



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

DISSERTATION

Flexible fitting of proteins and nucleic acids with Gaussian Mixture Models

Thomas Mulvaney

Integrative Virology

Department of Chemistry

Faculty of Mathematics, Informatics and Natural Sciences

Universität Hamburg

Hamburg, Germany

A thesis submitted for the degree of
Doctor rerum naturalium (Dr. rer. nat.)

Flexible fitting of proteins and nucleic acids with Gaussian Mixture Models

Dissertation submitted by: Thomas Mulvaney

Date of Submission: 21.10.2024

Date of Disputation: 24.1.2025

Supervisor: Prof. Dr. Maya Topf, Universität Hamburg

Co-supervisor: Prof. Dr. Kay Grünewald, Universität Hamburg

Committee:

1st Prof. Dr. Kay Grünewald, Universität Hamburg

2nd Prof. Dr. Maya Topf, Universität Hamburg

3rd Prof. Dr. Zoya Ignatova, Universität Hamburg

4th Dr. Jan Kosinski, European Molecular Biology Laboratory (EMBL), Hamburg

5th Dr. Ben Vollmer, Universität Hamburg

Reviewers:

1st Prof. Dr. Maya Topf, Universität Hamburg

2nd Prof. Dr. Zoya Ignatova, Universität Hamburg

Universität Hamburg, Hamburg, Germany
Faculty of Mathematics, Informatics and Natural Sciences
Department of Chemistry

Integrative Virology

The work described in this thesis was conducted between 1st of October 2021 and the 11th of October 2024. The work was performed at the Centre for Structural Systems Biology (CSSB) in Hamburg and funded by the Leibniz Institute for Virology (LIV).

Abstract

A fundamental aspect of cryo-electron microscopy (cryo-EM) is deriving atomic models from the 3D reconstructions. A number of advances have been made in detector technology and image processing that has enabled some dramatic improvements in resolution. Still, low resolutions continue to plague the field. Cryo-EM reconstructions, which rely on many images of molecules with different orientations, are limited by the noise and conformational heterogeneity often inherent to the molecules of interest. Classification of these structures into distinct classes, enables insights into the conformational space that they inhabit, and can indeed improve the quality of individual reconstructions. However, even within classes, atomic motions are an unavoidable fact of nature. The inherent noise in low-dose electron imaging, along with the number of particles in the sample, limits to what degree classes can accurately be assigned. It is thus an accepted matter, that many reconstructions will by their very nature, contain features from many molecules which are only approximately identical. This manifests itself in heterogeneous resolutions, where fluctuations can be attributed to changes in the local similarity of the averaged molecules.

To build atomic models from these low resolution reconstructions, a common approach known as flexible fitting, employs atomic structures which have been solved using high-resolution techniques. In this thesis, I explore a flexible fitting and refinement method which attempts to improve the interpretability of atomic models by estimating the local resolution of the atoms in the underlying data. Using a Gaussian Mixture Model, a single atomic model is used to describe the experimental map, with atoms modelled as isotropic three-dimensional Gaussians with widths determined from the reconstruction. This allows better interpretation of the atomic structure, as coordinate uncertainty is accounted for. Of course, isomorphic Gaussians are limited in their accuracy at representing atomic motions of bonded atoms. Instead, a second representation is derived by modifying the atomic models in the context of a molecular dynamics simulation with perturbations derived from the local resolution information. This representation, composed of an ensemble of atomic structures was shown to produce an improved fit with the data.

Unfortunately, the initial model which is fitted to the experimental data, often requires significant rearrangement of its coordinates in order to fit the cryo-EM map before local resolutions can be estimated. This is because high resolution starting models are typically derived from X-ray experiments, where crystallisation can result in structural change. Frequently, cryo-EM experiments are elucidating structures which have never been seen before. In such cases, structural predictions are used instead. These can be highly accurate, often down to the level of domains, but may require some rearrangement to better fit the data. Flexible fitting approaches are able to fit such models but require the use of restraints to prevent distortions. In the past, the RIBFIND approach has been successfully used to this

end, but has been limited to protein structures. RIBFIND2, which is presented in this thesis is able to decompose RNA structures into rigid bodies which can be restrained during flexible fitting procedures.

Combining the GMM method with these RIBFIND2 restraints enabled a diverse set of structural predictions from the recent CASP15 challenge to be flexibly fit into cryo-EM maps, resulting in models with similar quality to the target structures.

Zusammenfassung

Ein grundlegender Aspekt der Kryo-Elektronenmikroskopie (cryo-EM) besteht darin, atomare Modelle aus den 3D-Rekonstruktionen abzuleiten. In der Detektor-Technologie und Bildverarbeitung wurden zahlreiche Fortschritte erzielt, die signifikante Verbesserungen in der Auflösung ermöglicht haben. Dennoch bleibt eine geringe Auflösung ein ständiges Problem in diesem Bereich. Cryo-EM-Rekonstruktionen, die auf einer großen Anzahl Bildern von Molekülen in unterschiedlichen Ausrichtungen basieren, sind begrenzt durch das Rauschen und die konformationelle Heterogenität, die die Molekülen von Interesse oft aufweisen. Die Klassifizierung dieser Strukturen ermöglicht Einblicke in den konformationellen Raum, den sie einnehmen und kann tatsächlich die Qualität der einzelnen Rekonstruktionen verbessern. Doch selbst innerhalb der einzelnen Klassen sind atomare Bewegungen unvermeidbar. Das inhärente Rauschen bei Niedrigdosis-Elektronenmikroskopie sowie die Anzahl der Partikel in der Probe begrenzen das Ausmaß, in dem Klassen präzise zugeordnet werden können. Es ist daher allgemein akzeptiert, dass viele Rekonstruktionen Merkmale vieler Moleküle enthalten, die nur annähernd identisch sind. Dies zeigt sich in heterogenen Auflösungen, bei denen Schwankungen auf Änderungen in der lokalen Ähnlichkeit der gemittelten Moleküle zurückzuführen sind.

Um aus diesen Rekonstruktionen mit niedriger Auflösung atomare Modelle zu erstellen, wird ein gängiger Ansatz namens Flexible Fitting verwendet, der atomare Strukturen nutzt, die mit hochauflösenden Techniken ermittelt wurden. In dieser Arbeit untersuche ich eine Methode zur flexiblen fitting und Verfeinerung, die versucht, die Interpretierbarkeit atomarer Modelle zu verbessern, indem die lokale Auflösung der Atome in den zugrunde liegenden Daten geschätzt wird. Mithilfe eines gaußschen Mischverteilungsmodells wird ein einzelnes atomares Modell verwendet, um die Rekonstruktion zu beschreiben. Dabei werden Atome als isotrope dreidimensionale Gauss-Verteilungen modelliert, deren Verteilungen aus der Rekonstruktion bestimmt werden. Dies ermöglicht eine bessere Interpretation der atomaren Struktur, da die Ungenauigkeit der Koordinaten berücksichtigt wird. Natürlich sind isomorphe Gauss-Verteilungen in ihrer Genauigkeit zur Beschreibung atomarer Bewegungen von gebundenen Atome begrenzt. Stattdessen wird eine zweite Repräsentation abgeleitet, indem die atomaren Modelle im Kontext einer molekulardynamischen Simulation mit den Abweichungen modifiziert werden, die aus den lokalen Auflösungsinformationen abgeleitet sind. Diese Darstellung, bestehend aus einem Ensemble atomarer Strukturen, ermöglicht eine bessere Anpassung an die Daten.

Jedoch erfordert das Ausgangsmodell, das an die experimentellen Daten angepasst wird, oft eine erhebliche Neuordnung seiner Koordinaten, um den cryo-EM-Daten zu entsprechen, bevor lokale Auflösungen geschätzt werden können. Dies liegt daran, dass hochauflösende Ausgangsmodelle typischerweise aus kristallographischen Experimenten abgeleitet sind, bei denen die Kristallisation zu strukturellen

Veränderungen führen kann. Häufig zeigen cryo-EM-Experimente Strukturen, die zuvor noch nie beobachtet wurden. In solchen Fällen werden stattdessen Strukturvorhersagen verwendet. Diese können sehr ungenau sein und erfordern oft eine gewisse Anpassung, um besser zu den experimentellen Daten zu passen. Flexible Fitting-Ansätze können solche Modelle anpassen, erfordern jedoch den Einsatz von Restriktionen, um künstliche Verzerrungen zu vermeiden. In der Vergangenheit wurde der RIBFIND-Ansatz erfolgreich zu diesem Zweck verwendet, war jedoch auf Proteinstrukturen beschränkt. RIBFIND2, das in dieser Arbeit vorgestellt wird, kann lange RNA-Strukturen in einzelne rigide Domänen zerlegen, die während flexibler Anpassungsprozesse genutzt werden können.

Die Kombination der GMM-Methode mit diesen RIBFIND2-Restriktionen ermöglichte es, eine Vielzahl von Strukturvorhersagen des CASP15-Wettbewerbes, flexibel in cryo-EM-Karten einzupassen, wodurch Modelle von ähnlicher Qualität wie die experimentell-ermittelte Struktur erzielt wurden.

Contents

List of Publications	ii
List of Figures	iii
List of Tables	iv
List of Abbreviations	v
1 Introduction	1
1.1 Cryogenic-electron microscopy	2
1.1.1 Transmission electron microscopy	2
1.1.2 Electron scattering & Coulomb maps	4
1.1.3 Single Particle Analysis	6
1.1.4 Image processing	6
1.1.5 Reconstruction	8
1.1.6 Interpretation & validation	9
1.2 Flexible fitting and refinement	10
1.2.1 Structure Prediction	11
1.2.2 Assessing quality of fit	12
1.2.3 Molecular Dynamics	13
1.2.4 Biassing potentials	16
1.2.5 Convergence & restraints	19
1.3 RNA	21
1.3.1 RNA structure	23
1.3.2 Rationally designed functional nucleic acids	24
2 Aim	26
3 Publications	27
3.1 Publication in Nature Communications	27
3.2 Publication in Nucleic Acid Research	41
3.3 Publication in Proteins: Structure, Function & Bioinformatics	51
4 Discussion	69
4.1 Overfitting and force determination	69
4.2 Membranes and other unmodelled regions	70
4.3 Interpreting TEMPy-ReFF “B-factors”	70
4.4 Supported atomic models	70
4.5 Geometry refinement	71
4.6 Modelling of atomic charges	71

Contents	i
4.6.1 Is flexible fitting still the future?	72
4.7 Conclusion	72
References	74
5 Appendix	88

List of Publications

Dissertation Related Publications

- J. G. Beton*, T. **Mulvaney***, T. Cragolini, and M. Topf (Jan. 2024). Cryo-EM structure and B-factor refinement with ensemble representation. en. In: Nature communications 15.1, p. 444. ISSN: 2041-1723. DOI: 10.1038/s41467-023-44593-1. URL: <http://dx.doi.org/10.1038/s41467-023-44593-1>.
- S. Malhotra*, T. **Mulvaney***, T. Cragolini, H. Sidhu, A. P. Joseph, J. G. Beton, and M. Topf (Sept. 2023). RIBFIND2: Identifying rigid bodies in protein and nucleic acid structures. en. In: Nucleic acids research. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkad721. URL: <http://dx.doi.org/10.1093/nar/gkad721>.
- T. **Mulvaney**, R. C. Kretsch, L. Elliott, J. G. Beton, A. Kryshchuk, D. J. Rigden, R. Das, and M. Topf (Dec. 2023). CASP15 cryo-EM protein and RNA targets: Refinement and analysis using experimental maps. en. In: Proteins 91.12, pp. 1935–1951. ISSN: 0887-3585, 1097-0134. DOI: 10.1002/prot.26644. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26644>.

