samples are thinned by a focused ion beam (FIB) of Ga^+ in a multi-step process to ablate large volume of biological material down to 80–300 nm thin lamellae. A single lamella is obtained for each cell, representing ~0.3–3% of its volume, and is almost void of any sample preparation artifacts and thus is suitable for high-resolution cryo-ET. In addition, the sample is retained on the TEM grid during the whole process, which significantly facilitates the sample handling. FIB is typically combined with scanning electron microscopy (SEM) in a dual beam system, which enables simultaneous milling of the lamella with FIB and the visual inspection of the process by SEM [3]. Finally, the grid with several lamellas (typically 4–8) is transferred to TEM for cryo-ET data collection.

The whole cryo-FIBM workflow combines several instruments and requires transfer between them. Each of the steps comprises a risk of the sample damage, which eventually reduces the probability of obtaining high-quality lamella for cryo-ET downstream the sample preparation process. This is the major cause for the relatively low throughput of the cryo-FIBM workflow.

Here, we provide an optimized protocol for the lamella preparation which comprises all the steps from cultivation of the sample culture (or sample crystallization) through to the insertion of the sample into cryo-TEM for three different model specimens (Fig. 1): (1) A9 adherent mammalian cells, (2) *Saccharomyces cerevisiae* in suspension culture and (3) proteinase K crystals. The A9 cells represent large eukaryotic cells that adhere onto the surface of the TEM grid and where a single lamella is prepared from each cell. The yeast cells represent small suspension cells which are applied to the TEM grid before vitrification, and where a single lamella is usually milled over multiple cells. The proteinase K crystals are examples of medium-sized protein crystals that are too large for electron diffraction tomography without any further processing.

2 Materials

2.1 Preparation and Maintaining of the Mammalian Adherent Cell Culture

Essential requirement for a cell culture work is a sterile workspace for handling, incubation, and storage of the cell culture, reagents, and media. Isolated cell culture laboratory or designated workspace with cell culture hood is sufficient.

All solutions and equipment that are in contact with the cells must be sterile to avoid microbial contamination of the cell culture.

1. Biological material: *Mus musculus* A9 (APRT and HPRT negative derivative of Strain L) (ATCC[®] CCL-1.4[™]).
2. Dulbecco's Modified Eagle's Medium (DMEM)—high glucose.
3. Fetal bovine serum (FBS)—inactivated by heating for 30 min at 56 °C.
4. 0.025% Trypsin, 0.53 mM EDTA dissociation solution.
5. 70% ethanol disinfection solution.
6. Dulbecco's phosphate buffer saline (DPBS) without calcium chloride and magnesium chloride.
7. Plastic Petri dishes for cell culture (ø10 cm).
8. Cell culture flasks (25 cm³).
9. Micropipettes and sterile filter tips.
10. Sterile microtubes.
11. Serological pipette controller.

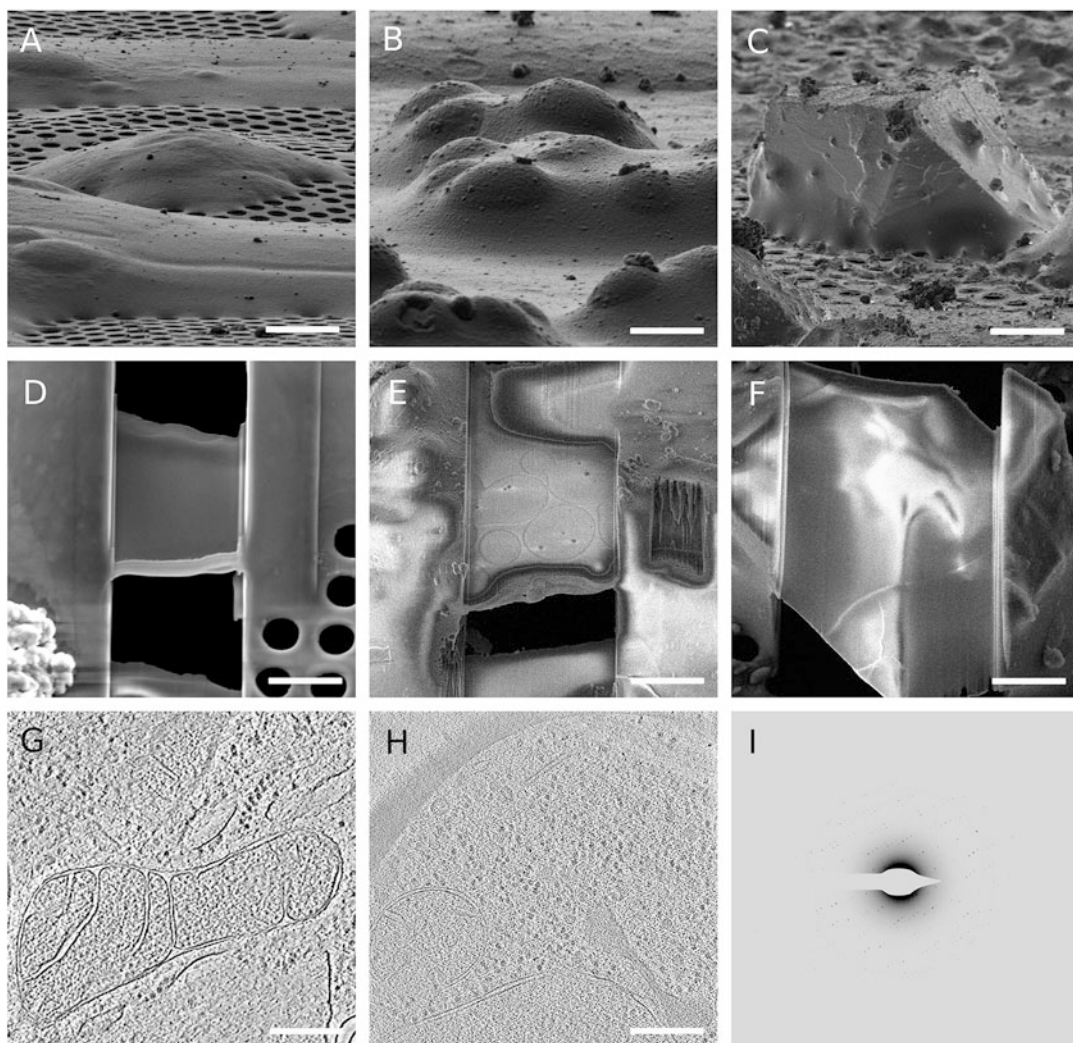


Fig. 1 FIB image of three model samples vitrified on TEM grids. **(a)** adherent mammalian cell, **(b)** *S. cerevisiae* cell cluster, **(c)** and proteinase K crystal, **(d)** SEM images of lamella milled from A9 cell, **(e)** multicellular lamella from *S. cerevisiae* cell cluster, **(f)** proteinase K single crystal lamella, **(g)** section from the tomograms reconstructed from the data collected on the A9 cell, and **(h)** *S. cerevisiae*, **(i)** electron diffraction tomography data collected on lamella of proteinase K crystal

12. Disposable serological pipettes.
13. CO₂-humidified incubator.
14. Inverted optical light microscope.
15. Water bath.
16. Automated cell counter with cell counting slides or Bürker counting slide chamber.
17. Laminar flow hood (Class II Biological Safety Cabinet).
18. Disposable gloves.

2.2 Preparation and Maintaining of *Saccharomyces cerevisiae* Cell Culture

1. *Saccharomyces cerevisiae* cell culture (strain BY4741) (ATCC[®] 201388[™]).
2. Liquid culture medium: Yeast extract (1.1% w/v) and Peptone (2.2% w/v).
3. Glucose (20% w/v).
4. Agar plates.
5. Spreading sticks.
6. Shaking incubator.
7. UV/VIS Spectrophotometer.
8. Sterile plastic Petri dishes.
9. Micropipettes and sterile filter tips.
10. Analytical balances.
11. Disposable gloves.
12. 100–500 ml glass bottles.
13. 100 ml Erlenmeyer flasks.
14. Laminar flow hood.

2.3 Protein Crystallization

1. Emerald BioStructures Combi Clover Crystallization Plate[™] (EBS plate) for sitting drop experiments.
2. 50 mM Tris-HCl pH 8.
3. 1.25 M Ammonium sulfate.
4. Proteinase K (lyophilized powder of recombinant Proteinase K, Roche).

2.4 Preparation and Vitrification of TEM Grids

1. Plasma cleaner.
2. Tweezers and precision tweezers.
3. Glass microscopy slides.
4. Quantifoil R2/1 or R2/4200 Mesh Au or Cu TEM grids.
5. 16-well chamber slide system (Lab Tek system or similar).
6. Vitrobot Mark IV (Thermo Scientific).
7. Filter papers.
8. PTFE (teflon) or other suitable non-adsorbing material.
9. Liquid nitrogen (LN₂).
10. Ethane gas.
11. Cryo-grid boxes.
12. Grid box openers.
13. LN₂ Dewar bottles.
14. Protective equipment for work with LN₂ and ethane (glass, face shield, cold-resistant gloves).

2.5 Mounting TEM Grids into the AutoGrid

1. AutoGrid assembly workstation.
2. AutoGrid box/boxes.
3. AutoGrids.
4. C-clips.
5. C-clip insertion tools.
6. AutoGrid tweezers.
7. Grid box openers.
8. Precision tweezers.
9. Dewar with fresh LN₂.
10. Heating plate with air flow or small oven drying of tools.

3 Methods

3.1 Preparation and Maintaining of the Mammalian Adherent Cell Culture

Following protocol is optimized for cultivation and maintaining of the mammalian adherent A9 cell line. In case of handling different cell lines, we recommend to follow the supplier instructions.

Keep standards of sterile work practice, wear gloves, use 70% ethanol for disinfection, and work in the laminar flow hood.

1. Pre-warm cultivation medium, FBS, trypsin-EDTA solution to 37 °C in water bath.
2. Switch on the laminar flow hood and disinfect the working surface with 70% ethanol.
3. Spray all media, reagent bottles and equipment with 70% ethanol before placing in the flow hood.
4. Complete the cultivation medium if needed (by supplementing DMEM with FBS—10% w/w) (*see Note 1*).
5. Take out A9 stock aliquot from the LN₂ storage Dewar and quickly thaw the cells in 37 °C water bath.
6. Spray the vial with 70% ethanol before placing in the flow hood.
7. Transfer the cell suspension into the cell culture flask, add 10 ml of the cultivation medium, and mix slowly.
8. Incubate the freshly seeded cell culture in a 37 °C, 5% CO₂-humidified incubator.
9. Check the state of the cell culture daily in inverted optical light microscope. Monitor health, growth rate, and cell confluence (*see Note 2*).
10. When the cells adhere to the surface of the culture flask, pipette out the cultivation medium and wash once with room temperature DPBS, add 10 ml of pre-warmed cultivation medium and return the flask back to the incubator (*see Note 3*).

11. Once the cells are ~80% confluent (*see Note 4*), transfer the flask with cell culture from the incubator into the flow hood, pipette out the cultivation medium using a sterile serological pipette.
12. Gently wash the adhered cells twice with room temperature DPBS (*see Note 5*).
13. Remove DPBS and add 1 ml of pre-warmed trypsin-EDTA solution (*see Note 6*) to the cell monolayer. Place the cell culture flask into the incubator and leave the cells to detach from the surface into solution for 10 min (*see Note 7*). The dissociation process can be observed in an inverted light optical microscope. Trypsinization is complete, when the cells are in the suspension and are rounded in shape.
14. Once most of the cells detach from the surface, place the flask back to the flow hood, inhibit the trypsin-EDTA solution by addition of 2 ml of pre-warmed complete cultivation medium, and gently resuspend the cells using serological pipette.
15. Pipette the cell culture into a microtube and determine the cell number per 1 ml using automated cell counter (or Bürker counting slide chamber).
16. Add $2\text{--}3 \times 10^6$ cells to 10 ml of the cultivation medium in a new Petri dish and return the cell culture into the incubator. Label the Petri dish with important information, e.g., cell line, passage number, date, etc. From now periodically passage the cells every 2 or 3 days until 30 passages are reaching (*see Note 8*).

3.2 Preparation of TEM Grids with Mammalian Adherent Cells

1. Place holey carbon 200-mesh Au grids (Quantifoil Au, 200-mesh, R2/1 or R2/4) on a glass slide facing carbon side up (*see Note 9*) and glow discharge them for 15 s (6–9 Pa pressure, 7 mA current). Store the slide with the grids in a Petri dish until applying the cell culture (the grids should be used within 1 h after plasma cleaning).
2. Transfer the glow-discharged grids into the flow hood for 15 min sterilization by UV.
3. Adjust a chamber slide (Fig. 2a) by removing the middle part of the chamber slide. Keep only bottom and top part of the chamber slide (Fig. 2b).
4. Place the grids into individual wells facing carbon side up with the tweezers (*see Note 9*).
5. Add 80 μL of the complete cultivation medium per each grid, cover the chamber slide, place it in the Petri dish, and incubate the grid with the medium for approximately 1 h (*see Note 10*).
6. Prepare the cells according to the points 1–14 of Subheading 2.2, **item 1** of this protocol.