(*) indicates co-first authorship

Other Publications

- J. G. Beton, T. Cragolini, M. Kaleel, T. **Mulvaney**, A. Sweeney, and M. Topf (Nov. 2022). Integrating model simulation tools and cryoelectron microscopy. en. In: Wiley interdisciplinary reviews. Computational molecular science. ISSN: 1759-0876, 1759-0884. DOI: 10.1002/wcms.1642. URL: <https://onlinelibrary.wiley.com/doi/10.1002/wcms.1642>.
- R. Das, R. C. Kretsch, A. J. Simpkin, T. **Mulvaney**, P. Pham, R. Rangan, F. Bu, R. M. Keegan, M. Topf, D. J. Rigden, Z. Miao, and E. Westhof (Oct. 2023). Assessment of three-dimensional RNA structure prediction in CASP15. en. In: Proteins. ISSN: 0887-3585, 1097-0134. DOI: 10.1002/prot.26602. URL: <http://dx.doi.org/10.1002/prot.26602>.
- L. R. Genz, T. **Mulvaney**, S. Nair, and M. Topf (Nov. 2023). PICKCLUSTER: a protein-interface clustering and analysis plug-in for UCSF ChimeraX. en. In: Bioinformatics 39.11. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btad629. URL: <http://dx.doi.org/10.1093/bioinformatics/btad629>.
- A. Sweeney, T. **Mulvaney**, M. Maiorca, and M. Topf (Jan. 2024). ChemEM: Flexible Docking of Small Molecules in Cryo-EM Structures. en. In: Journal of medicinal chemistry 67.1, pp. 199–212. ISSN: 0022-2623, 1520-4804. DOI: 10.1021/acs.jmedchem.3c01134. URL: <http://dx.doi.org/10.1021/acs.jmedchem.3c01134>.

List of Figures

1.1	Simplified schematic of TEM microscope	3
1.2	Comparison of X-ray and electron scattering curves	5
1.3	Vitrification captures ensembles	7
1.4	Resolution trends and model building	11
1.5	Improvements in contact prediction in CASP15	12
1.6	Local minima: An optimisation nightmare	17
1.7	Comparison of density and correlation force fields in 2D.	18
1.8	Recent growth of cryo-EM in nucleic acid structure elucidation	25

List of Tables

1.1	Table of biasing potentials	19
1.2	Table of molecular dynamics based refinement approaches.	22

List of Abbreviations

ADF	Atomic Displacement Factors
ART	Algebraic reconstruction technique
B	BFactor, see ADF.
CCC	Cross-correlation coefficient
CDMD	Correlation Driven Molecular Dynamics
CERES	the Cryo-EM re-refinement system .
Cryo-EM	cryogenic electron microscopy
Cryo-ET	cryogenic electron tomography
EMDB	Electron Microscopy Data Bank
EMPIAR	Electron Microscopy Public Image Archive
EM	Electron Microscopy
FBP	Filtered back projection
FSC	Fourier Shell correlation
GMM	Gaussian Mixture Model
IAM	Independent Atom Model
K	Kelvin
keV	kiloelectron volt
MD	Molecular Dynamics
MDFF	Molecular Dynamics Flexible Fitting
ML	Machine learning
PDB	Protein Data Bank
PHENIX	Python-based Hierarchical ENvironment for Integrated Xtallography
RIBFIND	RI gid B ody FIND er
RNA	Ribonucleic acid
SMOC	Segmented Manders Overlap Coefficient
SPA	Single Particle Analysis
SVA	Subvolume averaging
TAAM	Transferable Aspherical Atomic Model
TEM	Transmission Electron Microscopy

- TEMPy** Transmission Electron Microscopy Python library
- TEMPy-ReFF** **TEMPy** **R**esponsibility based **F**lexible **F**itting

1

Introduction

Contents

1.1	Cryogenic-electron microscopy	2
1.1.1	Transmission electron microscopy	2
1.1.2	Electron scattering & Coulomb maps	4
1.1.3	Single Particle Analysis	6
1.1.4	Image processing	6
1.1.5	Reconstruction	8
1.1.6	Interpretation & validation	9
1.2	Flexible fitting and refinement	10
1.2.1	Structure Prediction	11
1.2.2	Assessing quality of fit	12
1.2.3	Molecular Dynamics	13
1.2.4	Biassing potentials	16
1.2.5	Convergence & restraints	19
1.3	RNA	21
1.3.1	RNA structure	23
1.3.2	Rationally designed functional nucleic acids	24

Building atomic models from medium-resolution data remains a fundamental challenge of cryo-EM. At this range of resolution, existing structural models or computational predictions are used as starting points for a process called flexible fitting. One of the problems of these approaches is that while they may fit the atomic coordinates, they do not measure the local resolution, which may be an indicator of local flexibility. The second problem is that these approaches typically fail to converge on well-fitting models when the starting model is conformationally distant. Throughout this PhD, I have worked on methods for flexible fitting which avoid these limitations. The main result has been the development of a Gaussian Mixture Model based force field which uses expectation-maximisation to both improve the fit of the atomic coordinates and simultaneously determine the local resolution of each atom. This has resulted in a series of papers focused on fitting

and refining computational models into cryo-EM data. Each paper is dedicated to specific aspects of the fitting and refining problem, but they assume some prior knowledge of these topics already. In this chapter, I describe the background theory of cryo-EM single-particle analysis which is missing from these papers, the various ways in which we model aspects of the microscope and the molecules they image, as well the many shortcuts, assumptions, and well-known problems which are found in the field. To conclude, I highlight the role of cryo-EM in RNA structural biology where it is becoming increasingly applicable to the study of small (sub 200kDa) RNA molecules which had previously been off limits.

1.1 Cryogenic-electron microscopy

Cryogenic-electron microscopy (cryo-EM) has emerged as a dominant technique for studying a broad range of bio-molecules that have been inaccessible to methods such as X-ray crystallography and nuclear magnetic resonance imaging (NMR). Important molecules such as membrane proteins, which play a vital role in cellular organisms (and enveloped viruses), have been structurally determined using cryo-EM. Recent technological advances have pushed the resolution towards that of X-ray crystallography for certain molecules, opening the doors to image ligands and design drugs based on structural details. The, so called “resolution revolution” (Kühlbrandt 2014), has broken many theoretical barriers such as the 200kDa limit, allowing smaller molecules to be studied, including small RNA molecules.

Cryo-EM may refer to several transmission electron microscopy (TEM) based imaging approaches applied in the context of imaging cryogenically frozen samples. The first of the two main approaches is single-particle analysis (SPA), which is currently the best approach for achieving high-resolution images of molecules. The second approach is cryo-electron tomography (cryo-ET), where samples are rotated over a limited set of angles to produce a tilt-series that can be reconstructed, as per medical tomography. The latter approach does not currently afford the same levels of resolution, but can be used to study molecules within cells. The focus of the publications in this thesis are on methods for fitting atomic models to data obtained using the SPA method. For the rest of this thesis, the abbreviation cryo-EM will refer to this method, unless otherwise specified. What follows, is a brief overview focussing on critical aspects of SPA and the surrounding theory. For a broader introduction (Glaeser, Nogales, et al. 2021) is an excellent starting point.

1.1.1 Transmission electron microscopy

In the early 20th century it was realized that the wavelength of light was a limiting factor of the resolution obtainable using traditional light microscopy. Electron microscopes were developed in the 1930s, which instead took advantage of the significantly shorter wavelengths of high-energy electrons. The technique exists in three main modalities: transmission electron microscopy (TEM), scanning electron microscopy (SEM) and scanning transmission electron microscopy (STEM).

In TEM, the approach used in cryo-EM, electrons are fired at a thin sample from an electron gun (fig. 1.1A), a coherent source of electrons between 100 and 300keV depending on the microscope. The wavelength of an electron travelling

with a velocity v and mass m_e (accounting for relativistic effects due to velocity), is given by de Broglies formula (eq. 1.1), where \hbar is the Planck constant.

$$\lambda = \frac{\hbar}{m_e v} \quad (1.1)$$

Thus, at 300keV the wavelength of an electron is approximately 20 picometers, $1/50^{\text{th}}$ of an angstrom. In theory, this offers the ability to easily image at atomic resolutions, but as will be described, a number of factors make attaining atomic resolutions challenging.

Upon reaching the thin sample, incident electrons may pass straight through the sample (termed “direct” or “unscattered”). They may be elastically scattered due to interactions with positively charged nuclei of the atoms in the samples. The electrons may also impart some of their energy to the sample before being scattered (inelastic scattering). These events are depicted in the blow-up in fig. 1.1B. Inelastically scattered electrons damage the sample, as the energy imparted causes bonds to break, electrons to be ejected, heating and charging of the sample and a host of other errors which leads to a net loss of information.

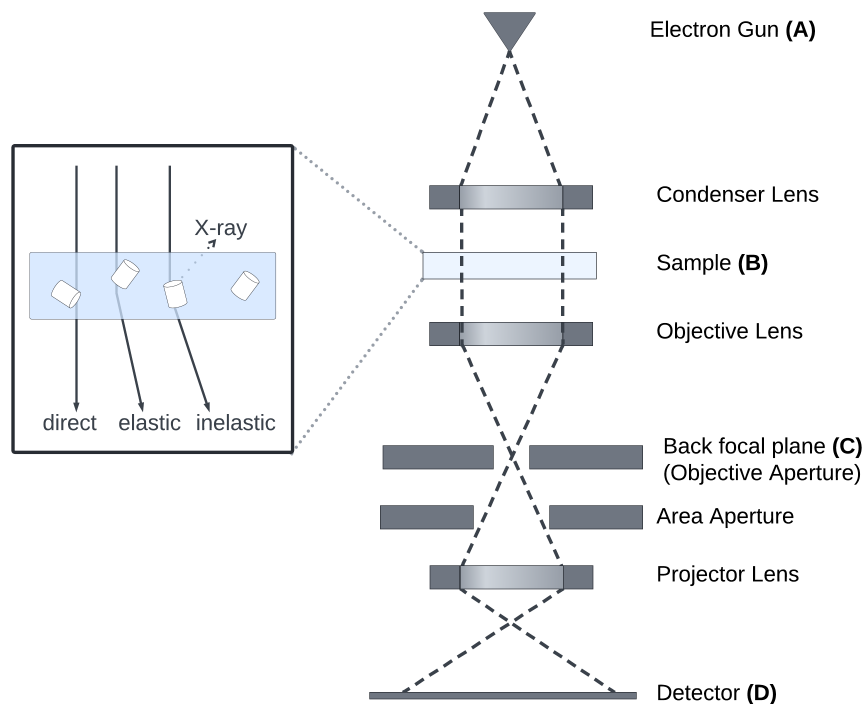


Figure 1.1: A simplified schematic representation of a transmission electron microscope showing key components. Electrons leave the electron gun (A) and pass through the condenser lens which focuses the beam on the sample (B). The electron interactions with the sample can be classified as direct, elastic and inelastic. Direct electrons pass directly through without interacting with the sample. Elastically scattered electrons are scattered but impart no energy to the sample. Inelastic electrons impart energy to the sample causing damage. The fourier transform of the coulombic potential is focused on the back focal plane of the microscope (C). The image formed on the detector (D) is a projection of the samples coulombic potential.

Because the density of biological molecules is not significantly different from the surrounding amorphous ice, and the majority of electrons pass directly through the sample, amplitude based imaging is not possible. Phase based imaging, which relies on interference between the direct and the elastically scattered electrons is used instead. Due to biological molecules being “weak phase objects”, phase contrast imaging requires biomolecules to be significantly out of focus. By defocusing the sample, the low-frequency components are enhanced whilst high-frequencies are attenuated. Some of the corruptions due to the defocus can be corrected for during the image processing (section 1.1.4).

Inelastically scattered electrons, not only cause damage to the sample, but will also interfere with the construction of phase images as they no longer have the same wavelength. They are thus filtered out using an energy filter. In order to limit damage to the sample a low dose of electrons is used (measured in electrons per unit area). In the past, film and CCD based detectors had a high background noise which meant the overall signal-to-noise ratio (SNR) was severely impacted by these low doses. The “resolution revolution” is largely attributed to the direct detector device (DDD) (Kühlbrandt 2014) which have much lower background noise levels making low electron doses less of a hindrance.

1.1.2 Electron scattering & Coulomb maps

Both X-rays and electrons are scattered by atoms in the sample. In the case of X-rays, scattering is determined by the electron density about the nuclei of atoms in the object of interest. The scattering of X-rays due to the electron cloud of an atom is approximated by the atomic form factor $f_x(s)$. Electrons on the other hand, are scattered by both the positively charged nucleus and the surrounding electron cloud. The images produced by electron scattering are thus a representation of the coulombic potential of the sample. The Mott-Bethe equation, $f_e(s)$ gives the electron scattering form factors of an atom due to its nuclear charge Z (its atomic number) and the shielding (hence subtraction) by the electron cloud, in terms of the X-ray atomic form factor $f_x(s)$:

$$f_e(s) = \frac{1}{8\pi^2 a_0} \frac{Z - f_x(s)}{s^2} \quad (1.2)$$

where s is given by the wavelength λ of the incident electron and the scattering angle θ :

$$s = \frac{\sin(\theta)}{\lambda} \quad (1.3)$$

The X-ray atomic form factors $f_x(s)$, are derived from X-ray diffraction experiments or computational approaches (e.g. Hartree-Fock method). They are usually approximated using a well established sum of Gaussians approach (Doyle et al. 1968) which has seen numerous updates in tabulated values and additional Gaussian terms (Peng et al. 1996; Peng 1998; Peng 1999; Yonekura et al. 2018) This model, known as Independent Atom Model (IAM), is based on the assumption that the atoms are not bonded.¹

¹More accurate methods which attempt to take into account multipole electron distributions exist such as TAAM (Kulik et al. 2022) although they are yet to be widely employed.

The X-ray scattering factors $f_x(s)$, are thus defined where a_k and b_k are parameters for the k^{th} Gaussian. An additional constant c may sometimes be included.

$$f_x(s) = \sum_{k=0}^K a_k e^{-b_k s^2} + c_k \quad (1.4)$$

There is an important difference in the atomic form factors for X-ray and electron scattering which are illustrated in figure 1.2. The X-ray scattering factors for neutral and charged atoms do not differ significantly. Electron scattering factors on the other hand, deviate significantly at low scattering angles when the atom has a partial charge. In the case of positive charges, the low scattering angle amplitudes are increased, while negative charges attenuate the low scattering amplitudes.

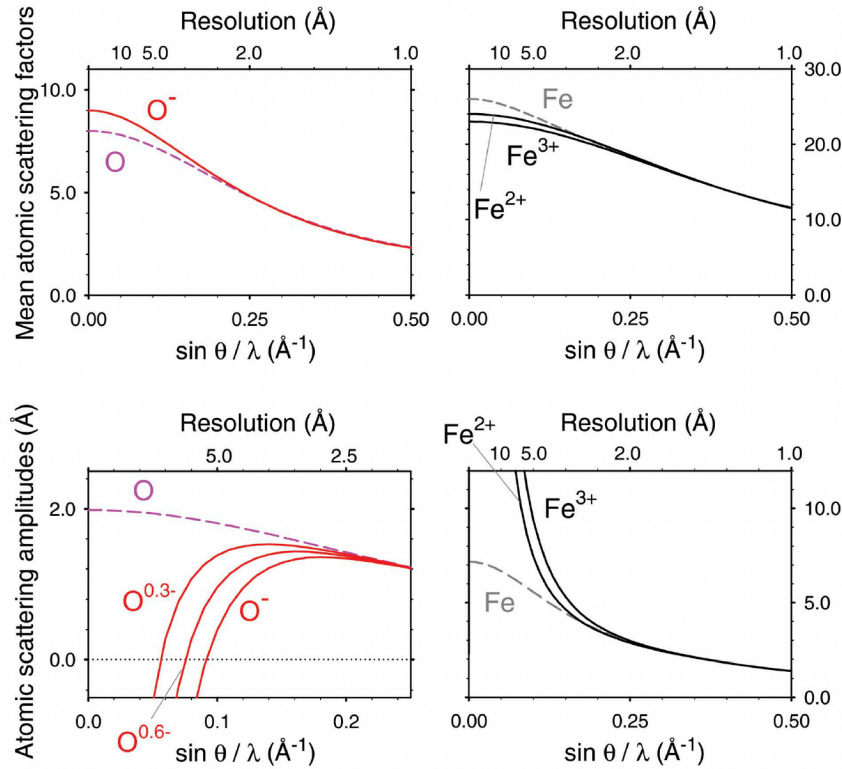


Figure 1.2: Scattering curves adapted from Yonekura et al. 2018 highlight the differences between X-ray scattering (top row) and the amplitude of electron scattering (bottom row) for oxygen and iron atoms of different charges. At low-resolutions the effects of partial charge causes strong deviations from the neutral atomic scattering amplitudes.

The scattered electrons form a diffraction pattern at the back focal plane of an electron microscope (fig. 1.1C). The image formed on the detector is a projection through the sample and thus contains contributions from many layers of atoms, flattened into a 2D image. The inverse Fourier transform, \mathcal{F}^{-1} of the Mott-Bethe equation gives the 3D coulomb potential ρ of the molecule. The 2D images are thus projections of this.

$$\rho = \mathcal{F}(f_e(s)) \quad (1.5)$$

1.1.3 Single Particle Analysis

Cryo-EM single particle analysis (SPA) allows molecules to be imaged in near native conditions and at the same time overcomes many of the apparent limitations TEM imposes when it comes to working with biological molecules. Plunge freezing methods allow biological molecules to be frozen in a amorphous ice (Adrian et al. 1984). This is important for two reasons: the sample must be placed in the vacuum chamber of the electron microscope, which would normally cause any water in the sample to vaporise. Secondly, the molecules need to be fixed to prevent their movement. In theory, the molecules of interest do not have time to settle in lower energy conformations, thus capturing them in near-native states (fig. 1.3). Indeed, vitrification of ribosomes from different starting temperatures has shown this to be true in practice (Fischer et al. 2010). Importantly, water also does not have time to relax and form crystalline ice which would cause strong patterns in the images. Instead, amorphous ice is formed, which fixes the molecules in place, and contributes weakly to the images.

The limitations of a low signal-to-noise ratio imposed by requiring low electron doses - which would otherwise make the endeavor of high-resolution biomolecular imaging futile - are overcome by the fact that the vitrified samples may contain millions of molecules in similar conformations (“single” particles). Despite their low signal-to-noise ratio and information loss due to radiation damage, the images can be aligned and averaged, boosting the overall signal (Rosenthal and Henderson 2003). Assuming radiation damage is uniformly distributed, these effects are also averaged out.

Finally, the thin frozen samples potentially contain molecules with a variety of orientations. This final point, allows SPA techniques to produce 3D reconstructions as per classic tomography (section 1.1.5). Samples containing molecules with a well-distributed set of orientations are not always obtainable as many molecules have strong preferences to air-water interfaces. This “preferred orientation” problem is a major challenge. Thin ice, which is desirable for having a lower background, can come at the cost of increasing the preferred orientation. Some of the preferred orientation issues can also be counteracted by tilting the sample, but this has the downside of increasing ice thickness at non-zero tilt angles (Tan et al. 2017).

1.1.4 Image processing

Besides having an improved signal-to-noise ratio compared to the older CCD and photographic film-based detectors, DDDs can record movies rather than one long exposure. Each movie is a set of images (or “frames”). Different detectors have different frame rates, which determines the exposure time of each frame.

This has led to some important innovations when it comes to post-processing. Images with a certain dose of electrons can be generated by averaging the appropriate number of “movie frames”. Initial frames can be dropped due to charge build-up and bulge (Brilot et al. 2012). Similarly, dose weighting (also called exposure filtering) can be applied across the images such that later exposures contribute fewer high-resolution details (Grant et al. 2015) which are gradually more corrupted due to radiation. Most importantly, stage and beam motion, which are perhaps unpreventable, are a significant cause of image blurring (Glaeser,

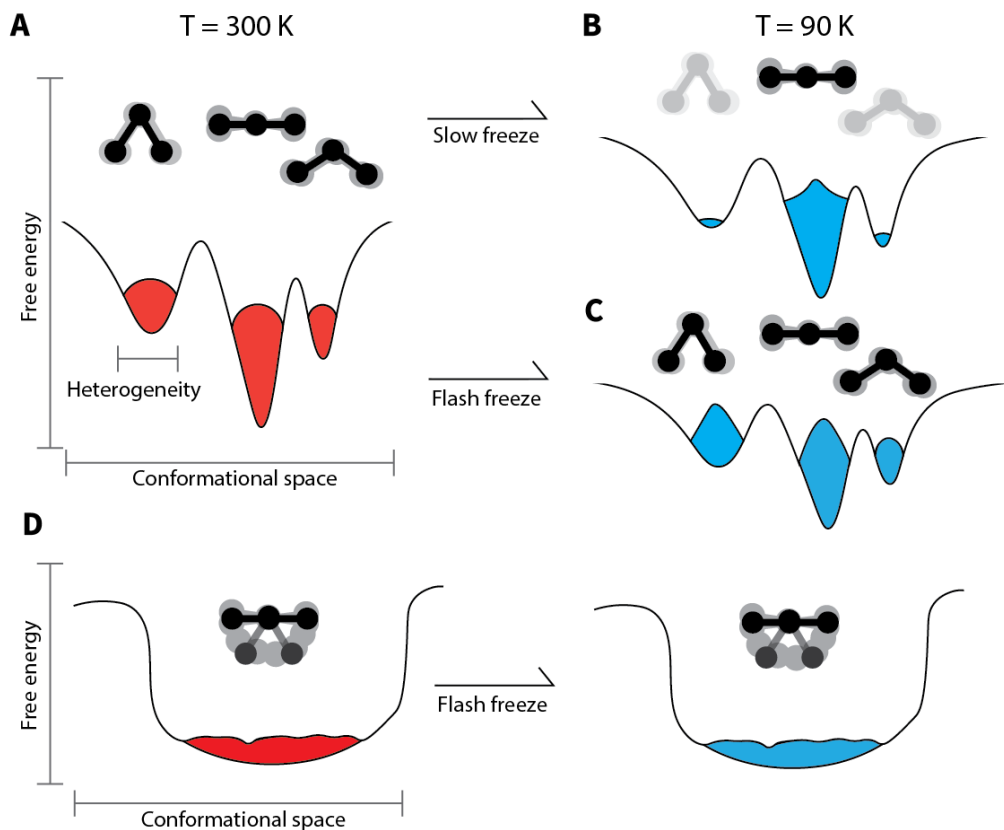


Figure 1.3: Vitrification aims to capture the molecules in their native populations. In (A) a sample containing a molecule which exists in three distinct conformations, each exhibiting some heterogeneity. In (B) slow freezing of the sample may lead to low energy conformations being more populated. Underpopulated conformations may not have sufficient particles to be identified. Flash freezing on the other hand, only causes small deviations in the ensemble (C) allowing the native populations to be captured in cryo-EM. Molecules which exist in a continuum of likely conformations are not so amenable to cryo-EM imaging (D).

McMullan, et al. 2011). With exposures broken down into frames, these motions can be accounted for, and the frames can be aligned and averaged to compensate for this (Li et al. 2013). Alignment may even be applied in sub-micrograph regions (or “patches”), as motions may occur at different rates throughout the frozen specimen (Brilot et al. 2012). One of many popular motion correction tools is MotionCor2 (Zheng et al. 2017).