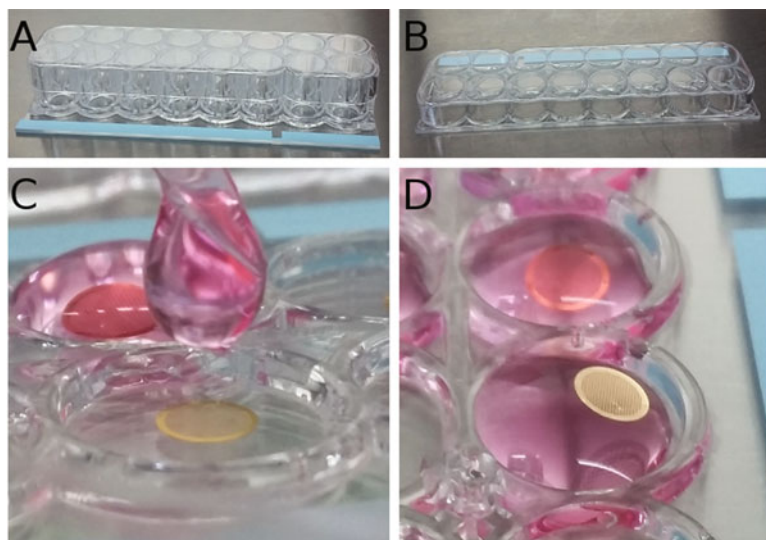


Fig. 2 (a) chamber slide system for adhesion of cells onto TEM grid, (b) bottom and top part of the chamber slide system used for cultivation of cells on TEM grids, (c) correct way for pipetting the cell culture onto the TEM grid, (d) grid position upside down on the drop of cell culture after incorrect application of the cells onto the grid

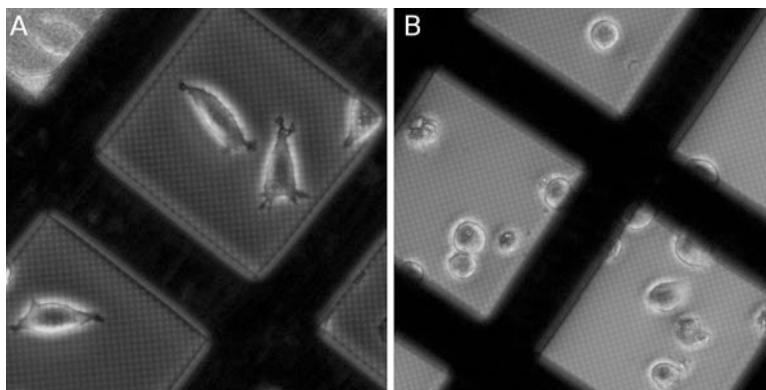


Fig. 3 (a) A9 cells that have correctly adhered onto the TEM grid, (b) dead cells on the grid

7. Dilute the cell suspension to the concentration of $1\text{--}1.5 \times 10^5$ cells/ml (see **Note 11**) using pre-warmed cultivation medium. Remove the cultivation medium from the wells with the pipette. Apply 80 μl of diluted cell suspension per one grid (see **Notes 12** and **13**). Cover the chamber slide and place it in a Petri dish. Allow the cells adhere to the surface of the grid overnight ($\sim 12\text{--}20$ h) in a 37°C , 5% CO_2 -humidified incubator.

8. Next day, check the status of the cells under the light microscope. If the cells adhere to the grid foil (*see* **Note 14**, Fig. 3a), continue with the plunge freezing.

3.3 Vitrification of the Mammalian Adherent Cells

Sterile conditions are no longer required. It is highly recommended to wear protective face mask or shield during vitrification procedure to prevent contamination from breathing to the specimen. Use dry tools to avoid accumulation of ice contamination.

1. Set the Vitrobot to following parameters: humidity: 100%, temperature: 4 °C, blot force: −5, blot time: 5 s, 2 blotting cycles (*see* **Note 15**), mount filter papers to both Vitrobot blotting pads.
2. Prepare liquid ethane for vitrification.
3. Wash the grid with cells twice using DPBS.
4. Mount the grid into the Vitrobot tweezers, blot the grid, and plunge immediately into liquid ethane (*see* **Note 14**).
5. Store the grid in the sealed cryo-grid box under LN₂ or directly clip the grid into the AutoGrid cartridge for loading into the FIB-SEM microscope (*see* **Note 15**).

3.4 Cultivation of *Saccharomyces* *cerevisiae* Suspension Cell Culture

The protocol is optimized for preparation of the sample for cryo-FIBM from suspension cell line *Saccharomyces cerevisiae* strain BY4741 [ATCC 4040002] or similar strains.

1. Take out new agar plate from 4 °C storage.
2. Take the *S. cerevisiae* glycerol stock from −80 °C deep freezer and place it into the freezing stand to avoid complete thawing of the stock.
3. Scrape off and transfer small culture with sterile inoculation loop (1–10 µl) to 50 µl of growth medium and mix thoroughly.
4. Transfer whole volume of mixed *S. cerevisiae* culture and disperse with sterile spreading stick over the surface of agar plate.
5. Incubate at 30 °C for approximately 48 h until 1.5–2 mm diameter colonies are formed.
6. Autoclave 50 ml Erlenmeyer (or similar) flask.
7. Work in hood or laminar flow hood.
8. Pipette 10 ml of the growth medium to sterile 50 ml Erlenmeyer flask.
9. Supplement the medium with 1 ml of filtered 20% glucose.
10. Pick one colony of yeast from agar plate with sterile, disposable inoculation loop (1–10 µl).
11. Place the Erlenmeyer flask in the incubator and culture at 30 °C with agitation (150–200 rpm) until exponential phase is reaching (approximately 7–15 h) (*see* **Note 16**).

12. Measure optical density (OD) of *S. cerevisiae* suspension culture at 600 nm using UV/VIS spectrophotometer.
13. OD₆₀₀ of *S. cerevisiae* culture should be in the range of 1–5 in the exponential phase of cell growth.
14. For EM grids preparation, dilute the cell suspension in growth medium to have final OD₆₀₀ = 1 (*see Note 17*).

3.5 Vitrification of *Saccharomyces cerevisiae* Cells on EM Grids

1. Glow discharge 200 mesh Cu or Au holey carbon grids (e.g., Quantifoil, Cu, 200 mesh, R2/1) on the glass slide facing carbon side up for 30–45 s (pressure: 6–9 Pa, current: 7 mA).
2. Set Vitrobot or another plunge freezing device to following parameters: temperature: 18 °C, humidity: 100%, blot time: 6 s, wait time: 5 s, blotting cycle: 1x, blot force: 5. (*see Note 15*).
3. Prepare liquid ethane for the sample vitrification.
4. Mount PTFE or different non-absorbent surface pad (0.2 mm thick) to the Vitrobot blotting pad facing the sample, use the filter paper for the other blotting pad. (*see Note 18*).
5. Pick the glow-discharged grid with the Vitrobot tweezers and mount it to the instrument. Apply 3.5 µl of *S. cerevisiae* suspension to the carbon side of the grid inside the Vitrobot climate chamber. (*see Note 19*, Fig. 4).
6. Plunge freeze the grid into the liquid ethane (*see Note 22*).
7. Store grids with vitrified cells under LN2 conditions or mount it into the AutoGrid cartridge for loading into the FIB-SEM microscope.

3.6 Preparation of Protein Crystal Samples for Cryo-FIB Workflow

Work in 4 °C cold room to prevent evaporation of proteinase K solution.

1. Dissolve 60 mg proteinase K powder in 1 ml of 50 mM Tris-HCl buffer, pH 8.
Mix well by pipetting.
2. Pipette 200 µl of 1.25 M Ammonium sulfate as a precipitating agent to reservoir of crystallization plate.
3. Pipette 2 µl of 1.25 M Ammonium sulfate to small well for seeding drop in crystallization plate.
4. Add 2 µl of 60 mg/ml solution of proteinase K to small well for seeding drop and mix properly by pipetting.
5. Cover wells with transparent adhesive tape to prevent evaporation of solutions and maintain balance between mixture.
6. Incubate at ambient temperature (21–23 °C) for 2–24 h to form proteinase K crystals in the size range of 10–100 µm (*see Notes 20 and 21*).

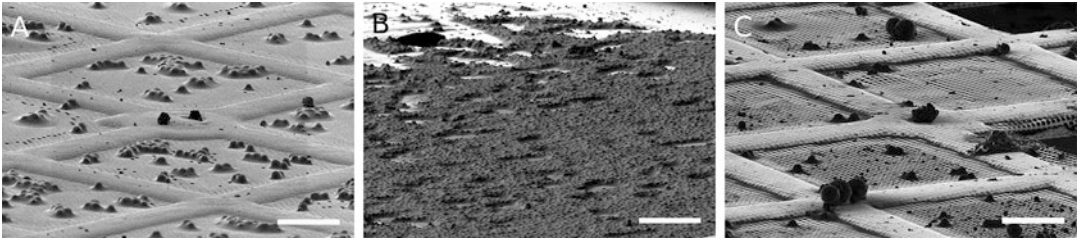


Fig. 4 Different density of *S. cerevisiae* cells plunge frozen on TEM grids. (a) Cell clusters sufficient for multicellular lamella preparation; (b) Cellular monolayer suitable for lamella preparation although proper vitrification of the cells and residual buffer around cells may not be attained; (c) Too low concentration of cells on the grid which does not allow to mill lamella. The scale bars correspond to 30 μm in **a**, 60 μm in **b**, and 30 μm in panel **c**

3.7 Vitrification of Proteinase K Crystals on EM Grids

1. Glow discharge 200 mesh Cu or Au holey carbon grids (e.g., Quantifoil, Cu, 200 mesh, R2/1) facing carbon side up for 30–45 s (pressure: 6–9 Pa, current: 7 mA).
2. Prepare liquid ethane.
3. Mix 4 μl of proteinase K crystals grown in small crystallization well or in the microtube with 12–16 μl of 1.25 M Ammonium sulfate (Fig. 5a, b, d, e).
4. Pick up glow-discharged TEM grid with the Vitrobot tweezers and hold it in horizontal position.
5. Apply 3.5 μl of proteinase K mixture to TEM grid and blot manually with blotting filter paper from reverse side outside the Vitrobot chamber.
6. Mount the tweezers to Vitrobot immediately after blotting and plunge the grid in the liquid ethane.
7. Store grids with vitrified crystals in storage Dewar or clip it into the AutoGrid cartridge for loading into the FIB-SEM microscope.

3.8 Mounting TEM Grids into the AutoGrid

The workflow described here utilizes the TEM grids mounted into the AutoGrid™ (Thermo Scientific) to facilitate sample handling and transfer between SEM and TEM microscopes. Other options are available when working with the instrumentation from other microscope manufacturers.

The AutoGrid assembly workstation is filled with LN_2 . The LN_2 level covers the grid box with the TEM grids, but the mounting of the TEM grids into the AutoGrid cartridge is performed in LN_2 vapor. It is highly recommended to wear protective facemask or shield during clipping procedure to prevent contamination from breathing to the specimen. Do not work with the tools that have accumulated the ice contamination.

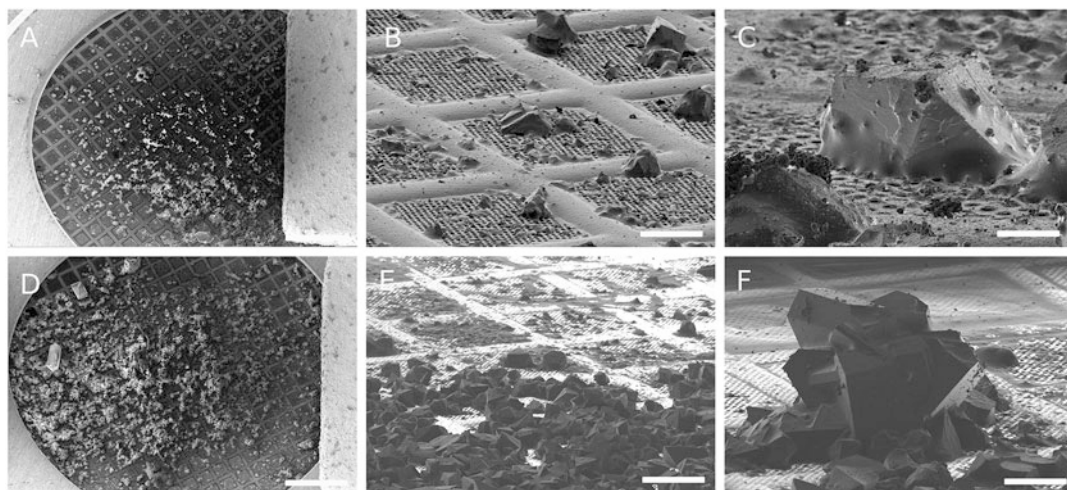


Fig. 5 (a) SEM overview of TEM grid with plunge frozen proteinase K crystals in density suitable for FIB milling, (b) FIB images of crystals dislocation on grid squares, and (c) single crystal with suitable size of 25 μm in diameter placed in the center of grid square, (d, e) SEM overview image and FIB image of the overloaded TEM grid with plunge frozen proteinase K crystals, (f) FIB image illustrating a cluster of proteinase K crystals with a large crystal ($>100\ \mu\text{m}$) not suitable for lamella preparation. The scale bars represent 400 μm in panels a and d, 30 μm in panels b and f, 10 μm in panel c, and 50 μm in panel e

1. Take the middle part of the C-clip with a precision tweezers and insert it vertically into the clipping tool.
2. Turn the clipping tool around and place it on a flat surface. Then, push the plunger gently down to align the C-clip properly horizontally to the edge of the clipping tool.
3. Dry all tools used for clipping to AutoGrid cassette before cooling down in LN_2 .
4. Put the AutoGrid assembly workstation together.
5. Place the AutoGrid into the clipping metal slot with the flat side facing down. Center it properly.
6. Fill AutoGrid assembly workstation with LN_2 to the level of the edge of the bottom disc. Avoid pouring LN_2 on the upper disc of the workstation.
7. Place the empty AutoGrid storage box and the grid box with vitrified sample grids into appropriate slots (*see Note 23*).
8. Unscrew the lid with pre-cooled grid box opener from the sample grid box (*see Note 24*).
9. Cool down clipping tool with the C-clip, AutoGrid tweezer, and precision tweezers (*see Note 5*).
10. Take one grid with pre-cooled precision tweezers out of the grid box and place it on the AutoGrid cartridge facing down.