The defocus required to get adequate contrast for particle picking and alignment, along with spherical aberrations due to the lenses cause the images to be corrupted. These corruptions are described by the contrast transfer function (CTF). By estimating the parameters of the CTF, some of these corruptions can be corrected. CTF estimations can be applied on a micrograph level, as per CTFFIND4 (Rohou et al. 2015). Theoretically, the defocus of each particle is unique, as particles are distributed throughout the ice and thus at different distances from the focal plane. Current best practices, are to perform per particle CTF correction (Relion, CryoSPARC, Xmipp). Given that particle motion may also occur along the Z plane, calculating on a per particle and movie frame basis as performed by gCTF

(K. Zhang 2016) may also be justified. It has been found that there is low agreement between the above listed CTF correction methods (Fernandez-Gimenez et al. 2023).

After the movies have been aligned (or “motion corrected”) and the CTF has been corrected, the next stage is locating (or “picking”) individual molecules of interest (the particles) in the cryo-EM micrographs. This is complicated by the fact that the images are very noisy, the particles are heterogeneous and their structure is not necessarily known. Particle extraction refers to cropping out “picked” particles of interest to build a collection of particles called a “stack”. Given that ice thickness varies and may be sub-optimal, particle picking can be performed on regions of optimal ice thickness using tools such as IceBreaker (Olek et al. 2022). The stack is then processed in a procedure called 2D classification, to identify particles of similar conformation and orientation (Scheres 2012). These particles can be further averaged (termed “2D class averages”).

Under the assumption that the noise in the images is Gaussian and white, averaging images of the same class yields an image with improved SNR. Theoretically, more images generally lead to better SNR (Rosenthal and Henderson 2003). However, this assumption assumes that the particles are homogenous and classification is without error. In practice, this is difficult due to the noise of the images being so high, which can lead to imperfect classification. Small amounts of conformational heterogeneity due to atomic motions are also unavoidable, and lead to class averages with lower resolution. The problem is compounded by the fact that not all orientations will have significant numbers of images due to preferred orientation. Not only that, but the quality of the ice, alignment errors and radiation damage will all contribute to the quality of the final 2D class averages (Beton et al. 2022).

To summarise, motion and CTF correction help to improve the resolution by counteracting particle drift and the corruption of signal due to defocus. 2D class averages aim to improve the signal-to-noise ratio but the averages will be blurred by any underlying local heterogeneity.

1.1.5 Reconstruction

In common tomography approaches such as those used in biomedical imaging, or even in cryo-ET, images of the target are acquired at different angles (a “tilt series” in cryo-ET). Given the angle of each image is known, it is possible to obtain a reconstruction by “back projecting” the images onto a 3D volume. In the case of cryo-ET, one can obtain 3D reconstructions of cells, or thin slices of cells, which can give insights into the spatial arrangement of subcellular features. For the reconstruction in SPA, the previously obtained stack of picked particles contains multiple projections of the molecule, but each with unknown orientation. In order to proceed with reconstruction, the relative orientations of the particles need to be established, and in the case where there may be multiple classes, they need to be identified and separated.

The most commonly applied method is known as “filtered back projection” (FBP) (the 2D images themselves are often called “projections”). Given the orientations of particles are not known, the reconstruction process may require a number of computational steps to be performed before sufficiently high resolution reconstructions are obtained. Usually, this would start with the production of one

or more low resolution “de novo” reconstructions, where particle angles are guessed using tools such as FREALIGN (Grigorieff 2016), RELION (Scheres 2012) or CryoSPARC (Punjani et al. 2017). It is a common choice to produce more reconstructions than the number of expected conformations (or classes). These additional reconstructions end up being “junk” classes, which “bad” particles get assigned to. Here “bad” particles refers to particles which may not be the molecule of interest or a particle which does not contribute to any 3D class in a reasonable way. Inspection of these reconstructions may give insights into the quality of the particles, whether there are multiple conformations, or the degree of bad particles in the stack. These reconstructions can then be used to better estimate the angles of the particles, and produce further refined reconstructions.

While there are many tools which are able to perform reconstruction and refinement of CTF parameters, it is not uncommon to find protocols which combine mixtures of these tools (DiIorio et al. 2022). Indeed, the field is full of software tools which seemingly overlap in their goals, but use different algorithms and methodologies. Hence, we end up with tools with slightly different characteristics, each amenable to different situations.

1.1.6 Interpretation & validation

The Fourier Shell Correlation (FSC) (Harauz et al. 1986) is computed from two independent 3D reconstructions of the cryo-EM data. The principle behind the FSC is that the correlation between spatial frequency shells gives an indication about the robustness of signal at these frequencies. The FSC is typically presented as an FSC curve, a plot of correlation vs spatial frequency. However, a number of attempts have been made to distill the FSC down to a single number which estimates the global resolution of the map. The “gold standard” approach is to identify the frequency at which the correlation drops to 0.143. This approach is both ubiquitous and is currently also reported by the Electron Microscopy Databank (EMDB) (Z. Wang et al. 2022). The choice of 0.143 as a cutoff is based on an attempt to make the resolution estimate comparable with those from X-ray crystallography (Rosenthal and Henderson 2003). Another common choice is 0.5 (Böttcher et al. 1997). Others have advocated against fixed cutoffs, offering alternatives based on standard deviations (Heel et al. 2005). Regardless of the choice of cutoff, these methods assume that the two datasets used to compute the half-maps are indeed independent. This is not always the case if a tight mask is involved in reconstruction, which could lead to resolution being overestimated.

It has been noted (Rosenthal and Rubinstein 2015), that resolution estimates using the “gold standard” approach, may only hold for some parts of the map - it is not a homogenous quantity. Local resolution estimates attempt to assess resolution on a per-voxel basis. Rather than rely on half-maps, ResMap (Kucukelbir et al. 2014) instead finds the highest frequency sinusoidal wave that fits a given voxel, whilst accounting for false discovery rate. Blocres (Cardone et al. 2013) and Phenix (Adams et al. 2010) on the other hand, rely on half-maps to perform a local windowed version of the FSC method.

The 3D reconstruction is the culmination of the various image processing steps described so far. Understanding the differences in resolution across regions of the reconstruction can offer insight into the nature of the imaged molecules. Higher

resolutions are typical for regions which are rigid, compared to mobile elements such as loops. When images of flexible molecules are aligned, it tends to be that the largest rigid parts of the molecules end up being better aligned, whilst smaller domains may end up being less well aligned. This leads to core regions with high resolution, while domains which may move in relation to it end up with lower resolution. Thus, variation in local resolution can be an indication of flexibility. Other interpretations are also possible: poor rotational alignment of particles is likely to produce lower resolution at the periphery where small angular differences lead to large errors in cartesian coordinates.

In extreme cases, flexible domains will have very low resolution, or may not be visible at all. Such molecules likely undergo continuous movements (fig. 1.3D), in which case focused reconstruction approaches might be required to resolve the domains individually, however this is beyond the scope of this brief introduction.

Intrinsically disordered regions have too many distinct conformations to ever be sufficiently resolvable. However, identifying their existence – regions with low resolution or complete lack of density – in a molecule may be of interest in its own right (Uversky 2016).

1.2 Flexible fitting and refinement

The ultimate goal of cryo-EM experiments is to answer questions about the spatial arrangement of molecules. Resolution dictates how detailed these answers can be. At resolutions worse than 10Å, cryo-EM maps are only able to indicate the positions of domains or individual units in a complex. At such low resolutions, models can be rigidly fitted (sometimes referred to as docking) to the density. Even at these resolutions, which are today considered low (fig 1.4b), useful biological questions can be answered about the spatial arrangement of proteins in complexes (Ranson et al. 2001). At resolutions better than 4Å it is sometimes possible to do without any starting model and build a model directly from the density, using tools such as ModelAngelo (Jamali et al. 2024). Such resolutions are becoming increasingly common, but are not necessarily the norm.

Today, many cryo-EM reconstructions fall between 4Å and 10Å resolution (fig 1.4a). Such resolutions enable domains and secondary structure elements (SSEs) to be distinguished (fig 1.4b). Discrepancies between the positioning of these elements in the atomic model with respect to experimental data may justify altering their relative orientation. A class of computational approaches known as “flexible fitting” or “refinement” can perform such structural optimizations automatically.

These terms, flexible fitting and refinement, do not have well defined meanings. However, it is fair to say that refinement typically refers to small optimizations in atomic coordinates and atomic displacement parameters (ADPs, B-factors). These generally do not have a wide radius of convergence, which is to say, models must be relatively close to their ideal conformation for refinement to produce satisfactory results. They also have their origins in X-ray crystallography, where the phases must also be refined simultaneously. Generally, a reasonable estimation of the phases is required, which again means having a model close to the ideal coordinates.

Flexible fitting, on the other hand, typically refers to the case where large conformational changes are required to optimize the fit of the initial model to the

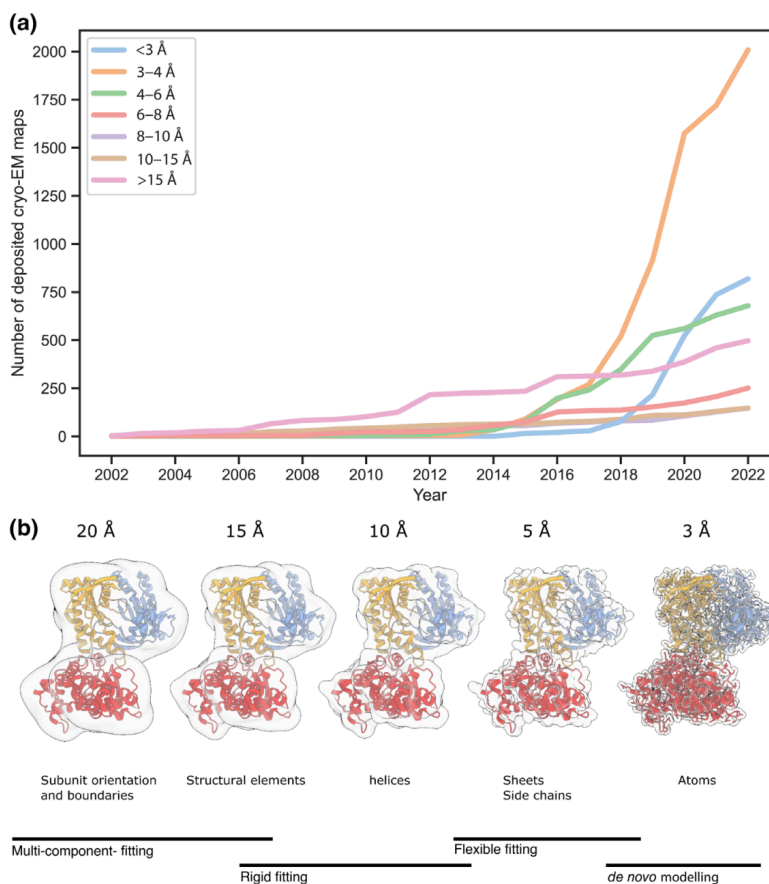


Figure 1.4: (a) Resolution has been rapidly improving with the majority of recently deposited maps in the EMDB falling in the 3-4Å range. (b) The visible features at different resolutions determine the type of modelling approach. Adapted from Beton et al. 2022.

data. Often, flexible fitting tools will employ a final refinement stage to further optimize the local geometry (Kidmose et al. 2018; Kim, Moriarty, et al. 2019) or such protocols are performed in an ad-hoc manner (Kryshtafovych, Malhotra, et al. 2019).

TEMPy-ReFF, (TEMPy REsponsibility-based Flexible-Fitting), is the basis of one of the publications in this thesis (Beton*, Mulvaney* et al 2024). The method is in keeping with the flexible fitting tradition. Namely, a combination of scoring functions, molecular dynamics, biasing potentials and restraints. The rest of this section is dedicated to exploring these fundamental details and the diverse array of approaches in the flexible fitting field.