11. Center the upper disc of the clipping assembly workstation over the grid using the AutoGrid tweezers.
12. Put the clipping tool with a C-clip on top of the grid and clip slowly.
13. Turn the disc back and take the AutoGrid assembly with the AutoGrid tweezer out of the slot and store it in the AutoGrid box (*see Note 25*).
14. Place the grid on the site groove and flip it over to see if the grid is properly mounted.
15. Position the clipped AutoGrid cartridge into the AutoGrid box.
16. Continue with the clipping or close the AutoGrid box with pre-cooled lid (*see Note 25*).
17. Store cartridges in the Dewar with LN₂ or load it into the FIB-SEM microscope.

3.9 Sample Manipulation in FIB-SEM Microscope

Strong charging effect is observed when imaging frozen hydrated biological material in SEM. In addition, imaging biological samples with FIB (even at low currents) induces fast sample damage. Therefore, additional coating of the sample is performed inside the FIB-SEM microscope, in order to protect the specimen surface and increase its conductivity.

1. For deposition of the protective layer with the Gas Injection System (GIS) (*see Note 26*), set the sample to eucentric height.
2. Tilt the stage back to 0° (sample tilted to 45° with respect to electron beam).
3. Move stage in *z* axis 4 mm below eucentric height (*see Note 27*).
4. Set the GIS needle to 26–30 °C.
5. Deposit ~20 nm of the GIS layer on the grid with biological specimen (Corresponds to 10–30 s of the GIS deposition).
6. For Sputter coating of the specimen surface with conductive metal layer, deposit ~10 nm of the metal layer (Ir, Au, Pt, etc.) to the grid with biological specimen (*see Notes 28–30*).
7. Set imaging and milling parameters: FIB—high voltage = 30 kV, current = 10 pA (imaging), 10–300 pA (FIB-milling) SEM—high voltage = 2–5 kV (Fig. 6), spot size = 4.5, current = 8–27 pA. Scan rotation: 180° Stage tilt – milling angle 6°–11° (+7° stage pre-tilt, (*see Note 31*)).

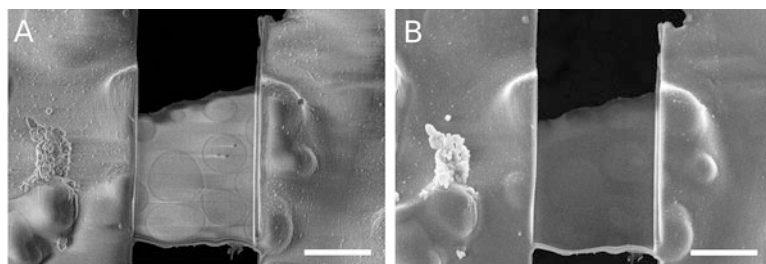


Fig. 6 Comparison of SEM imaging of *S. cerevisiae* lamella taken on ETD detector in FEG accelerating voltage (a) 5 kV and (b) 2 kV (scale bars = 4 μm)

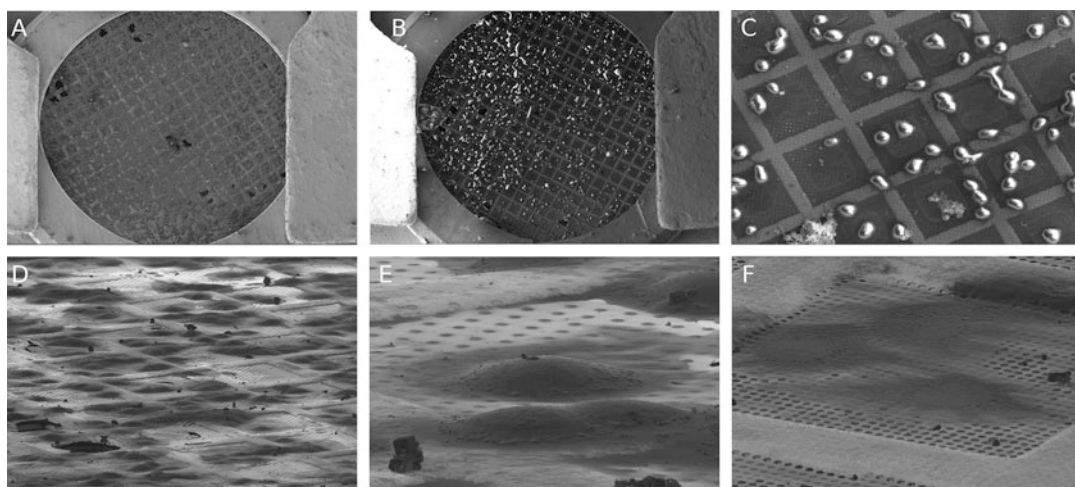


Fig. 7 Cryo-SEM images of TEM grid with vitrified A9 mammalian adherent cells. (a) Electron beam images of A9 cells showing overloaded grid, (b and c in detail) optimal concentration of the cells on the grid, (d) optimal concentration of the cells on the grid imaged by ion beam, (e) ideal positioning and spatiality of the cell, (f) grid with shriveled cells

3.10 SEM Imaging of the Specimen and Grid Quality Control before FIB Milling

1. Optimal cell or protein crystal concentration is reached when the specimen is individually distributed on the grid without any visible clustering (Fig. 7a–d).
2. The specimen should be positioned in the center of the grid square (*see* Note 32, Fig. 7c).
3. The specimen that should be surrounded by visible holey carbon holes showing the ice thickness of the grid square is sufficiently thin for effective milling.
4. The cells should be free of surface ice contamination coming from air humidity or breathing.
5. The cell, cell clusters, or protein crystal (Fig. 7e–f) should be at least 1.5 μm high and have more than 8 μm in X or Y dimension.
6. The holey carbon foil around the selected cells should be free of cracks (*see* Note 33).

3.11 Preparation of Cellular or Protein Crystal Lamella

When all requirements for the quality of the grid (see above) are fulfilled, select the object for lamella preparation.

1. Select the milling object nearby center of the grid (*see Note 32*).
2. Select the milling object at the center of the grid square (*see Note 33*).
3. Select the milling object at the grid square with compact holey carbon foil without cracks (*see Note 34*).
4. Select the milling object on the grid properly blotted from both sides without additional water around milling object and on the reverse side of the grid (*see Note 34*).

The milling pattern is generated (Fig. 8) and centered with respect to the region of interest. Cryo-FIBM is performed sequentially with multiple milling steps performed at different FIB settings. The lamella with roughly 2 μm thickness is initially milled using high current (300 pA). The lamella is subsequently gradually thinned further to the thickness of 500 nm. The fine-milling step at low current (10 pA) is used to finalize the lamella to ~200 nm thickness (Fig. 9).

1. Create upper pattern—box rectangle pattern above the region of interest with scan direction Top to Bottom.
2. Create lower pattern—box rectangle pattern below the region of interest with scan direction Bottom to Top.
3. Create middle pattern—deselected box rectangle pattern covering the region of interest providing rough estimate of the lamella thickness (this pattern is not milled during lamella preparation).
4. Mark all patterns and set the lamella width (x dimension). The width of the milling pattern should not exceed two-third of the cell width (*see Note 35*). This corresponds to 8–15 μm wide lamella in most of the cases.
5. Set parameters for rough milling steps (Table 1):
 - (a) FIB current: 300 pA; final lamella thickness: 1.5–2 μm ; width of the FIBM area: 8–12 μm ; stage-tilt: 13–17° (**Note 31**); active milling patterns: Upper and lower.
 - (b) FIB current: 100 pA; final lamella thickness: 1 μm ; width of the FIBM area: 7.5–11.5 μm ; stage-tilt: 13–17° (**Note 31**); active milling patterns: Upper and lower.
 - (c) FIB current: 30 pA; final lamella thickness: 0.5 μm ; width of the FIBM area: 7.5–11.5 μm ; stage-tilt: 13–17° (**Note 31**); active milling patterns: Upper and lower.
6. Set parameters for fine milling step (Table 1):

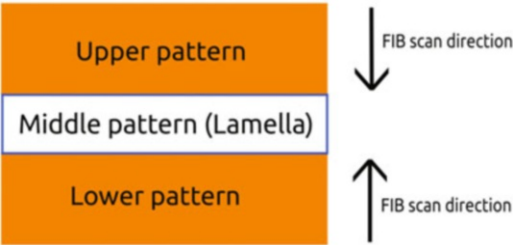


Fig. 8 Illustration of the pattern for lamella preparation. Arrows on the right side indicate recommended direction of the FIB scan direction. Dimensions for upper, middle, and lower pattern in individual milling steps are listed in Table 1

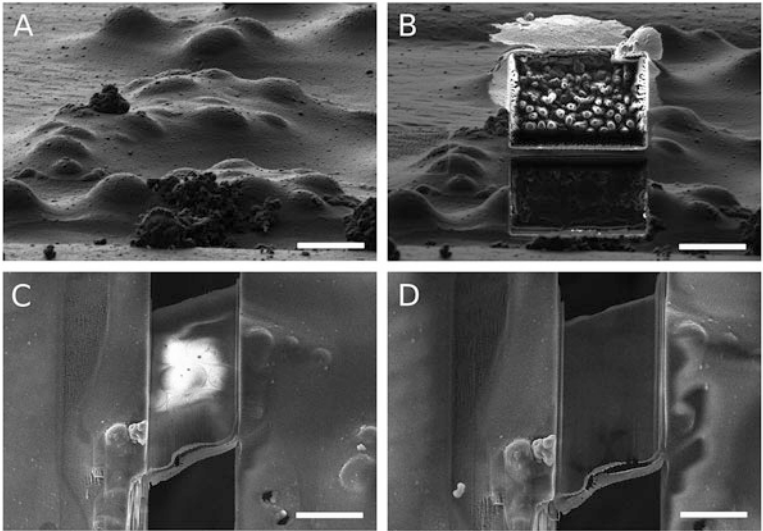


Fig. 9 FIB image of *S. cerevisiae* cell cluster plunge frozen on TEM grid and selected for (a) lamella preparation, (b) FIB image of the lamella after rough milling, (c) SEM image, (d) SEM image of *S. cerevisiae* lamella polished to the final thickness of ~200 nm. The scale bars represent 4 μm in panels a and b, 6 μm in panel c, and 5 μm in panel d

Table. 1
Summary of the FIB milling settings for individual steps of lamella preparation

Milling step	Current (pA)	Overtilt	Lamella thickness (um)	Lamella width (X) shrink (um)	Upper pattern	Lower pattern
1.	300	–	1.5–2	0	Allowed	Allowed
2.	100	–	0.8–1	–0.5	Allowed	Allowed
3.	30–50	–	0.5	–0.5	Allowed	Allowed
4.	10	+ 1°	<0.5	–0.5	Allowed	Disabled

7. FIB current: 10 pA; final lamella thickness: 0.2 μm ; width of the FIBM area: 7–11 μm ; stage-tilt: 13–17° (+1°, **Note 31**); active milling patterns: Upper.
8. Prepare properly dried Dewar and fill it with LN2 (*see Note 36*).
9. Unload the samples with lamellae from the FIB-SEM microscope under cryo-conditions with care, transfer it to a grid box, and store in LN2 storage Dewar for long-term storage. Alternatively, directly load the grids into the cryo-TEM.
10. Correct orientation of the lamella with respect to the cryo-TEM stage rotation axis is important. The milling direction of prepared lamellae needs to be perpendicular to cryo-TEM stage-tilt axis.
11. When using the dose-symmetric scheme (cit.) for cryo-ET data acquisition, pre-tilt the cryo-TEM stage to compensate for the lamella tilt (*see Note 31*).
12. Asymmetric angular distribution is advisable to use when collecting electron diffraction tomography data.