1.2.1 Structure Prediction

Before discussing the intricacies of flexible fitting in detail, it is important to acknowledge that while there have been advances in flexible fitting approaches, including those introduced in this thesis in (section 3.1) (Beton*, Mulvaney*, 2024), the biggest leap in recent years is how initial models for flexible fitting are obtained.

Arguably the best sources of high resolution, reliable structural models are still X-ray and NMR experiments. However, cryo-EM is able to study structures which

have never been crystallized or are too large for NMR. Computational structural predictions have been a stalwart of cryo-EM modelling. Traditionally, these have involved homology modelling (Schwede et al. 2003), and modelling with restraints (Sali et al. 1995; Fiser et al. 2003). An important advance was the use of 2D contact maps as a stepping stone to 3D structure. Residue coevolution based methods, (Hopf et al. 2012; Jones et al. 2012; Morcos et al. 2014), attempt to statistically determine pairs of residues which coevolve. The idea being that residues which are involved in important contacts in the protein structure will mutate together in such a way that they maintain their interactions. Meanwhile, other residues will mutate independently. These methods are only possible because of the large databanks of sequence information, which can be constructed into the deep multiple sequence alignments (MSAs) required for determining pairs of mutating residues. More recently, this has culminated in the AlphaFold2 (Jumper et al. 2021) deep learning approach, whose “evoformer” network relies heavily on coevolution information from an input MSA. This approach, which stood out in the 14th critical assessment of structure prediction (CASP14) challenge (Kryshtafovych, Schwede, et al. 2021), has become the basis for an expanding repertoire of structure prediction methods. The training of such neural networks requires not just an expansive set of sequences for building MSAs, but also experimentally solved structures to train the network against. The ideas behind coevolution extend beyond intra-protein residue contact prediction. Indeed, coevolution information can be used to determine protein assemblies by effectively concatenating the MSAs of the proteins of interest. With this approach, CASP15 saw a similar advance (Ozden et al. 2023) in multimeric structure prediction (fig. 1.5).

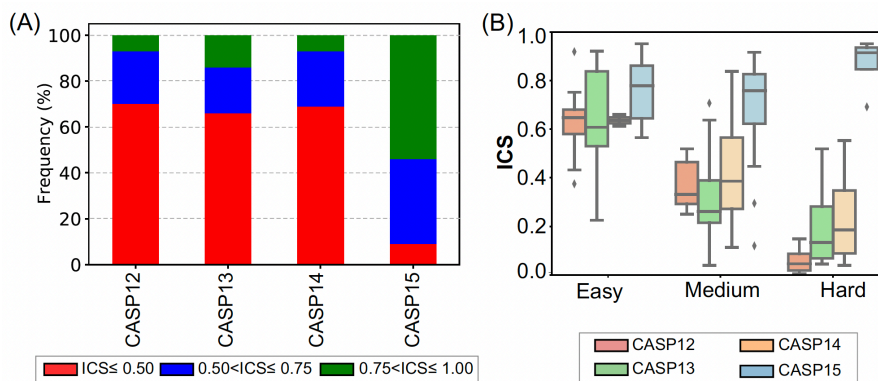


Figure 1.5: The improvement of protein-protein interfaces as determined by the interface contact similarity (ICS) score. CASP15 saw a large improvement in ICS scores, with more than 50% of the predictions having ICS scores above 0.75. When compared across difficulty levels of targets, the “Hard” category where no solved structures exist in public databases, saw a huge jump in accuracy. Figure adapted from (Ozden et al. 2023).

1.2.2 Assessing quality of fit

In order to improve the fit of an atomic model to the data, a way of measuring the goodness-of-fit is required. Given we have established how a molecule scatters electrons and thus derived the coulombic potential (eq. 1.5), we can simulate the coulombic potential for a given atomic model and compare it against the

experimental data. While the multi-gaussian model of an atom is often used in X-ray refinement such approaches are considered to be superfluous at resolutions typically seen in cryo-EM. Indeed, many of the methods that this thesis builds upon, assume a simple one Gaussian model to describe the intensity profile (Orzechowski et al. 2008; Topf et al. 2008; DiMaio, Song, et al. 2015; Igaev et al. 2019; Pintilie et al. 2021). Thus, the approximate simulated coulombic potential at a given location r is given by the function $\rho_{sim}(r)$ which is defined as the sum of the coulombic contributions of all N atoms in a molecule. Here Z_n and r_n are the atomic number and atomic coordinates of the n^{th} atom. The parameter σ is dictated by the resolution and must be estimated.²

$$\rho_{sim}(r) = \sum_n^N Z_n \exp\left\{\frac{(r_n - r)^2}{\sigma^3}\right\} \quad (1.6)$$

From this simulated coulombic map, it is possible to measure how well the atomic model reflects experimental data. One such measure of fit, important to this thesis, is the cross correlation (CC) between the simulated map and the experimental one. The cross correlation is computed across the corresponding voxels v of the two simulated and experimental maps.

$$CC(\rho_{sim}, \rho_{exp}) = \frac{\sum_v^V (\rho_{sim}(v) - \overline{\rho_{sim}})(\rho_{exp}(v) - \overline{\rho_{exp}})}{\sqrt{\sum_v^V (\rho_{sim}(v) - \overline{\rho_{sim}})^2 (\rho_{exp}(v) - \overline{\rho_{exp}})^2}} \quad (1.7)$$

Like, in the case of global resolution determined by FSC (see 1.1.6), the CC score returns a single value which is indicative of overall fit to the experimental data, but does not indicate where problematic regions may be arising. A number of alternatives exist, such as SMOC (Cragolini, Sahota, et al. 2021; Joseph et al. 2016) and the residue CC of Phenix (Adams et al. 2010) which provide local correlation based scores.

1.2.3 Molecular Dynamics

Classical molecular dynamics (MD) can be used to simulate how certain aspects of the molecular system evolve over time. The equations of motion are applied to all atoms in the system and integrated over a period of time. Under the ‘‘classical’’ MD scheme, the bonds are maintained throughout the course of the simulation and chemical reactions do not take place.

The forces ‘‘experienced’’ by atoms are defined by potentials or ‘‘force fields’’. These force fields are sets of parameters that have been empirically determined or parameterised with the help of quantum mechanical computations. Commonly used force fields include AMBER (Case et al. 2024) and CHARMM (Brooks et al. 2009). In order to accurately describe the behaviour of atoms in biological molecules, the parameters of atoms (even of the same element) will be different based on the chemical environment they are in. This has important implications: molecules must be unambiguously defined in the input for common MD programs like NAMD (Phillips et al. 2020), GROMACS (Bekker et al. 1993) and OpenMM (Eastman

²Some common approaches are to multiply the global resolution as determined by the FSC by a scaling factor such as 0.187 (Wriggers et al. 1999) or 0.225 (Pettersen et al. 2004).

et al. 2017), to correctly parameterise the forces for each atom in the system. If the chemical environment of one of the atoms is ambiguous because the residue it is part of is missing atoms or incorrectly named, the simulation can not proceed, as it will be unable to “decide” what parameters those atoms should be given.

Molecular dynamics is well established in the field of structure determination. First, it was applied in the determination of NMR structures (Brünger, Clore, et al. 1986), then in the refinement of X-ray crystal structures (Brünger, Kuriyan, et al. 1987) where it was noted for being able to escape some of the local minima inherent in the least squares approaches popular at the time.

Flexible fitting: Steered molecular dynamics

In flexible fitting methods, rather than use MD to study the evolution of a molecular system over a period of time, the force field acts to maintain physiological features of the biological molecule while a “biassing” potential (section 1.2.4) adjusts the structure to improve overall fit. This biassing term has no physiological basis and the trajectories of a molecule under its influence does not necessarily reflect its natural modes of motions. The overall energy term E for such an MD system is given by the sum of the kinetic energy E_k , the energy terms of the force field E_{ff} and the additional energy term due to the biassing potential E_{bias} :

$$E = E_k + E_{ff} + E_{bias} \quad (1.8)$$

Enumerating all the potential energy terms from a force field such as CHARMM or AMBER would not be illuminating. Instead, two key potential energy forms are described. The harmonic potential E_{bond} , which behaves like a spring, forms the basis for modelling the behaviour of bonds. Here, r_0 is the ideal bond length, r is the current bond length and k is the strength of the potential.

$$E_{bond} = k(r - r_0)^2 \quad (1.9)$$

The harmonic potential can be used to describe other aspects of an atomic model, such as bond angles and torsion angles. It is also commonly used as an external non-physiological restraint in protein modelling, for example to maintain proximity of certain atoms or maintain secondary structure.

The van der Waals (vdW) force is commonly modelled using the Lennard-Jones “12-6” potential, where r is the interatomic distance of the atoms, σ is their combined vdW radii and ϵ is the well depth. The weak attractive term models electrostatic attraction due to polarisation, and the hard repulsive term is due to Pauli exclusion principle. The low clash scores observed in MD based refinements have been attributed to this repulsive term (Y. Wang et al. 2018).

$$E_{LJ} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad (1.10)$$

Integration

In order to determine the motions of the atoms in the system, the change in velocity and position of each atom must be computed. For simple two body systems, there are analytical solutions, but for more interesting systems the “many body” problem

is encountered. Here, the solution to equations of motions are computed over small timesteps. The smaller the timestep, the more accurate the integration, at the expense of computing time. Various algorithms have been established for performing integration, with different cost/accuracy tradeoffs. The Velocity Verlet (Swope et al. 1984) method is a popular choice and is described as follows: Given a particle with a position x , a velocity v , experiencing an acceleration a (determined trivially by the relationship established by Newton, $F = ma$), the approximate new position x and velocity v of the particle after a small fraction of time, δT , is given by the expression:

$$x' = x + v\delta T + \frac{1}{2}a\delta T^2 \quad (1.11)$$

$$v' = v + \frac{1}{2}(a + a')\delta T \quad (1.12)$$

Because the acceleration is position dependent, the updated velocity v' is approximated by using the average acceleration experienced by the particle at its initial location a , and its new location a' . Due to the errors associated with these approximations, the total energy of the system is not well conserved. This can be accounted for by adjusting the velocities of the particles at given intervals so that the total energy is approximately constant.

Given, that small timesteps are required for accurate integration, where each timestep involves computing forces on many atoms, it is easy to see why MD-based methods are computationally expensive. Fortunately, flexible fitting approaches typically run over short time periods and likely force conformational changes over timescales which are not physiologically reasonable. Still, attempts have been made to improve the performance of flexible fitting approaches such as by coarse-grained dynamics (Grubisic et al. 2010), where residues are reduced to $C\alpha$ centered particles with torsion angles that enforce sensible stereo-chemistry. Common reduced representation force fields include the ‘‘Go’’ model (Go 1983) and Martini (Souza et al. 2021).

Ensembles

The MD system described thus far, where the total energy of the system is maintained, is known as the ‘‘NVE’’ ensemble. This nomenclature refers to the fact that the number of particles (or atoms) N , the volume V and the total energy E are constant. In practice, it is usually a constant temperature T rather than energy which is desired in biological simulations, the so called ‘‘NVT’’ ensemble. Because temperature is a statistical quantity related to the average kinetic energy \overline{E}_k of the particles in the ensemble it must be constantly adjusted.

$$\overline{E}_k = \frac{3}{2}Nk_B T \quad (1.13)$$

Imagine for a moment, a simulation of a protein in an NVE system with an initial temperature of 300K. As the protein relaxes into a low energy conformation, it transforms its potential energy into kinetic energy, heating up the system. In the case of flexible fitting, the external biasing potential can also impart a large out

amount of kinetic energy to the system, which needs to be dissipated. A number of thermostats exist, whose purpose is to adjust the average kinetic energy of the system such that a given temperature is maintained, for example the Andersen thermostat (Andersen 1980). Another popular approach is Langevin dynamics (Pastor 1994), here the velocities of particles are constantly dampened by friction or accelerated by collisions with an imaginary solvent. Finally, an important technique used to improve the chances of finding a global energy minimum is to gradually cool the simulation down by controlling the thermostat, a technique called simulated annealing, explored in section 1.2.5.