4 Notes

1. A typical cell culture media consist of amino acids, glucose, vitamins, inorganic salts, and serum as a source of hormones, growth factors, and adhesion factors. Moreover, the media also help to keep pH and osmolality. Phenol red as a medium component serves as a pH indicator. There are also extra media supplements that can help to optimize cell growth and viability. For example, antibiotics are often used to inhibit bacterial and fungal contamination. L-glutamine is important for energy production and for protein and nucleic acid synthesis as well. 2-Mercaptoethanol as a reducing reagent is utilized to avoid toxic levels of oxygen radicals. In addition, glucose sodium pyruvate is added as a carbon source. Non-essential amino acids can replace depleted amino acids during the cell growth.
2. The cells look healthy when they are mainly attached to the bottom surface of the culture flask, round and plump, or elongated in shape (A9 cells possess elongated star-like shape), and refracting light around their membrane. Media should be pinkish-orange in color. On the other hand, the dead cells detach from the surface and/or look shriveled and dark in color or if they are in quiescence—not growing at all.
3. Purpose of the medium change in this step is the removal of DMSO used as a protective agent for cell freezing.

4. When the cells are ~80% confluent, they still shall be in their log phase of growth and in the best condition for passaging. It is not recommended to let the cells over-grow as this may negatively affect cellular expression of certain genes and viability of cells.
5. The purpose of the washes with DPBS is to remove any traces of serum, calcium, and magnesium that can inhibit the Trypsin-EDTA dissociation action in the following step.
6. Trypsin can be toxic to some cell lines. In such cases, cells can be dissociated by gentle scraping or using a very mild dissociation reagent.
7. Dissociation time can vary between cell lines. Optimize the incubation time needed for complete dissociation of the cells. However, do not expose the cells to trypsin solution for periods longer than 10 min because trypsin causes damage to the cells.
8. The number of passages should be recorded and shall not exceed 30. This is protection against use of the cells undergoing genetic drift and other variations.
9. When using Quantifoil grids, “carbon side up” position of the grids in the chamber slide can be checked under the light microscope. The Quantifoil grids have a “1” mark at the edge. If a mirrored “1” is observed, then the grids are placed carbon side up.
10. Fetal bovine serum (FBS) included in complete medium contains adhesion-promoting molecules such as fibronectin and vitronectin and thus helps in adhesion of cells to the surface of the grid.
11. Cell concentration used for cell adhesion on the surface of the grid have to be optimized for each cell line.
12. We have found that 80 μl volume of cell suspension per well is optimal. The volumes below 50 μl are not sufficient for the cells to survive in the incubator overnight.
13. Do not go too close to the grid with the pipette when applying cell suspension, as having the tip too close to the grid would cause its adhesion to the pipette tip and improper positioning of the grid at the surface of the media (Fig. 3d). Therefore, apply the cell suspension from a distance (Fig. 2c). If the grids end-up on the top of the drop facing upside down, return it back to its original position using tweezers.
14. The A9 cell are adhered when they have a star-like shape, and they are dead and not adhered to the grids when they stay round. There might be several reasons for cells not to adhere on the grids, for example, the cell suspension was contaminated during the preparation or low volume of cell suspension was applied on the grids.

15. Vitrification parameters must be experimentally determined for each instrument because they can vary among different plunge freezing devices (even in the case of using the same type of instrument).
16. Start of exponential phase is dependent on the age of culture on agar plate, use agar plates that are no more than 6 weeks, then prepare new agar plate with *S. cerevisiae* cell culture.
17. Yeast optical density $OD_{600} = 1$ corresponds to 10^7 cells/ml.
18. Non-absorbent surface is used as an alternative material for samples adhesive to filter paper. On the other hand, filter paper on back-side of the grid aspirates any excess liquid from the grid.
19. To maximize quality and reproducibility of *S. cerevisiae* specimen, mix the cell suspension properly before every grid preparation. The *S. cerevisiae* cells quickly pellet to the bottom of the tube. Figure 6 shows different cell density on prepared TEM grids.
20. Proteinase K forms uniform sized crystals in the range of 5–10 μm approximately after 2 h of crystallization process.
21. This size of crystals is suitable for micromachining of lamella in crystal clusters and collection of diffraction data in multiple places on lamella. On the other hand, crystals growing in 24 h are much larger with large variability in size (5–40 μm). Single crystals of proteinase K suitable for FIB milling grow to the maximal size 10–50 μm (Fig. 5c). Proteinase K rarely forms crystals >100 μm inappropriate for lamella preparation (Fig. 5f).
22. Blot force is a very instrument-specific parameter. We recommend testing different blot force settings beforehand and select the optimal value.
23. We recommend to carefully check the position of the TEM grid in the AutoGrid before clipping. Misalignment will cause grid deformation resulting in sample damage.
24. Care should be taken when opening the grid box, as grids may eventually stick to the lid of the grid box.
25. Ice contamination increases with time. Therefore, clip maximally eight grids at once.
26. GIS embedded inside of the microscope chamber is used to treat the biological sample with a thin layer that protects the sample from radiation damage during FIB milling and imaging with ion beam.
27. GIS contains a reservoir of organic metal compound of platinum (Pt), carbon (C), palladium (Pd), or tungsten (W) in a liquid state. Organic metal solution is injected by the

micromanipulator with the needle in the form of aerosol, which solidifies on the specimen surface and forms a solid layer—"crust." During milling is the layer evident as the so-called lamella front.

28. Sputter coating is a process using gaseous plasma for deposition of the metal layer to the specimen. Deposited metal creates a thin conductive layer on its surface and facilitates SEM imaging.
29. The most common metals for sputter coating of biological material are gold, platinum, iridium, or carbon.
30. Sputter-coating conditions such as current, voltage, material type, chamber pressure, and duration of sputtering determine the final thickness of the metal layer. We recommend to carry out a test to calibrate the sputter coater which is not equipped with the thickness sensor. Procedure: take the empty TEM grid covered with foil with defined thickness and defined size of holes. Choose sputter conditions (e.g., sputter time 30s, pressure 8×10^{-2} to 2×10^{-2} mbar, sputtering voltage 0.1–3 kV, current 10pA) and measure thickness of the metal layer intensity loss in TEM or using electron tomography.
31. Grids in the special holder "shuttle" are mounted and pre-tilted at 45° angle. When the cryo-stage in the SEM chamber is at 0° tilt angle, grid is oriented in the angle 45° to the electron beam and in -7° relative to the ion beam, respectively, in case of the Versa 3D-microscope (ThermoScientific) used here. The specimen is mostly hidden behind AutoGrid edge at 0° stage-tilt when imaging by FIB. It is important to choose a suitable stage-tilt angle to rotate the sample visible by FIB without interfering of the AutoGrid edge. Perfectly handled and clipped grid in AutoGrid cartridge should be uniformly flat. The minimal stage-tilt for milling allowed by currently used AutoGrids is 11° (4° between the FIB direction and the grid surface). However, due to imperfect geometry of the grid or in order to access milling positions closer to the AutoGrid edge, it is usually needed to choose higher tilt angles. Nevertheless, milling tilts higher than 20° are not recommended for two reasons: (1) lamella might get too short (2) the range of tilts available for tomogram collection will be much narrower.
32. Milling closer to the base of the object will result in longer lamellae (sampling more of the cell volume). Objects close to the grid bars are harder to mill and will have narrower range of tilt series. Objects sitting directly on grid bars are ignored. Milling objects that satisfy all of the above criteria and are located near or at the center of grid are preferred over the objects closer to the edge of the grid.

33. Any crack around the milled object would cause sample vibrations during tomographic data acquisition.
34. Milled lamella have to be supported by the mass of non-milled part of cell; otherwise, it is unstable and prone for bending or rupture.
35. Better results of fine and smooth lamella surface without curtaining (e.g., Fig.9d) can be achieved when the upper (not lower) part of the lamella is subjected to “fine” milling and thinning (in this case lamella is more stable); this process is carried out in parallel milling mode (both upper and lower patterns are active simultaneously), but the lower pattern is disabled.
36. We have observed a recurring phenomenon, that contamination from air humidity is readily deposited on milled lamella. Obtaining high-quality cryo-ET data on lamella covered with ice contamination is considerably more difficult or impossible.

Acknowledgments

This work was supported by Instruct-ULTRA (Coordination and Support Action Number ID 731005) to further develop the services of Instruct-ERIC and iNEXT-Discovery (Grant ID 871037), both funded by the Horizon 2020 program of the European Commission. We acknowledge the support obtained from Thermo Fisher Scientific, Brno.

References

1. McMullan G, Faruqi A, Clare D, Henderson R (2014) Comparison of optimal performance at 300keV of three direct electron detectors for use in low dose electron microscopy. *Ultramicroscopy* 147:156–163
2. Zivanov J, Nakane T, Forsberg BO et al (2018) New tools for automated high-resolution cryo-EM structure determination in RELION-3. *elife* 7:e42166
3. Villa E, Schaffer M, Plitzko JM, Baumeister W (2013) Opening windows into the cell: focused-ion-beam milling for cryo-electron tomography. *Curr Opin Struct Biol* 23:771–777
4. Schur FK (2019) Toward high-resolution in situ structural biology with cryo-electron tomography and subtomogram averaging. *Curr Opin Struct Biol* 58:1–9
5. O'Reilly FJ, Xue L, Graziadei A et al (2020) In-cell architecture of an actively transcribing-translating expressome. *Science* 369:554–557
6. Nannenga BL, Gonen T (2014) Protein structure determination by MicroED. *Curr Opin Struct Biol* 27:24–31
7. Nannenga BL, Gonen T (2016) MicroED opens a new era for biological structure determination. *Curr Opin Struct Biol* 40:128–135
8. Al-Amoudi A, Norlen LP, Dubochet J (2004) Cryo-electron microscopy of vitreous sections of native biological cells and tissues. *J Struct Biol* 148:131–135
9. Al-Amoudi A, Studer D, Dubochet J (2005) Cutting artefacts and cutting process in vitreous sections for cryo-electron microscopy. *J Struct Biol* 150:109–121
10. Pierson J, Fernández JJ, Bos E et al (2010) Improving the technique of vitreous cryo-sectioning for cryo-electron tomography: electrostatic charging for section attachment and implementation of an anti-contamination glove box. *J Struct Biol* 169:219–225

11. Dubochet J, Zuber B, Eltsov M et al (2007) How to "read" a vitreous section. *Methods Cell Biol* 79:385–406
12. Hsieh C, Schmelzer T, Kishchenko G et al (2014) Practical workflow for cryo focused-ion-beam milling of tissues and cells for cryo-TEM tomography. *J Struct Biol* 185:32–41
13. Rigort A, Bauerlein FJ, Villa E et al (2012) Focused ion beam micromachining of eukaryotic cells for cryoelectron tomography. *Proc Natl Acad Sci USA* 109:4449–4454
14. Schaffer M, Engel BD, Laugks T et al (2015) Cryo-focused ion beam sample preparation for imaging vitreous cells by Cryo-electron tomography. *Bio-Protoc* 5:e1575



Protein and Small Molecule Structure Determination by the Cryo-EM Method MicroED

Emma Danelius and Tamir Gonen

Abstract

Microcrystal Electron Diffraction (MicroED) is the newest cryo-electron microscopy (cryo-EM) method, with over 70 protein, peptide, and several small organic molecule structures already determined. In MicroED, micro- or nanocrystalline samples in solution are deposited on electron microscopy grids and examined in a cryo-electron microscope, ideally under cryogenic conditions. Continuous rotation diffraction data are collected and then processed using conventional X-ray crystallography programs. The protocol outlined here details how to obtain and identify the nanocrystals, how to set up the microscope for screening and for MicroED data collection, and how to collect and process data to complete high-resolution structures. For well-behaving crystals with high-resolution diffraction in cryo-EM, the entire process can be achieved in less than an hour.

Key words MicroED, 3D Electron Crystallography, Cryo-EM, Electron Diffraction, Protein Structure, Grid Preparation, Microcrystal, Nanocrystal, Transmission Electron Microscopy

1 Introduction

Modern cryo-EM encompasses at least five techniques: cryo-electron tomography [1], single-particle cryo-EM [2], helical reconstruction [3], electron crystallography of 2D crystals [4], and MicroED [5] (Fig. 1). In all these methods, frozen-hydrated samples [6] are examined in a transmission electron microscope (TEM); however, the data collection mode differs where the first three techniques use imaging of the biological sample exclusively, and the last two, use diffraction in addition to imaging.

In MicroED, high-resolution electron diffraction data are collected from 3D micro- and nanocrystals, a billionth the size of those used for X-ray diffraction [5]. The use of such vanishingly small crystals has opened up for a new era in protein structural elucidation as previously unattainable “difficult to crystalize” samples are now accessible for investigation [5]. This includes several highly important targets such as various membrane proteins [7]

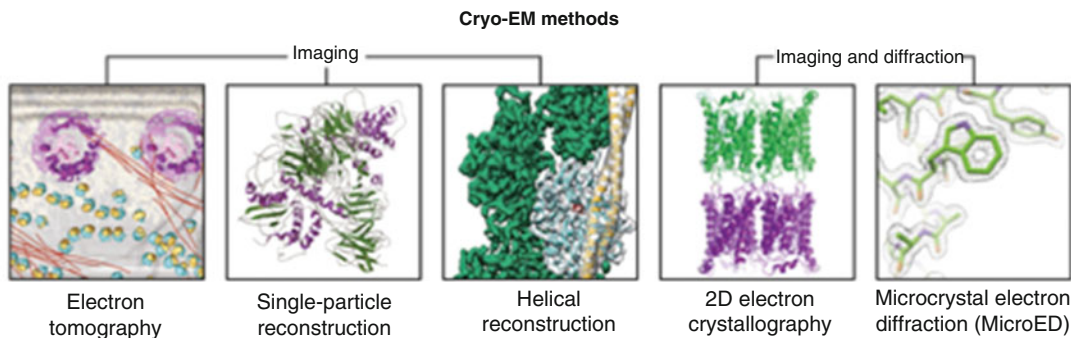


Fig. 1 Cryo-EM can be used to provide structural information from a wide range of samples. There are five major techniques: Cryo-electron tomography was developed to study whole cells and large organelles at resolutions of about 1 nm [30]. In single particle reconstruction isolated single particles are imaged and near atomic structures can be determined, [31] and similarly for helical assemblies helical image reconstruction can be used [32]. In 2D electron crystallography, the structures are obtained from the diffraction of highly ordered 2D crystals containing only one layer of the protein of interest [33] and in MicroED from the diffraction pattern of nano- or micro-sized three-dimensional crystals [8]

and protein complexes [8], as well as small molecules [9] and natural products [10]. As compared with X-ray-free electron lasers (XFELs) [11] which also obtain data from microcrystals, MicroED is less expensive, requires substantially less sample, crystals can be substantially smaller, and the equipment needed is more accessible and easier to maintain and handle. Moreover, structures in MicroED can be completed with a single nanocrystal whereas in XFELs many thousands are necessary further complicating the procedure. A large part of the success of MicroED is owing to the use of continuous rotation during data collection [12], leading to well-defined rocking curve parameters and enabling data processing by well-established X-ray diffraction programs including MOSFLM [13, 14], XDS [15], the HKL suite [16], DIALS [17], CNS, [18, 19] PHENIX [20], Buster [21], SHELX [22], and the CCP4 [23] suite. Another important feature of MicroED is the use of an extremely low electron exposure (dose rate typically $\sim 0.01 \text{ e}^-/\text{\AA}^2/\text{s}$), minimizing radiation damage [24] during data collection and allowing for micro- and nanosized crystals to be examined. Since its initial demonstration in 2013 various proteins [7, 12, 25], protein complexes [8], protein–ligand complexes [26], peptides [27, 28], small molecules [9], natural products [10], and inorganic material [29] have been described by MicroED (Fig. 2). Future directions include developing imaging methods for phasing and high throughput structure determination [5].

Herein, we describe step-by-step protocols with all the necessary details to prepare samples, collect data and solve structures using MicroED. The data is collected from vanishingly small crystals, with a thickness of $\sim 500 \text{ nm}$ or less. This makes the method highly attractive for anyone having difficulties in crystallizing their

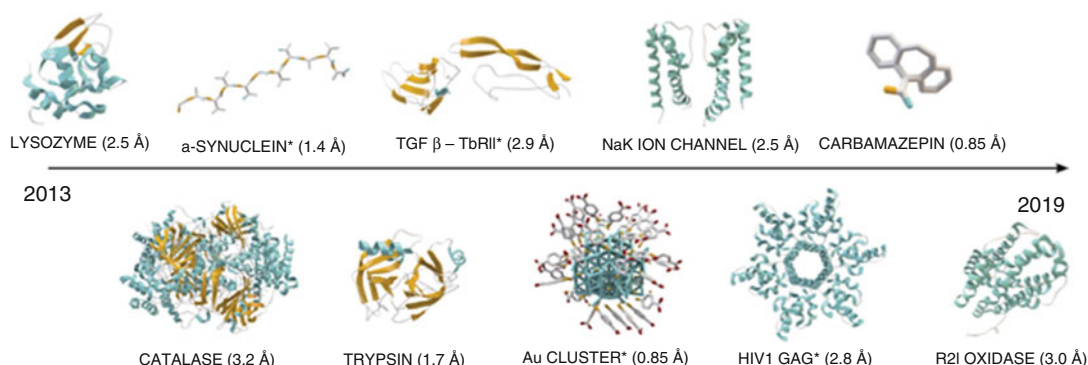


Fig. 2 The array of MicroED structures published so far includes lysozyme [12], catalase [25], the amyloid core of α -synuclein [27], proteinase K [8], trypsin, [8], the complex of TGF- β m bound to its receptor T β RII [8], an inorganic gold cluster [29], the NaK ion channel [7], the HIV1 GAG protein bound to its ligand bevirimat [26], the R2lox enzyme [34], and several small organic molecules including carbamazepine [9]. Novel structures are marked with asterisk. There are today over 70 published MicroED structures

target, but MicroED can also be used for thicker crystals once they are broken into smaller crystalline domains, as described in Sub-heading 3.1. A general overview of the workflow is given in Fig. 3. The crystallization of microcrystals is carried out by the same methods as used for X-ray crystallography, such as vapor diffusion by hanging drop or sitting drop and liquid–liquid diffusion. If the crystals are too small to be verified by standard light microscopy, other methods such as negative staining EM can be used. The protocol below also explains how focused ion beam (FIB) milling in a scanning electron microscope (SEM) can be used to prepare crystalline lamella of desired thickness for MicroED. For grid preparation, the solution of microcrystals is deposited on cryo-EM grids followed by blotting and vitrification by plunge-freezing into liquid ethane. The grids are then screened, and data are collected and processed as described below. The structures can be solved from single crystals or by merging several data sets. For a crystal with high-quality diffraction, the entire process of data collection and processing can be achieved in less than an hour.

2 Materials

2.1 Preparing Crystals for MicroED

1. Protein sample.
2. Micropipette, capable of pipetting 1–10 μ L solution.
3. Pipette tips.
4. Ultrasonic water bath.
5. Microcentrifuge tubes.
6. Parafilm.

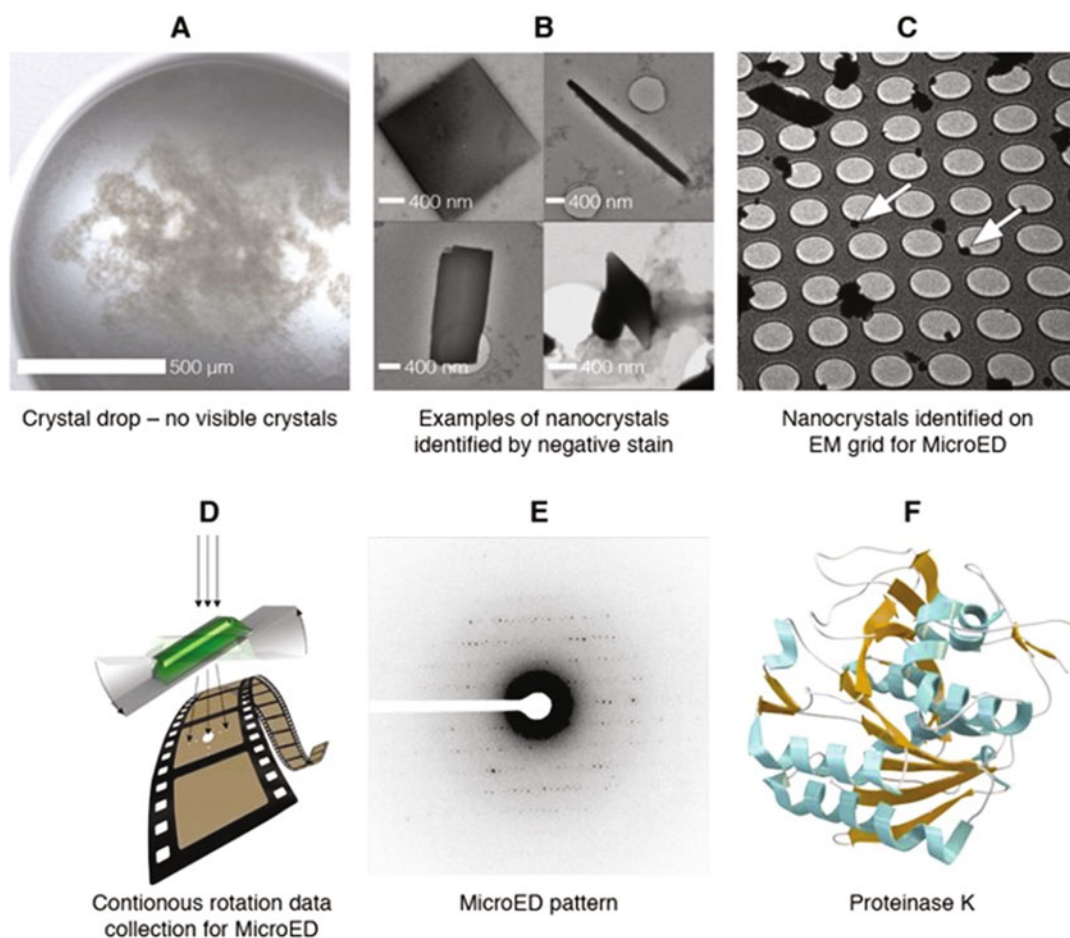


Fig. 3 Overview of workflow. Crystal drops typically contain invisible crystals (**a**) that can be identified by, for example, negative stain (**b**). The nanocrystals are placed onto an electron microscopy grid (**c**, white arrows), and MicroED data are collected from a single crystal (**d**). The collected data (**e**) are processed using standard crystallographic software, yielding the structure, here exemplified by proteinase K (**f**) [8]

7. 0.5 mm glass bead.
8. Vortex mixer.
9. Cryo-FIB DualBeam system, for example Aquilos from ThermoFisher Scientific.
10. Liquid nitrogen.

2.2 Identification of Microcrystals

1. Uranyl formate.
2. Buntzen burner.
3. MilliQ water.
4. Glass test tube.
5. Magnetic stirrer.

6. Aluminum-wrapped beaker.
7. 5 M NaOH.
8. 0.22 μ m syringe filter.
9. Aluminum-wrapped 8 mL falcon tube.
10. Anticapillary reverse (self-closing) tweezers.
11. Glow discharge cleaning system (e.g., PELCO easiGlow).
12. Parafilm.
13. Glass slide.
14. Whatman filter paper 1.
15. Cryo-grid storage boxes.

2.3 Grid Preparation

1. Glow discharge cleaning system (e.g., PELCO easiGlow).
2. Anticapillary reverse (self-closing) tweezers.
3. Micropipette, capable of pipetting 1.5–3 μ L solution.
4. 300-Mesh copper holey carbon EM grids (Quantfoil, SPI supplies).
5. FEI Vitrobot Plunge-freezer System including Vitrobot coolant container.
6. Standard Vitrobot filter paper.
7. Locking Tweezers Assembly for Vitrobot.
8. Slide Warmer, for example, Premiere XH-2002.
9. Liquid nitrogen.
10. Ethane gas.
11. Cryo-grid storage boxes and gripper tool.
12. Whatman Filter paper 1.

2.4 Data Collection and Processing

1. Gold/graphitized calibration grids (Ted Pella, prod. no. 638).
2. Microscopes: Titan Krios or Talos Arctica transmission electron microscopes, both from ThermoScientific.
3. Cameras: Falcon 3EC Direct Electron Detector (ThermoFisher Scientific), K3 Direct Electron Detector (Gatan), Ceta 16 M camera (ThermoFisher Scientific), TVIPS TemCam-F416.
4. Image conversion tools: the conversion tools necessary for MRC or TVIPS file format to SMV are available on the Gonen lab webpage <https://cryoem.ucla.edu/downloads>
5. Software: EPU-D, CCP4, XDS, iMOSFLM, and Phenix.

3 Methods

3.1 *Preparing Crystals for MicroED*

Micro- and nanocrystals are prepared by the same crystallization methods used for larger crystals such as vapor diffusion by hanging drop or sitting drop and liquid–liquid diffusion. Crystals for MicroED should ideally be thinner than ~500 nm. If required, crystal growth can be restricted by using high salt concentration. In addition, larger crystals can easily and mechanically be broken into smaller crystal domains by pipetting, ultrasonication, or vortexing, as described below.

3.1.1 *Fragment Crystals by Pipetting*

1. Place the crystal drop under the light microscope.
2. Use a micropipette with a new tip and pipette up and down a couple of times directly into the drop while viewing the crystals in the microscope. Be careful not to create air bubbles.
3. Add well buffer/mother liquor if needed to dilute the sample.