Implementation

Fortunately, most of the complexities associated with building efficient MD simulations are taken care of by packages such as GROMACS, NAMD and OpenMM, the latter being the method used in the software developed in this thesis (see section 3.1). These software packages allow scientists to focus their efforts on defining mathematical expressions for forces, whilst efficient integration is handled by dedicated routines. Recently, graphics processing units, (GPUs), which as the name suggests were once devoted to computer graphics tasks, have become an important part of accelerating scientific computational workloads which are made up of highly parallelizable tasks. The above mentioned MD packages all support GPU acceleration, enabling simulations which previously would have required dedicated scientific computing resources to be run on “commodity” hardware.

1.2.4 Biassing potentials

Flexible fitting methods are based on adding a potential to the standard force fields of a molecular dynamics simulation. This potential is designed to improve the fit of the atomic model to the experimental map. Unlike the potentials of the force fields such as AMBER and CHARMM, this potential has no physiological basis. The biassing potentials can be divided into three different main categories as described below and summarised in Table 1.1.

Density driven

Under this scheme, a potential is derived from the experimental map which pushes atoms in to regions of high density. Concretely, atoms experience a force proportional to the negative gradient of the experimental map. This approach was popularised by MDFF (Trabuco et al. 2008), and has spawned a number of other methods which extend this with various enhancements (Singharoy et al. 2016; Vuillemot et al. 2022; Dahmani et al. 2024; Croll 2018). The approach is straight forward to implement, it can be accurate at high-resolutions as density peaks are centered around atoms. However, at lower resolutions, it is less accurate as the density in of coulombic potential map, will no longer be centered on atoms.

Correlation driven

Correlation driven approaches such as (Orzechowski et al. 2008), Flex-EM (Topf et al. 2008), and CDMD (Igaev et al. 2019), try to fit the model in a way that

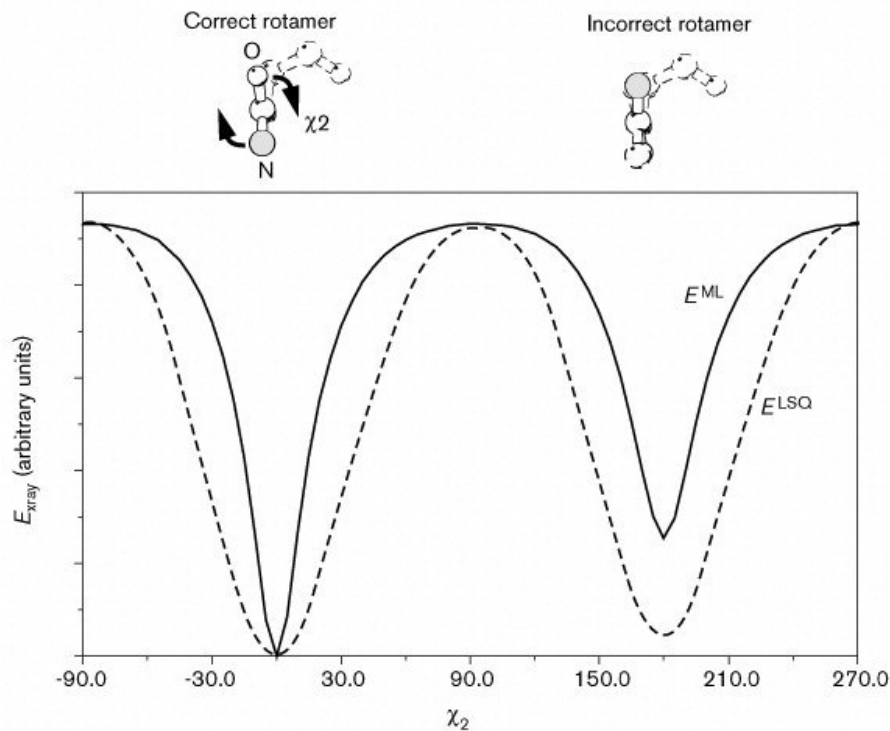


Figure 1.6: The problem of local minima is pervasive and can be a problem at multiple scale, from the fitting of domains down to adjusting rotamers. The potential energy surface of landscape is the sum of all the different potentials that make up a system. In this figure from Brünger, Adams, et al. 1997, we see how the conformation of a rotamer can become trapped by the fitting potential. The difference between the potential energy curves of the biasing potentials can impact how well a system is able to escape local minima and converge on the more optimal solution. In the figure, a least-squares and maximum-likelihood potential are compared.

increases correlation. The potential energy function is given by E_{cc} , where k is a scaling factor which determines the strength of the potential and CC is the cross-correlation as defined in (eq. 1.7).

$$E_{cc} = k(1 - CC(\rho_{sim}, \rho_{exp})) \quad (1.14)$$

Given that computing CC requires computing a simulated map (eq. 1.6), the force field must be regularly updated as the model changes, making it computationally expensive. The force field has the advantage of being theoretically more accurate than the density based approach described previously (see 1.2.4) as is illustrated in fig. 1.7.

Gaussian mixture model based

One limitation of the previously described cross-correlation biasing potential is that in order to compute the CC , an estimation of global resolution is required. It is also well established that a global resolution does not describe the heterogeneity of cryo-EM reconstructions. In the paper in section 3.1, we implement a new

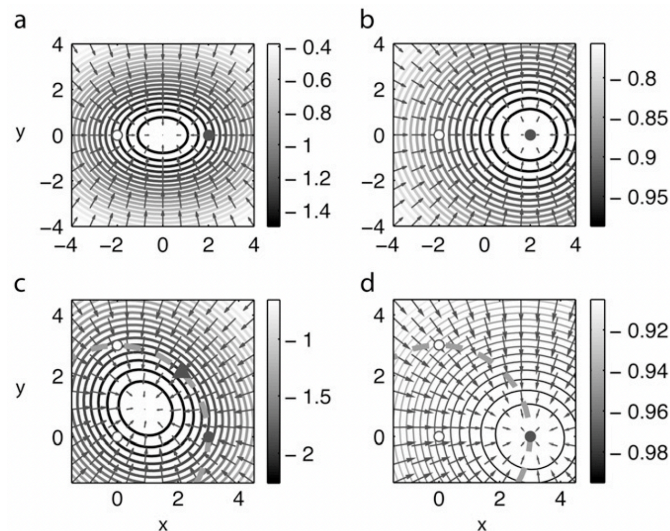


Figure 1.7: A comparison of the forces experienced by atoms in density and correlation based force fields. In each of the panels, the circles represent atoms and the arrows represent the direction of the forcefield. In (a) and (b) two atoms are present with a Gaussian width of $\sigma = 4$ and positions $(-2, 0)$ and $(2, 0)$ respectively. In the density based force field (a), the force experienced by the shaded atom is towards the center of the two atoms. In the correlation based force field (b) the atom experience no net force in its original position. The same phenomenon is seen in the three atom case in (c) and (d). Adapted from Orzechowski et al. 2008.

biassing potential which does not rely on a fixed resolution σ as per (eq. 1.6). Instead, we use a value B for each atom.

$$\rho_{sim}(r) = \sum_n^N Z_n \exp\left\{\frac{(r_n - r)^2}{B_n^3}\right\} \quad (1.15)$$

The parameters B_n and the atomic positions r_n are optimized using the expectation-maximisation algorithm as described in the paper.

Further more, we show that using B we can derive an ensemble of models which describe the experimental density. While B can describe heterogeneity its interpretation is not trivial and is the subject of the discussion.

Determining biassing strength

Unfortunately, every cryo-EM reconstruction is unique, with their own background noise, resolution profiles and artifacts. It is thus difficult to define a scaling factor for the biassing force which can be applied for all model fitting scenarios which produces both good fit to the experimental data whilst avoiding overfitting and distorting the model. One approach proposed in (DiMaio, J. Zhang, et al. 2013) was to build models into two independent reconstructions using Rosetta, with increasing contributions from the fit to map terms. This is very much in keeping with the R_{free} approach established in crystallography (Brünger 1992).

The optimal biassing strength is at the point just before overfitting occurs, which is when the two FSC or real-space correlation curves begin to diverge. The optimal biassing strength would be at the point just before these correlation

Potential	Res. required?	Comp. cost	Accuracy	Fits
Density	No	Cheap	Low	Coordinates
Correlation	Yes	Expensive	High	Coordinates
GMM	No	Expensive	High	Coordinates + Bfactor

Table 1.1: Three main classes of biasing potentials have been developed for flexible fitting, including the GMM based potential introduced in this thesis. Here the benefits and tradeoffs of each are listed.

measurements diverge. This approach has also been applied to correlation based flexible fitting software such as (Igaev et al. 2019) to determine the fitting strength.

1.2.5 Convergence & restraints

The biasing potentials described perform well in the case where small conformational changes are required, but are subject to a local optima and structural distortions when the initial structures require significant conformational changes. To overcome such issues additional methods are commonly employed which restrain aspects of the geometry, smooth the potential energy landscape of the biasing potential, or sample alternative starting conformations. These approaches are described below along with their appearances in fitting software.

Geometry and rigid body restraints

It is noted by authors of the popular MDFF flexible fitting software that stereochemical violations can occur (Schreiner et al. 2011). They recommend harmonic restraint to preserve chiral center at each peptide bond whilst applying fitting potential. Similarly, it is common practice to apply restraints to secondary structures elements (SSEs) during fitting (Trabuco et al. 2008; Topf et al. 2008). It is generally assumed, that SSEs and their local environment are structurally stable (or well predicted in the case of computational models) and should be preserved during fitting.³

Besides preserving geometry, such restraints can also improve convergence. During flexible fitting, atoms in the model being fit will feel forces due to local density. Depending on how far the atoms are from their target density, these local density pockets can trap regions of the model. This can lead to distortions as neighbouring atoms are pulled in opposing directions.

A common solution to avoid these distortions is to treat parts of the structure as rigid until they have moved closer to the correct density, before allowing the fit of their internal structures to be optimized. One way of imposing rigidity, is to apply additional harmonic distance restraints between residues which constitute SSEs or larger areas of interest. RIBFIND (A. P. Pandurangan and Topf 2012) is a tool for finding rigid bodies in proteins, starting at the level of SSEs. These are progressively combined into ever larger rigid bodies based on the strength

³There are always exceptions. One extreme case, is the pore forming toxins in the MACPF family. These undergo large conformational changes where the α -helices extend to form complementary β -strands of the large β -barrel which makes up the pore.

of their interactions with one another. Appropriately chosen rigid bodies were shown to improve geometry and convergence of models fit using Flex-EM software (Pandurangan et al. 2012b) However, in its original implementation, Flex-EM has only been applied to protein structure refinement as RIBFIND is not able to cluster RNA secondary structure. In (Malhotra*, Mulvaney*, et al.) which is introduced in (section 3.2), a new implementation of RIBFIND is described which supports RNA and DNA structures.

An alternative to fully rigid body restraints are adaptive distance restraints (ADRs). These have been implemented in a variety of refinement softwares such as BUSTER (Smart et al. 2012), and REFMAC (Nicholls et al. 2012). Under these schemes, restraints behave roughly harmonically, while the atoms are close to the idealised distance. At larger deviations they stop behaving harmonically (the potentials “top out”), allowing greater freedom of movement. In addition to “topping out” at large distances, the adaptive restraints implemented in ISOLDE (Croll 2018) “bottom out”, in other words apply no force when the atom is sitting in the well close to the target position. The width of the “bottom” is a tunable parameter. The theoretical advantage of bottom out potentials, is that restraints derived from low-resolution information likely have an error associated with them (which is hopefully the size of the well bottom), they thus allow the high-fidelity molecular force field to better govern the atoms coordinates when they are close to the target geometry (Croll and Read 2021).