3.1.2 *Fragment Crystals by Sonication*

1. Prepare an ultrasonic water bath by adding water to the container.
2. Transfer the crystal solution to a new microcentrifuge tube.
3. Add well buffer/mother liquor if needed to dilute the sample.
4. Close the microcentrifuge tube and wrap it with parafilm.
5. Set the sonicating bath to run continuously and dip the bottom end of the microcentrifuge tube into the water, make sure that the surface of the crystal solution is below the surface of the water bath (*see Note 1*).

3.1.3 *Fragment Crystals by Vortexing*

1. Transfer the crystal solution to a new 1.5 μ L microcentrifuge tube.
2. Add well buffer/mother liquor if needed to dilute the sample.
3. Add a glass bead of 0.5 mm diameter to the microcentrifuge tube. Close and wrap with parafilm.
4. Set the vortex mixer to auto mode at the highest speed setting. Use the flat rubberized head platform and place the bottom of the microcentrifuge tube in the center. Vortex for 2 s (*see Note 2*).

3.1.4 *FIB Milling*

1. Prepare the Aquilos by purging the lines with 10 L/min nitrogen gas flow, for at least 0.5 h.
2. Cool down the Aquilos and wait for the temperatures to stabilize, according to manufactures' instructions.
3. Prepare grids according to Subheading 3.3.
4. Load grids into Aquilos under liquid nitrogen temperature.

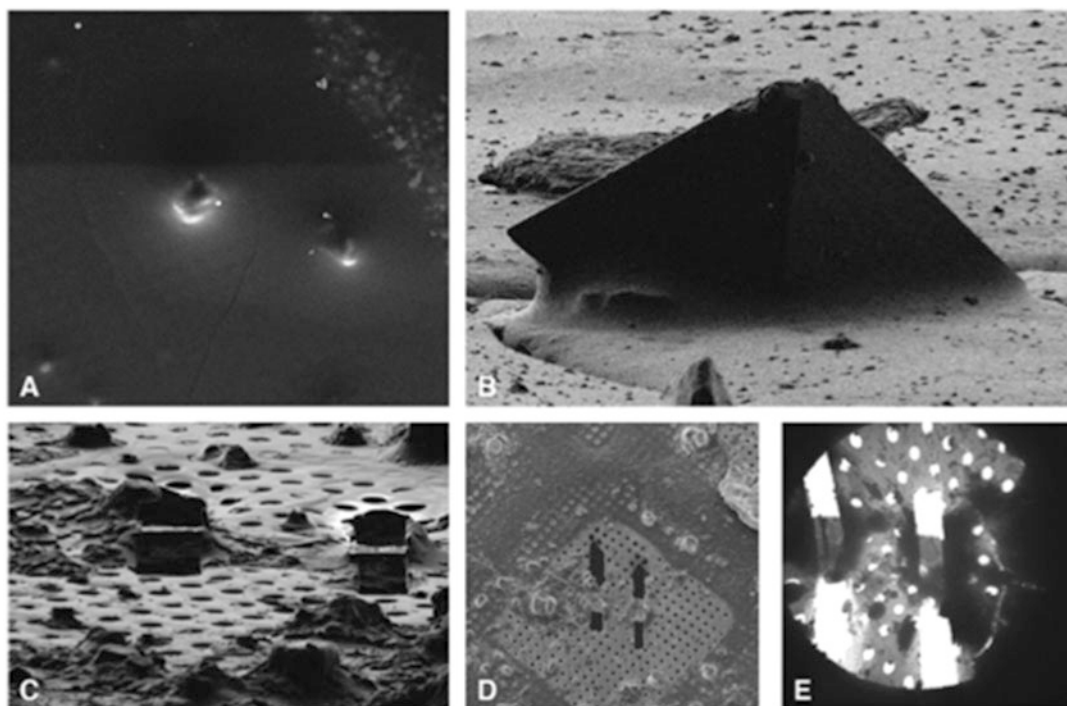


Fig. 4 Images of crystals in SEM (a) and FIB (b) which are stepwise milled to a final thickness of about 200–300 nm, generating lamella here shown in FIB (c), SEM (d), and in the cryo-TEM (e)

5. Select a grid and move to deposition position to prepare for sputtering.
6. Select the time and current for sputtering. We usually use 7 mA, for 30 s, at 1 kV and pressure of 10 Pa.
7. When finished, move grid to mapping position to screen for crystals. We usually use 3.1 pA (5 kV) in SEM for screening (Fig. 4 a).
8. When a crystal has been identified, set the eucentric height by tilting to 18° and adjust with the stage Z bar.
9. Link Z height to FIB at 18 degrees and move to the FIB to mill crystal (Fig. 4b–c).
10. Place two rectangular milling patterns on the crystal and select a milling direction of top to bottom for the upper box, and bottom to top for the lower box. The milling is a stepwise process and the power and time is dependent on crystal size. In the first step, the boxes should be 3–5 μm apart. A good starting point is 0.5 nA in the first round and then gradually reduce the current to 30 pA to get the final thickness of 200–300 nM (see **Note 3**).

11. When the lamella has been milled to the desired thickness, move to the next crystal until all sites have been milled. Keep the grids under liquid nitrogen temperature when transferring to the cryo-electron microscope (Fig. 4 e).

3.2 Identification of Microcrystals

Crystals that due to their small dimensions are invisible by optical microscopy can be identified by ultraviolet (UV) light, fluorescence, or Second Order Nonlinear Imaging of Chiral Crystals (SONICC), using for example a Rock Imager, or by negative stain. Outlined below is our protocol for identification of microcrystals using negative stain, and in Subheading 3.5 it is described how the cryo-electron microscope can be used to screen for micro- and nanocrystals.

3.2.1 Identification of Microcrystals Using Negative Stain EM

1. Weigh 50 mg uranyl formate (*see Note 4*) into a beaker. Use a Bunsen burner to heat up MilliQ water in a glass test tube (*see Note 5*) and dissolve the uranyl formate in 5 mL of the water. Stir the mixture under an aluminum-wrapped beaker for 5 min.
2. Add 5 μ L 5 M NaOH into the mixture and stir for another 5 min.
3. Filter the uranyl formate solution through a 0.22 μ m filter and transfer to aluminum-wrapped 8 mL falcon tube (*see Note 6*).
4. Use tweezers to transfer carbon-coated grids onto glass slides wrapped in parafilm. Place the glass slide with grids in a glow discharge chamber and glow discharge for 30 s (*see Subheading 3.3.1*).
5. Prepare three drops of MilliQ water and 2 drops of uranyl solution for each grid sample on parafilm.
6. Use tweezers to pick up a grid and add 2 μ L sample to the carbon-coated side. Wait 20 s and then blot excess solution onto filter paper.
7. Wash the grid by gently touching the carbon-coated side of the grid on the surface of the first MilliQ water drop, without disrupting the surface tension (*see Note 7*). Blot on filter paper and repeat for the other two drops.
8. Wash the grid in the same way as in the first uranyl drop and blot on filter paper.
9. Stain the grid by gently rotating the grid in the last uranyl drop for 20 s, again without disrupting the surface tension.
10. Dry the grid by aspirating off the side of the grid for a few seconds using a vacuum. Store the grids in grid boxes.

3.3 Grid Preparation for MicroED

3.3.1 Glow Discharge

Place grids with the carbon side up in a glow discharge cleaning system and run it with negative polarity to make grids more accessible for the crystal drop. We typically use 1×10^{-1} mbar vacuum and 15 mA power and glow discharge with negative polarity for 30 s.

3.3.2 Automated Vitrification

The procedure described below is a typical procedure using the Vitrobot from ThermoFisher Scientific; however, other vitrification systems can be used as well.

1. Assemble the Vitrobot coolant container by placing the cryo-gridbox and the inner ethane container into the box holder and put the aluminum bridge on top (Fig. 5) and add liquid nitrogen to the outer reservoir to cool it down (*see Note 8*).
2. Attach filter paper to the blotting pads of the Vitrobot and set the temperature to and humidity (*see Note 9*). Put all tools that will be used on a 38 °C slide warmer.
3. Slowly fill the inner brass reservoir of the Vitrobot coolant container with ethane gas, the low temperature will cause it to condense (*see Note 10*).
4. When the reservoir is filled, wait for the ethane to begin freezing and then remove the aluminum bridge and place the container into the Vitrobot (*see Note 11*).
5. Use the locking tweezers to pick up a glow discharged grid and place in the Vitrobot with the carbon side to the left. Insert grid into the chamber of the Vitrobot.
6. Carefully pipette 1.5–3 μ L microcrystal solution onto the carbon side of the grid. Select blotting time and force and blot the sample to remove excess solution (*see Note 12*).
7. Plunge freeze into the liquid ethane and carefully remove the tweezers and quickly transfer the grid to a cryo-gridbox in the liquid nitrogen outer reservoir. Make sure to keep grid under liquid ethane or nitrogen all the time.
8. The grids can either be transferred to the microscope for examination or they can be stored in liquid nitrogen.

3.3.3 Manual Blotting and Freezing

1. Prepare the coolant container as described in Subheading 3.3.2, preferably in a fume hood.
2. Carefully pipette 1.5–3 μ L microcrystal solution to a glow discharged grid and blot by placing it carbon side up on a filter paper.
3. When the excess liquid has been blotted, quickly freeze in the liquid ethane and transfer to cryo-gridbox as described in Subheading 3.3.2.

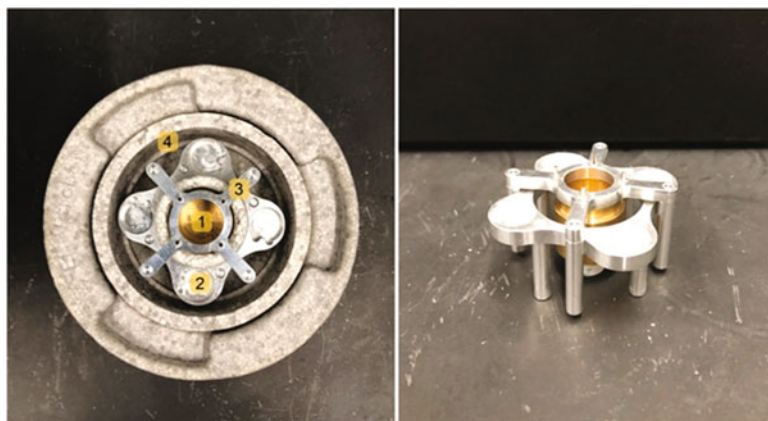


Fig. 5 Assembled Vitrobot coolant container, with inner brass ethane container (1) cryo-gridbox holder (2), aluminum bridge (3), and outer reservoir for liquid nitrogen (4)

3.4 Setting up the Microscope (ThermoFisher Cryo-TEMs Such as Glacios, Talos Arctica, and Titan Krios)

3.4.1 Calibrating the Electron Dose of the Microscope

4. The grids can either be transferred to the microscope for examination or they can be stored in liquid nitrogen.
1. Find a large space such as an area of broken carbon film or unload the grid.
2. Go to exposure/bright field mode and keep the same illumination settings (C2 condenser and intensity) as will be used in diffraction mode. The beam must be parallel and illuminate the entire sensor.
3. Use a direct electron detector such as the Falcon 3EC and record a 10 s exposure.
4. Note the dose rate and convert to $e^-/\text{\AA}^2/\text{s}$.

3.4.2 Calibrating Detector Distance

1. Load a TEM grid with a standard sample for calibration, for example, gold/graphitized grids (Ted Pella, prod. no. 638).
2. Align the microscope as described in Subheading 3.4.3.
3. Measure the diffraction pattern and calculate the distance according to the formula $L\lambda \approx Rd$, where L is the calibrated detector distance, λ is the relativistic wavelength of electrons (0.0251 \AA at 200 kV acceleration voltage), R is the radius of the diffraction ring (see Note 13), and d is the known lattice spacing of the standard sample.

3.4.3 Setting up the Microscope for Low-Dose Electron Diffraction (Talos Arctica)

1. Load the frozen grids into the microscope under liquid nitrogen temperature.

2. Set the microscope to diffraction mode and activate the low-dose mode (*see* **Note 14**). The low-dose mode consists of search, focus, and exposure modes.
3. In search mode, use the magnification control to set the camera distance. Set the spot size and the C2 condenser. These settings will be different for every microscope (*see* **Note 15**).
4. With the microscope in search mode and the screen inserted, find a feature on the grid such as a thick ice or black blob and center on it.
5. Switch to exposure mode. Use the magnification control to set the detector distance; for protein samples, we typically use 2–3 m and for small molecules 0.85–1.1 m (*see* **Note 16**).
6. Center the beam using the diffraction shift controls and make it as small as possible using the intensity control.
7. Switch back to search mode and center to the burn mark using the diffraction shift controls.
8. Repeat **steps 6** and **7** until the beam stays centered after changing between the two modes.
9. In search mode, decrease the intensity so that the image of the grid is slightly larger than the outermost of the red circles.
10. Using the focus control, reduce focus until an oval shape is visible within the red circles.
11. Using the beam shift, adjust the beam to the center of the oval. Decrease the focus further until a three-pointed star is visible and align the star using the stigmator controls.
12. Make the beam as small as possible using the focus and then reset defocus. Set the defocus to 1.2 e^{-6} .
13. Set the intensity to its original value.
14. Repeat **step 6–7** to check that the beam did not move.