Smoothing the density

A well known trick in the flexible fitting community to improve convergence when large conformational changes are required, is to blur the map. This makes the potential energy landscape less rugged, with fewer minima to become trapped in. After performing flexible fitting with the blurred map, the original map is used to further optimize higher-resolution features. This kind of ad-hoc protocol is described in (Kidmose et al. 2018) and by others (Kim and Sanbonmatsu 2017; Casañal et al. 2020).

It is conceivable, that multiple levels of blurring might be needed depending on the starting model and the cryo-EM reconstruction. In Igaev et al. 2019 and Singharoy et al. 2016, blurring and progressive sharpening are performed automatically as part of the refinement protocol. I refer to this approach as “progressive sharpening”.

Normal mode analysis

Normal mode analysis (NMA) provides a mathematical framework for determining the low-frequency modes of motion in molecules (Hinsen 1998; Tama et al. 2004). Sampling from conformations from these motions has been used to find models which better fit the density (Hinsen et al. 2010). This approach has been applied in conjunction with molecular dynamics, where NMA is used to find a more optimal starting model, before standard biased MD approaches refine the model further (Dahmani et al. 2024; Vuillemot et al. 2022).

Simulated Annealing

One of the advantages of molecular dynamics over gradient-based methods such as conjugate gradient descent that was noted early in crystallographic refinement methods (Brünger, Kuriyan, et al. 1987) was MDs ability to overcome some local minima. This can be further enhanced using simulated annealing. Here, the systems ability to climb over energy barriers is improved by increasing the temperature and thus the available energy of molecules in the system. By gradually cooling the system, the model is unlikely to revisit high energy states.⁴

This method has been applied to cryo-EM flexible fitting (Topf et al. 2008; Igaev et al. 2019) and is employed in the Phenix real-space-refinement tool (Adams et al. 2010).

Other sampling methods

Even with simulated annealing, some barriers are simply too unlikely to be crossed. Rotamers can become trapped in energy minima which are difficult to escape as illustrated in figure 1.6. While blurring the experimental map can smooth the contributions of the map potential, the various potentials from the force fields such as AMBER contribute to the rugged terrain. Atomic models may have sub-optimal torsion angles, but finding the optimum angles may require overcoming a steep energy barrier. An approach used in the ISOLDE software (Croll 2018) is to allow user intervention to correct for such cases. Phenix (Adams et al. 2010) uses heuristics like the grid search proposed by (Oldfield 2001) to sample common torsion angles which make up protein backbones. RNA backbones have seven torsion angles compared to the two which describe proteins (described later in RNA structure) making the nucleic acid geometry particularly challenging. QRNAS (Stasiewicz et al. 2019) applies additional restraints along with the AMBER force field in an attempt to regularise RNA backbones to better fit with known backbone conformers. Rosetta (Simons et al. 1999), uses Monte Carlo sampling an alternative to MD, and may be able to sample more conformational candidates using fragment searches. ERRASER (Chou et al. 2013) is a Rosetta based tool for the refinement of RNA structures. This tool is used in the refinement pipeline described in the paper (Mulvaney et al. 2023) (see section 3.3) of this thesis. Here, RNA predictions from CASP15 needed additional backbone corrections as their starting geometry was suboptimal. This approach is similar to an approach used for fitting proteins which combined flexible fitting with Rosetta (Lindert et al. 2015).

1.3 RNA

Our understanding of the role of RNA in biology has transcended early views of the molecule as a mere messenger encoding protein sequences (Crick 1970). RNAs are capable of performing complex regulatory and enzymatic roles in cells. Even messenger RNA (mRNA) molecules are able to regulate their own transcription through structures formed in their untranslated regions (UTRs). Examples include the iron responsive element (IRE) (Address et al. 1997) and the cobalamin

⁴Ironically, annealing is essentially, the process that is avoided by plunge freezing. Here, microscopists aim to prevent the molecules from finding their global energy minima.

Name	Fitting force	Restraints/Sampling
MDFF (Trabuco et al. 2008)	Density	SSE restraints
ReMDFF (Singharoy et al. 2016)	Density	Progressive sharpening
NMMD (Vuillemot et al. 2022)	Density	Normal mode sampling
MDFF-NM (Dahmani et al. 2024)	Density	Normal mode sampling
CDMD (Igaev et al. 2019)	Correlation	Progressive sharpening
Flex-EM (Topf et al. 2008)	Correlation	Rigid body restraints
ISOLDE (Croll 2018)	Density	User guided
TEMPy-ReFF (Beton* et al. 2024)	Density + GMM	Hier. rigid body restraints

Table 1.2: This table attempts to summarise the numerous flexible fitting and refinement tools which have been developed. Each chooses slightly different approaches to fitting in order to avoid local minima and aid convergence.

riboswitches whose conformational states have recently been elucidated with cryo-EM (Ding et al. 2023). Many viruses must hijack the hosts translational facilities in order to replicate. Often this involves the evolution of specific features in the 5' UTR of the viral RNA. Those of coronaviruses, including SARS-CoV-2, have been studied using both NMR (Vögele et al. 2023) and cryo-EM (Kretsch et al. 2024). Other viral adaptations include structures for protecting polyA tails from deadenylation (Mitton-Fry et al. 2010). Viral RNA, and RNA in general, is thus not just of interest because of the genetic material it encodes, but also as a complex structured entity capable of carrying out a range of functions.

One of the most studied RNA containing complexes by cryo-EM are ribosomes which perform an essential role as protein production machines. Owing to their large size, these have been particularly amenable to the EM based techniques including in-situ cryo-ET (Erdmann et al. 2021). The unique structure of bacterial ribosomes compared to eukaryotic ones, has made it the target of an ever-growing number of antimicrobial agents, many of which no longer work due to resistance. Anti-microbial resistance (AMR) was estimated to have caused 1.27 million deaths in 2019 (C. J. L. Murray et al. 2022). Understanding the ribosome at a structural level is thus essential to combat the ever evolving AMR pandemic.

Small RNA molecules occur throughout nature and include molecules such as ribozymes, the Group I, II and III introns. Despite their size, they have the ability to perform functions such as splicing and catalysis - pointing towards the possible RNA origins of life. Like ribosomes, introns are also of interest from a global health perspective. While bacterial ribosomes are sufficiently different from those in eukaryotes to develop targeted drugs, fungal ribosomes are similar. Introns are present in many fungi and may be a potential target for future drugs (T. Liu et al. 2024).

However, until recently, these smaller RNA molecules have been off limits due to the poor SNR of earlier detectors. Previously, molecules smaller than 200kDa did not offer sufficient contrast to allow the alignment and determination of orientation with sufficient accuracy to produce reconstructions (Henderson et al. 2011). Besides better direct electron detectors, some additional methodologies have been introduced for stabilising these molecules. ROCKS for example introduces carefully engineered hairpin loops into the RNA structure to encourage multimerisation

through the creation of kissing loops (D. Liu et al. 2022). Scaffolding approaches where by the RNA target is embedded in a larger structural design have also been applied (Sampedro Vallina et al. 2023).

These advances, and the growing interest in RNA, have culminated in an explosion of nucleic acid structures being solved using cryo-EM (fig. 1.8).

1.3.1 RNA structure

RNA is composed of four units, the purines adenine (A) and guanine (G), and the pyrimidines cytosine (C) and uracil (U). Despite this simplicity, RNA is capable of forming diverse secondary and tertiary structures. This is defined predominantly by helical motifs formed by canonical base pairing between C-G, U-A nucleotides. Canonical base pairing involves the formation of bonds between the Watson-Crick edges of the respective bases. Some alternative (non-canonical) base pairing such as wobble pairing between G-U also exist. Besides the canonical Watson-Crick edges, the Hoogsteen and sugar edges are also possibilities for base pairing. Various attempts have been made to systematize these interactions including the nomenclature introduced in (Leontis and Westhof 2001) which offers excellent insight into the complexities of RNA base pair interactions.

The helical forms which dominate RNA structures occur in three major types, A, B and Z, with A being the most common. Protein tertiary structure is determined by the network of interactions formed between amino acid side chains. In RNA, tertiary structures are predominately determined by long distance inter-helical interactions formed between exposed nucleotides in hairpin loops. A plethora of tertiary structures arise from these such as kissing-loops, pseudoknots, kinks (which involve non-canonical G-A interactions) to name a few. Due to the negatively charged phosphate groups, coordinated metal ions, enable the backbone to come into proximity. The coordination of metal ions is also important for catalytic and sensing functions in many RNA. A number of reviews exist which focus on aspects of secondary, tertiary, quaternary and backbone structure (Leontis, Lescoute, et al. 2006; Butcher et al. 2011).

Protein backbones structure is described by the ψ and ϕ torsion angles formed by the amino acid (Ramachandran et al. 1963). Statistics about the frequencies at which these torsion angles occur for different amino acids is important for detecting geometry outliers which may be indicative of modelling errors. These statistics also inform the development of force fields and refinement methods. RNA backbones are complex to characterise due to the large number of rotatable bonds. One approach is by defining “suites”, composed of the seven torsion angles between adjacent sugars (L. J. W. Murray et al. 2003). Within this seven-dimensional torsion space, 46 different clusters have been enumerated (Richardson et al. 2008). Refinement tools such as ERRASER (Chou et al. 2013) and QRNAS (Stasiewicz et al. 2019) attempt to regularise the backbone geometry to fit these suites.

Base-pairing is energetically favourable, making RNA and DNA helices highly stable, but also leading to kinetic traps: misfolded intermediates may themselves be stable and slow down the correct folding process. Excitingly, cryo-EM has been used to observe the folding intermediates of a group I intron (Bonilla et al. 2022) and in a designed RNA system (Sampedro Vallina et al. 2023).

1.3.2 Rationally designed functional nucleic acids

Cryo-EM experiments are creating insight into the world of RNA and nucleic acid structures. At the same time, designed nucleic acids are finding exciting roles in the experimentalists toolbox. One such example, are origami nanostructures composed of rationally designed DNA molecules. These have been used as signposts for labelling otherwise difficult to discern molecules in cryo-ET data (Silvester et al. 2021). They have also been used as scaffolds for small proteins which would otherwise be difficult to resolve (Martin et al. 2016). The discovery of green fluorescent protein (GFP) by Shimomura et al. 1962 was revolutionary, enabling live cell imaging of fluorescently tagged proteins (Chudakov et al. 2010). Only in recent years have equivalent tags become available for studying RNA molecules. Aptamers such as “Spinach” (Paige et al. 2011), “Broccoli” (Filonov et al. 2014) and “Pepper” (Huang et al. 2021), enable RNA to be studied throughout its cellular lifecycle with light microscopes. In virology, understanding the movements of viral RNA through out the cell is of particular interest, with broccoli aptamers already being used to study human immunodeficiency virus (HIV) (Burch et al. 2017) and alphavirus RNA (Nilaratanakul et al. 2017; Nilaratanakul et al. 2020). Recently, it was shown that pairs of RNA aptamers can be employed for FRET experiments, their structures solved using cryo-EM combined with RNA scaffolding techniques (Sampedro Vallina et al. 2023).

It is clear, that nucleic acids are interesting as structured molecules which carry out important functions. Cryo-EM is offering insights into these structures and their slow folding intermediates. Our increasing understanding of these molecules and their dynamics is leading to the rational design of functional nucleic acid polymers that can serve as fluorophores, probes and scaffolds which will enable new avenues of inquiry into both RNA and protein structure.