3.5 Grid Screening

For low magnification screening of the grids, an Atlas can be collected using EPU-D (Subheading 3.5.1). The grids can also be screened using the search mode (Subheading 3.5.2), within the low-dose settings (Fig. 6).

3.5.1 Atlas

1. Align the beam in imaging mode. Set the spot size, magnification, and C2 condenser. Again, these will be different for different microscopes (*see* **Note 17**).
2. In search mode, decrease the intensity so that it is slightly larger than the middle of the red circles.
3. Center using the beam shift.
4. Increase the intensity so that it fits the outermost of the red circles.

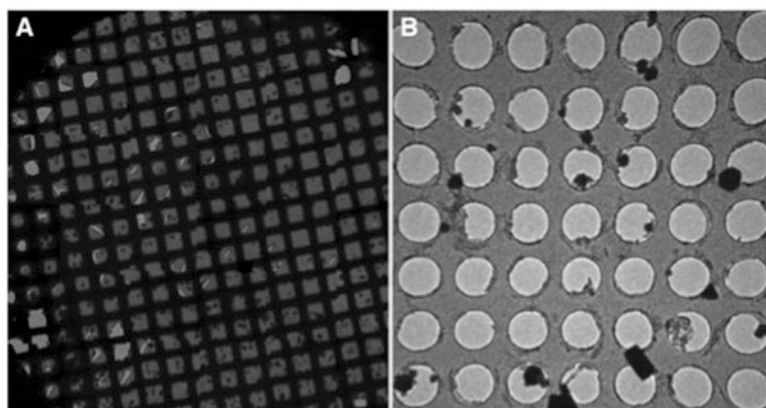


Fig. 6 To identify crystals on the grids, an overview or Atlas (a) can be collected or the grids can be screened using the search mode (b)

5. Center to the ring using the C2 lens aperture controls.
6. Repeat **steps 2–5** until and check that the beam did not move.
7. Lift the screen and start EPU Atlas.
8. Under the preparation tab set presets to Atlas, binning to 1, readout to full, exposure time to 3, noise reduction to yes, frames summed to 1 and intensity to 100. Also set the magnification and spot size to the values used for imaging mode (*see Note 17*).
9. Use the preview to quickly check that the settings are fine.
10. Go to atlas acquisition and acquire to record the full atlas.
11. Use “move stage to” and “save position” to save the positions of the potential microcrystals.

3.5.2 Grid Screen in Search Mode

1. In search mode, move around to look for areas of thin ice and well-separated microcrystals.
2. To check for diffraction, center on the potential crystal and insert the selected area aperture.
3. Blank the beam and go to exposure mode.
4. Insert the beam stop and lift the screen.
5. Under low dose, set the integration time to 3, sampling to 2, and read out area to full.
6. Unblank the beam and press acquire.
7. To continue to search for crystals, go back to search mode, remove the selected area and the beam stop, and insert the screen.

3.6 Data Collection

As the stage in MicroED data collection is continuously rotated, the frame rate of the detector has to be fast enough. Therefore, charge-coupled device (CCD) detectors are not generally recommended

and complementary metal-oxide-semiconductor (CMOS)-based detector (e.g., TVIPS F416) or direct electron detectors (e.g., Falcon 3EC) should be used. To set the rotation, we use the Delphi scripting interface, but this can also be done in EPU or TEMspy.

1. In search mode, center on the crystal for data collection and set the eucentric height by changing the α -tilt between the desired angles for recording data and adjusting with the Z-height.
2. Insert the selected area aperture and check the α -tilt again.
3. Set the α -tilt value of the stage to the starting value for data collection.
4. Blank the beam and go to exposure mode.
5. Insert the beam stop.
6. Set the rotation to continuous and the rotation speed to $0.3^\circ/\text{s}$ in Delphi.
7. Set the final tilt value that will be used.
8. Lift the screen and unblank the beam.
9. Press acquire and start the rotation to record data.

3.7 Data Processing

MicroED data are most commonly processed with the CCP4 programs, XDS, SHELX, and Phenix although other programs, such as HKL-2000/3000 and DIALLS, can be used as well. For the processing of MicroED data, MRC or TVIPS file format has to be converted to SMV. The conversion tools necessary are available on the Gonen lab webpage <https://cryoem.ucla.edu/downloads>.

3.7.1 Indexing and Integration with iMosflm

1. Start iMosflm and select processing options from the settings menu. Go to the indexing tab and uncheck automatically index after spot finding.
2. Select add images from the session menu. Load the dataset by double-clicking one of the images.
3. Go to settings and experiment settings in the main iMosflm window. If the rotation rate during data collection was negative, check the reverse direction of spindle rotation box in the experiment tab.
4. Go to the detector tab and set the gain and ADC offset according to the detector.
5. Go to settings and processing options. In the integration section of the advanced integration tab, change null pixel threshold to -1 .
6. In the image display window, adjust the red, green, and blue masks (*see Note 18*).

7. Go to the indexing task. Choose a set of images spanning a wedge of approximately 10° – 20° by adding them in the image window (*see* **Note 19**).
8. Initially, use the automatic estimation for mosaicity and press index. Then change it to make sure, there are no full reflections in the integration.
9. Go to the integration task. Change the filename and check the fix tilt and twist boxes in the fix column.
10. Process the data by pressing process. Adjust the mosaicity and optionally the mosaic block size in the images task if needed.
11. Perform a QuickScale.

3.7.2 Indexing and Integration with XDS

1. Start xdsGUI and select the desired directory using the choose or create new folder button.
2. Go to the frame tab and click load. Go to the project folder, select an image, and click open.
3. Click generate XDS.INP to generate the initial input file for XDS.
4. Go to the XDS.INP tab to view and adjust the input file (*see* **Note 20**).
5. Adjust values for the beam center assigning the corresponding x and y values in **ORGX=x** and **ORGY=y**.
6. Set **REFINE(CORRECT)=CELL BEAM ORIENTATION AXIS**.
7. Add **OFFSET=adc** offset used for image conversion.
8. Set **ROTATION_AXIS** according to the rotation direction. For example, for forward direction the default value of **ROTATION_AXIS=1 0 0** is used.
9. Save the modified input file and click run XDS.

3.7.3 Merging and Phasing Preparation

1. Start the CCP4 interface.
2. Go to directories & project dir. and fill in the project. Browse for the data to be used in the “uses directory” field.
3. Download the coordinates and structure factors for the molecular replacement search model, i.e., PDB-formatted coordinates and CIF-formatted structure factors. Move the downloaded files to the project directory.
4. Convert the CIF structure factors to MTZ file format. Go to the reflection data utilities task in the CCP4 interface and choose convert to/modify/extend MTZ. Choose import

reflection file in mmCIF format and create MTZ file. Set the in path to that of the downloaded CIF file. Change the out path and choose run now from the run drop-down menu.

3.7.4 Merging iMosfilm Processed Data with AIMLESS

1. Go to the data reduction and analysis task in the CCP4 interface and select symmetry, scale, merge (aimless).
2. Set the path of HKLIN #1 to the output MTZ file from the integration step described above using the browse button.
3. Set the project, crystal, and dataset names.
4. Check to ensure unique data and add FreeR column for 0.05 fraction of the data box.
5. Check copy FreeR from another MTZ box.
6. Set the MTZ with FreeR to the MTZ file created previously from the CIF file (see Subheading 3.7.3) using the browse button.
7. Click run now from the run drop-down menu.

3.7.5 Merging XDS Processed Data with XSCALE

1. Go to the XSCALE tab in XDS. Add an **INPUT_FILE** line with the path to **XDS_ASCII.HKL** for each additional dataset to be included.
2. If the data will be phased by molecular replacement, convert the merged intensities to MTZ file format. Go to the XDSCONV tab and set **OUTPUT_FILE** to, e.g., **temp.hkl CCP_I + F**. Click run XDSCONV to produce temp.mtz in the project directory.
3. To add the free flags from the molecular replacement search model, open CCP4 and go to reflection data utilities and then merge MTZ files (cad). Use the browse buttons to give the merged MTZ as the first file. Click add input MTZ file to add the MTZ from the search model as the second file. Use all columns for the merged data and selected columns (R-free-flags) for the search model's data.

3.7.6 Molecular Replacement Using MOLREP

1. Start the CCP4 interface.
2. Go to the molecular replacement task and select run Molrep - auto MR.
3. Set data to the MTZ file with the merged intensities from Subheading 3.7.4 or Subheading 3.7.5. Set model to the coordinate file downloaded in Subheading 3.7.3. Set solution to sample_name.pdb.
4. If the space group was not defined during any of the previous steps, click on search options and set SG to use Laue class instead of the default as is. The correct space group should

then be found during molecular replacement. Then rerun MOLREP setting SG to use the correct space group.

5. Click run now from the run drop-down menu.
6. When MOLREP has finished, inspect the log file by double-clicking the job in the central panel of the main CCP4i window. Check the contrast (*see Note 21*).

3.7.7 Molecular Replacement Using Phaser

1. Start the Phenix graphical user interface.
2. Go to new project button and set the project ID and select the project directory.
3. Go to molecular replacement in the right part of the frame, and select Phaser-MR (simple one-component interface).
4. Add the downloaded search model file (*see Subheading 3.7.3*) and the MTZ data file to search (*see Subheadings 3.7.4 and 3.7.5*) in the input files tab.
5. Go to the search options tab and click on other settings. Set scattering form factor type to use electron scattering.
6. If the space group was not defined during any of the previous steps, try alternative space group drop down menu and select “all possible”. The correct space group should be identified during molecular replacement.
7. Click the run icon and select **Run Now**.
8. When Phaser is finished, check the translation function Z-score (TFZ) in the run status tab (*see Note 22*).

3.7.8 Refinement Using REFMAC

1. Start the CCP4 interface.
2. Go to the refinement task and choose run refmac5. Assign paths to the MTZ and PDB input files using the browse buttons, and set where the MTZ out and PDB out files are to be written.
3. Select Run&View Com File from the run drop-down menu. A window will open where the command line arguments and the input script can be edited. To enable electron-scattering factors, open a new line before any keyword and enter source EC MB. Click **Continue** without display.

3.7.9 Refinement with phenix.refine

1. Start the Phenix graphical user interface.
2. Choose refinement in the right part of the frame, and select phenix.refine.
3. Go to the input data tab. Add the molecular replacement solution (*see Subheadings 3.7.6 and 3.7.7*) and the data to refine against (*see Subheadings 3.7.4 and 3.7.5*).

4. Go to the refinement settings tab and set the scattering table to electron in the other options section.
5. Click Run. When refinement is done, inspect the R-factors. After the first round of refinement, they may be high.

4 Notes

1. The intensity and time used have to be optimized for each sample but a good starting point is 1–10 s, at the lowest intensity setting.
2. Ensure that bead and tube temperatures are stabilized at sample temperature before starting. The speed and time used might have to be optimized depending on sample.
3. The lamella will usually be thinner than what is indicated in the settings; therefore, we usually set the lamella thickness to ~350–300 nm to obtain a final thickness of ~200 nm.
4. Uranyl formate is light sensitive, keep dark.
5. Keep tube pointed away from you, might splash.
6. The uranyl solution should be a light yellow color. It can be kept for 4–5 days. After this, it will begin to precipitate and should be discarded.
7. Do not dip the entire grid into the drop which will disturb the carbon coating.
8. During grid preparation, make sure the level of liquid nitrogen does not fall too low and fill the reservoir with additional liquid nitrogen whenever necessary.
9. The temperature and humidity for blotting has to be optimized for each sample, 17 °C and 50% humidity are good starting points.
10. Be careful, liquid ethane is explosive. Make sure to include the aluminum bridge, it will transfer heat from the ethane to the liquid nitrogen.
11. The ethane needs to be almost frozen, not just a liquid. Before plunging the grid, check that the top of the ethane is not completely frozen.
12. Good starting points are blotting force of 1 and blotting time of 5 s. However, it is recommended to try a range of blotting times, for example 2–12 s, to obtain the optimal ice thickness. Alternatively, the grids can be blotted manually from the copper side using a filter paper strip. The filter paper strip is folded into a tear drop shape and held with the tweezers by its thinner end. The grid is then blotted gently with the thicker end.

13. The radius of the diffraction ring can be obtained by measuring the radius in pixels and then multiplying it by the pixel size of the detector.
14. The microscope is operated in low-dose mode, meaning a low number of electrons per \AA^2 per second will minimize exposure of the sample to damaging radiation.
15. The settings are different for every microscope and depend on what the target dose is. We usually set the camera distance to 6.9 m, spot size to 9 and C2 to 100 μm in search mode.
16. The detector should be close enough to capture all high-resolution data, yet at a distance long enough to keep the reflections well separated.
17. Use the biggest aperture and the lowest magnification possible and the largest spot size. We usually set the spot size to 11, magnification to LM155X and the C2 condenser to 100 μm for collecting an atlas.
18. The red mask should be centered on the halo around the circular part of the beam stop shadow. The red rectangle corresponds to the area used for estimation of the image background and should not cover any areas that deviate significantly from the overall image, such as the beam stop shadow. The green mask should be adjusted so it covers the shadow of the beam stop. The blue mask indicates the resolution limits. For initial assessment of a high-resolution dataset, it may be set to the edge of the image, whereas for final integration, it may be dragged off the image entirely.
19. The number of spots in the wedge should be in the hundreds. If the wedge is too small, autoindexing might fail. If the wedge is too large, experimental errors may accumulate and prevent successful indexing.
20. The input file is a plain text file with keywords and values separated by an equals (=) sign. Any text following an exclamation mark (!) is a comment that will be ignored.
21. According to the MOLREP manual page [<http://www.ccp4.ac.uk/html/molrep.html#score>], a contrast greater than 3 means that MOLREP has found a solution.
22. According to the PhaserWiki [http://www.phaser.cimr.cam.ac.uk/index.php/Molecular_Replacement], a value greater than 8 means Phaser has found a solution.

Acknowledgments

We thank all members of the Gonen laboratory, current and past and all collaborators who worked with us on MicroED applications. This work was supported by the National Institutes of Health P41GM136508. The Gonen laboratory is supported by the Howard Hughes Medical Institute.

References

1. Koning RI, Koster AJ, Sharp TH (2018) Advances in cryo-electron tomography for biology and medicine. *Ann Anat* 217:82–96
2. Glaeser RM (2019) How good can single-particle Cryo-EM become? What remains before it approaches its physical limits? *Annu Rev Biophys* 48:45–46
3. Fromm SA, Sachse C (2016) Chapter twelve - Cryo-EM structure determination using segmented helical image reconstruction. In: Crowther RA (ed) *The resolution revolution: recent advances in cryoEM*, vol 579. Academic Press, pp 307–328
4. Righetto RD, Biyani N, Kowal J et al (2019) Retrieving high-resolution information from disordered 2D crystals by single-particle cryo-EM. *Nat Commun* 10:1722
5. Nannenga BL, Gonen T (2019) The cryo-EM method microcrystal electron diffraction (MicroED). *Nat Methods* 16:369–379
6. Iancu CV, Tivol WF, Schooler JB et al (2006) Electron cryotomography sample preparation using the Vitrobot. *Nat Protoc* 1:2813–2819
7. Liu S, Gonen T (2018) MicroED structure of the NaK ion channel reveals a Na⁺ partition process into the selectivity filter. *Commun. Biol* 1:1–6
8. De La Cruz MJ, Hattne J, Shi D et al (2017) Atomic-resolution structures from fragmented protein crystals with the cryoEM method MicroED. *Nat Methods* 14:399–402
9. Jones CG, Martynowycz MW, Hattne J et al (2018) The CryoEM method MicroED as a powerful tool for small molecule structure determination. *ACS Cent Sci* 4:1587–1592
10. Ting CP, Funk MA, Halaby SL et al (2019) Use of a scaffold peptide in the biosynthesis of amino acid-derived natural products. *Science* 365:280–284
11. Chapman HN (2019) X-ray free-electron lasers for the structure and dynamics of macromolecules. *Annu Rev Biochem* 88:35–58
12. Nannenga BL, Shi D, Leslie AG, Gonen T (2014) High-resolution structure determination by continuous-rotation data collection in MicroED. *Nat Methods* 11:927–930
13. Battye TG, Kontogiannis L, Johnson O et al (2011) iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr* 67:271–281
14. Leslie AG, Powell HR (2007) Processing diffraction data with mosflm BT. In: Read RJ, Sussman JL (eds) *Evolving methods for macromolecular crystallography*. Springer Netherlands, pp 41–51
15. Kabsch W (2010) XDS. *Acta Crystallogr Sect D* 66:125–132
16. Otwinowski Z, Minor WB (1997) Processing of X-ray diffraction data collected in oscillation mode. In: *Macromolecular crystallography part a*, vol 276. Academic Press, pp 307–326
17. Waterman DG, Winter G, Parkhurst JM et al (2013) The DIALS framework for integration software. *CCP4 Newsl PROTEIN Crystallogr* 49:16–19
18. Brünger AT et al (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr Sect D* 54:905–921
19. Brunger AT, Adams PD, Clore GM et al (2007) Version 1.2 of the crystallography and NMR system. *Nat Protoc* 2:2728–2733
20. Adams PD, Afonine PV, Bunkóczi G et al (2010) PHENIX: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallogr Sect D* 66:213–221
21. Blanc E, Roversi P, Vonrhein C et al (2004) Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallogr Sect D* 60:2210–2221
22. Sheldrick GM (2010) Experimental phasing with SHELXC: combining chain tracing with density modification. *Acta Crystallogr Sect D* 66:479–485
23. Winn MD, Ballard CC, Cowtan KD et al (2011) Overview of the CCP4 suite and

- current developments. *Acta Crystallogr D Biol Crystallogr* 67:235–242
24. Nannenga BL, Gonen T (2014) Protein structure determination by MicroED. *Curr Opin Struct Biol* 27:24–31
 25. Nanneng BL, Shi D, Hattne J et al (2014) Structure of catalase determined by MicroED. *elife* 3:e03600
 26. Purdy MD, Shi D, Chrustowicz J et al (2018) MicroED structures of HIV-1 gag CTD-SP1 reveal binding interactions with the maturation inhibitor bevirimat. *Proc Natl Acad Sci U S A* 115:13258–13263
 27. Rodriguez JA, Ivanova MI, Sawaya MR et al (2015) Structure of the toxic core of α -synuclein from invisible crystals. *Nature* 525:486–490
 28. Sawaya MR, Rodriguez J, Cascio D et al (2016) Ab initio structure determination from prion nanocrystals at atomic resolution by MicroED. *Proc Natl Acad Sci U S A* 113:11232–11236
 29. Vergara S, Lukes DA, Martynowycz MW et al (2017) MicroED structure of au 146 (p-MBA) 57 at subatomic resolution reveals a twinned FCC cluster. *J Phys Chem Lett* 8:5523–5530
 30. Mahamid J, Pfeffer S, Schaffer M et al (2016) Visualizing the molecular sociology at the HeLa cell nuclear periphery. *Science* 351:969–972
 31. Kasinath V, Faini M, Poepsel S et al (2018) Structures of human PRC2 with its cofactors AEBP2 and JARID2. *Science* 359:940–944
 32. von der Ecken J, Müller M, Lehman W et al (2015) Structure of the F-actin–tropomyosin complex. *Nature* 519:114–117
 33. Gonen T, Cheng Y, Sliz P et al (2005) Lipid–protein interactions in double-layered two-dimensional AQP0 crystals. *Nature* 438:633–638
 34. Xu H, Lebrette H, Clabbers MT et al (2019) Solving a new R2lox protein structure by microcrystal electron diffraction. *Sci Adv* 5: eaax4621

INDEX

A

- Antibody
 - conjugate 176
- Autographa californica* 142

B

- Baculovirus
 - expression 83, 108, 131, 137, 141, 142
 - genome engineering 141–150

C

- Cell-free 175–189
- Cells
 - adherent 83, 92, 100, 102, 108, 303, 308, 314
 - high five 129–139
 - insect 84, 106, 107, 130, 131, 138, 141, 193
 - K562 155, 156, 163, 165, 166, 172
 - mammalian 83–103, 107, 108, 125, 130, 133, 138, 141, 193, 194, 303, 308, 314
 - yeast 193, 303
- CRISPR/cas9 155, 159, 162, 163
- Critical assessment of techniques for protein structure
 - prediction (CASP) 13, 24, 31, 34, 36, 38–41, 43, 44, 46, 49
- Cryo-electron microscopy
 - single particle 257–286
- Cryo-focused ion beam milling
 - (cryo-FIBM) 302, 303, 308, 310, 315
- Crystallography
 - time resolved 17, 206, 209

D

- Data
 - archiving 5–10, 17
 - collection 69, 70, 107, 203, 204, 208–219, 229, 230, 239, 240, 247–250, 252, 253, 291–298, 301, 302, 319, 323–325, 327, 334, 335
 - processing 117, 218, 229, 230, 247–252, 257–286, 293, 294, 324, 325, 327, 335–339
 - standards 5, 6, 8–10, 230, 325, 326
- Detergents 107, 110, 113, 120–125, 173
- Docking
 - ligands 44, 55–57
 - peptides 59–61

- programs 44–46, 54, 56, 60, 61
- protein-protein 63–64
- protein-nucleic 68–69

E

- Electron diffraction 205, 206, 303, 304, 317, 323, 332, 333
- EM grids 310, 311, 327
- Enzyme
 - catalysis 204, 216
- Expi293F 105–126, 130

F

- Flow cytometry 84, 95, 96, 100, 102

G

- GPCR 106, 107, 109, 122

H

- His-tag 123, 133, 139, 195–197

I

- Image processing 15, 230, 247, 248, 258, 272, 277, 285, 292, 293
- Immunopurification 168–172
- IntFOLD 28, 34, 38–39
- Isotope 195, 196
- I-Tasser 28, 39–41, 47, 48

L

- Lamella
 - protein crystal 315–317
 - cellular 315–317
- Lentivirus 83–103

M

- Membrane proteins v, 58, 95, 98, 101, 105–126, 204, 212, 323
- Metalloenzymes 217, 218
- Microcrystal 204, 206–208, 210, 214–217, 219, 324–327, 329, 331, 334
- MicroED 323–341
- Microtubule 71, 193–200

Modeling

- ab initio 26, 36–38, 41, 43
- predictive 37
- protein-peptide 58–61
- protein-nucleic 54, 65–69
- template-based 24, 27, 29–38, 46

N

- Nanocrystal 206, 323, 326, 328, 329
- NMR spectroscopy
 - solid-state 193–200
- Non-canonical amino acids (NCAA) 175–189

O

- Oligonucleotides 114, 154–157, 159–165

P

- Phase contrast 291
- Photosensitive 210–215
- Phyre2 27, 28, 40, 41, 47
- Protein
 - classification 24–26, 40, 269
 - complexes 3, 46, 53–74, 86, 87, 130, 138, 153, 154, 160, 170, 173, 193, 215, 301, 302, 324, 325
 - crystal 43, 46, 58–60, 203, 204, 207, 208, 215, 302, 303, 310, 314, 315, 323, 324
 - data bank 3–18, 24–26, 106, 203, 204
 - dynamics 31, 61, 194, 204, 215
 - expression 83–87, 92, 97–99, 102, 103, 114, 117, 129, 130, 133, 138, 153, 160, 168, 193
 - interactions 23–49, 53, 58, 60, 61, 65, 87, 194
 - modelling 24, 26, 27, 30, 31, 34, 35, 40, 43, 45, 53–73
 - production 83–86, 102, 105–109, 117, 130, 131, 141, 142, 155, 194, 207, 316
 - purification 124, 125, 133, 139, 153, 155, 157, 159, 160, 163, 194, 196, 197
- Proteinase K 303, 304, 310–312, 319, 325, 326

R

- Red/ET recombination 147
- Ribosome 41, 71, 87, 296
- Robetta 28, 36, 37, 41, 42, 47

S

- Saccharomyces cerevisiae* 303, 304, 308, 310, 311, 314, 316, 319
- Sample
 - delivery 84, 206, 208, 209, 213, 216
 - manipulation 115, 312
 - preparation 84, 189, 196, 215, 229–238, 257, 302, 303, 308
- Scipion 258, 266, 267, 269, 272, 273, 281, 283, 285
- Sequence
 - alignment 25, 27, 29, 35, 36, 41, 42, 60, 62, 65
- Strep-tag 123, 124, 177, 182
- Structure
 - archives 3–5, 8–12, 14, 16, 17
 - quaternary 11, 12, 44–47
 - tertiary 24–26, 28, 36, 38, 43
- SWISS-MODEL 25, 28, 42, 43, 45

T

- Transduction 83, 84, 86, 88, 92, 94, 95, 101, 102, 107, 108
- Transient gene expression (TGE) 129–139
- Transmission electron microscope (TEM) 231, 238–240, 242, 302–304, 306–309, 311, 312, 314, 316, 319, 320, 323, 332
- tRNA 176, 178–180, 182–184

V

- Vectors 36, 88, 95, 101, 109, 110, 114, 115, 125, 130–132, 139, 141, 161, 165, 177, 178, 187, 188
- Vitrification 229, 230, 233, 234, 303, 304, 308, 310, 311, 319, 325, 331
- Volta phase plate (VPP) 231, 291–298

W

- Worldwide protein data bank (wwPDB) 4–18, 26

X

- X-ray emission spectroscopy (XES) 206, 213, 214, 217, 218
- X-ray free electron laser (XFEL) 17, 203–220