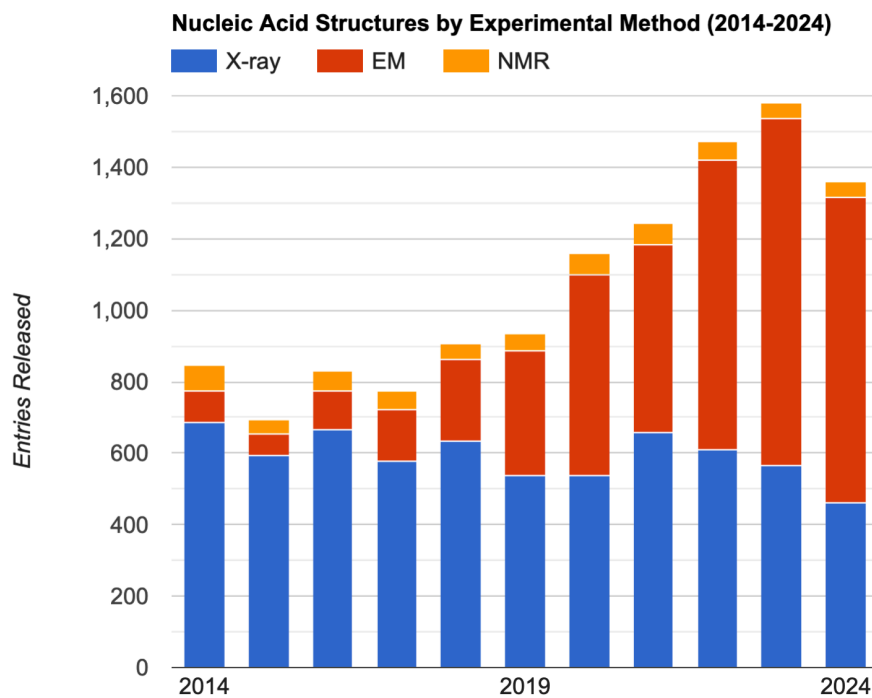


Figure 1.8: The growth of cryo-EM has been rapid with more than two-thirds of deposited models solved using cryo-EM in 2024.

2

Aim

Given the important role cryo-EM is playing in solving the structure of biomolecules and the ever improving methods for structure prediction, it is clear that better tools are needed for fitting these structure predictions to experimental data.

The aim of this work was to develop programs for fitting protein and RNA predictions to experimental cryo-EM data at a range of resolutions with the ability to handle diverse starting models. Thus, such tools must be able to perform significant conformational changes to find an optimal fit. Given, the heterogeneity of experimental maps described earlier, such software models should be able to account for the dynamics or model uncertainty.

These three objectives were realized and are described in the publications in the following sections.

- 3.1 TEMPy-ReFF (Beton*, Mulvaney* et al.) a program for flexible fitting and refinement of atomic models at high to intermediate resolutions which uses Gaussian Mixture Model (GMM) derived force field for flexible fitting and simultaneous local resolution estimation. Furthermore, these local estimations can produce perturbed models to better describe the underlying data as an ensemble.
- 3.2 RIBFIND2 (Malhotra*, Mulvaney* et al.) a program for hierarchical decomposition of protein and RNA structures into rigid-bodies enables large conformational changes.
- 3.3 An application of the previous mentioned software to the refinement of predictions from CASP15 (Mulvaney et al.) demonstrated the efficacy of the method when faced with diverse sets of predictions.

3

Publications

3.1 Publication in Nature Communications

Flexible fitting approaches based on either density guided or correlation driven approaches are well established in the field, and are continuously being optimized and adapted (table 1.2). However, they suffer from a number of shortcomings. Density guided approaches by design steer atoms towards high density and not towards where they are most correlated. This is nicely illustrated by (Orzechowski et al. 2008) and adapted in figure 1.7. Whilst the cross-correlation maximisation approach does push atoms towards correlation it requires an estimation of global resolution in order to proceed. In the following paper (Beton*, Mulvaney*, et al. 2024) which I share first authorship with Dr. Joseph Beton, a Gaussian Mixture Model (GMM) based approach to flexibly fitting models, which we call TEMPy-ReFF, is introduced. One of the main advantages of the approach is that the local resolution of each atom is estimated during the refinement procedure. My own contributions to the project include developing the GMM biasing potential, implementing the bulk of the software including the various force fields, managing the the MD simulations and establishing interfaces for monitoring simulation progress and performing analysis. The majority of the software was written in Python and relies on the OpenMM (Eastman et al. 2017) library for implementing the molecular dynamics components.


The other major aspect of the paper was benchmarking our method against the Phenix (Adams et al. 2010) refinement program. This was made possible by of the CERES (Liebschner et al. 2021) project which offers a repository of models refined by Phenix. Efforts such as CERES are valuable to the method development community and therefore we have deposited our refined models in a Zenodo archive. Such large benchmarks like CERES (and now ours) are rare amongst flexible fitting tools, with most publications providing just a handful of case studies.

Cryo-EM structure and B-factor refinement with ensemble representation

Received: 11 July 2022

Accepted: 20 December 2023

Published online: 10 January 2024

 Check for updatesJoseph G. Beton^{1,3}, Thomas Mulvaney^{1,3}, Tristan Cragnolini^{1,2} & Maya Topf¹ 

Cryo-EM experiments produce images of macromolecular assemblies that are combined to produce three-dimensional density maps. Typically, atomic models of the constituent molecules are fitted into these maps, followed by a density-guided refinement. We introduce TEMPy-ReFF, a method for atomic structure refinement in cryo-EM density maps. Our method represents atomic positions as components of a Gaussian mixture model, utilising their variances as B-factors, which are used to derive an ensemble description. Extensively tested on a substantial dataset of 229 cryo-EM maps from EMDB ranging in resolution from 2.1–4.9 Å with corresponding PDB and CERES atomic models, our results demonstrate that TEMPy-ReFF ensembles provide a superior representation of cryo-EM maps. On a single-model basis, it performs similarly to the CERES re-refinement protocol, although there are cases where it provides a better fit to the map. Furthermore, our method enables the creation of composite maps free of boundary artefacts. TEMPy-ReFF is useful for better interpretation of flexible structures, such as those involving RNA, DNA or ligands.

Cryo-electron microscopy (cryo-EM) can resolve the structure of bio-molecules at an ever-improving resolution. Larger complexes can now be visualised as 3-dimensional density maps at near-atomic resolutions, and in various conformations. The interpretation of those maps often hinges on fitting atomic models of the different macromolecules present in the complex^{1–3}. This procedure is often difficult and requires the user to provide accurate models, and a well-estimated resolution (which can vary at different parts of the map). Pre-existing experimental or predicted atomic models may be in a different conformation, and converging to a well-fitted one may require significant sampling.

Several methods are commonly used for this procedure. To improve the map fit, the map can be treated as a scalar field, for which a gradient can be used as a force^{4,5}. Optimisation of the position against the correlation coefficient (CCC) has also been proposed⁶, or by Bayesian expectation-maximisation (EM) against the density observed in the map^{7,8}. The sampling itself is usually based on either molecular dynamics (MD)^{4,9}, minimisation¹⁰, normal mode analysis and/or

gradient following techniques^{11,12}, or Fourier-space-based methods². Manual inspection and modification of the structure, or targeted sampling for specific parts of the structure, are also common, especially at high resolutions^{13–15}.

Molecular dynamics-based refinement methods have the advantage of wider sampling but may result in locally distorted structures. This can usually be fixed by either clustering the resulting data⁹ or by minimising the structures at the end of the run⁶. The use of a force field (such as CHARMM¹⁶ or AMBER¹⁷) have the added benefit of ensuring that clashes are generally absent from the structure since they include parameterised van der Waals repulsion terms.

Virtually all methods rely on blurring the model (globally or locally)¹⁸ to compare against the experimental map, which poses an additional challenge for maps of flexible systems that will often exhibit significant resolution heterogeneity between flexible and rigid regions. This heterogeneity in the map can also result from adding up density maps from different reconstructions (e.g., result of multibody or focused refinement) into a so-called composite map^{19,20}. However, a

¹Leibniz Institute of Virology (LIV) and Universitätsklinikum Hamburg Eppendorf (UKE), Centre for Structural Systems Biology (CSSB), 22607 Hamburg, Germany.

²Institute of Structural and Molecular Biology, Birkbeck, University of London, London, UK. ³These authors contributed equally: Joseph G. Beton, Thomas Mulvaney. ✉ e-mail: maya.topf@cssb-hamburg.de

systematic way to combine multiple maps into a composite map has not been proposed yet.

Flexibility is intrinsic to biomolecular systems, which presents a challenge for methods that tend to rely on a single structure representation. Methods using a population of models^{21–26} can provide an improved understanding of the fit between map and models²⁷. Mixture modelling is a powerful framework to represent arbitrary density probabilities comprising several parts: by iteratively estimating the model parameters, and then re-computing the expected distribution, a (locally) optimal model can be generated⁸. We use this approach to estimate both the local spread of density around atomic positions and the background noise level.

Here, we propose TEMPy-ReFF (REsponsibility-based Flexible-Fitting)—an MD-based refinement guided by an EM scheme that uses a Gaussian Mixture Model (GMM) to provide self-consistent estimates for the atomic positions and local B-factors (Fig. 1). We show that the method can accurately treat maps with highly heterogeneous resolution. To assess the quality of the refined models, we have developed a measure that estimates the quality-of-fit of every residue to the local density and allows us to compare the fit of different parts of the model in regions of varying resolution. We demonstrate on a large dataset (from the CERES database <http://cci.lbl.gov/ceres> and additional cases from the Protein Data Bank²⁸ (PDB) and Electron Microscopy Data Bank²⁹ (EMDB)), that our approach produces single fits of similar quality compared to state-of-the-art methods, such as Phenix³⁰ although it can sometimes provide improved ones. Importantly, we show that our B-factor refinement approach not only allows for the generation of an ensemble of atomic models to better represent the density information but also enables the generation of more reliable composite maps.

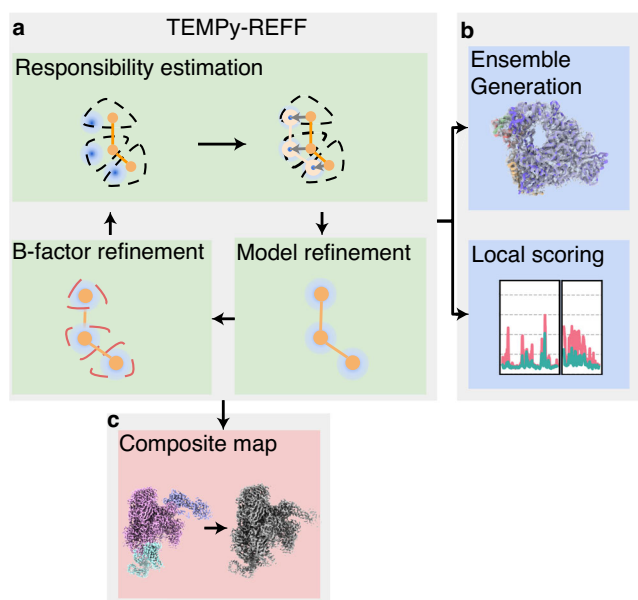


Fig. 1 | Flow chart summarising the steps in the TEMPy-ReFF algorithm. **a** The EM (Expectation-Maximisation) algorithm. Responsibility is an estimation of the part of the data that is represented by a given component in the mixture. New parameters (the mean and variance of each component corresponding to the position and B-factor value) for each component (e.g., for each atom) are then re-estimated using this responsibility and the experimental data. **b** After refinement, an ensemble can be generated based on the local variance; local scoring provides a view of the quality of fit of all regions of the map, irrespective of the local resolution. **c** By considering the sum responsibility of all the atoms in a chain, we obtain a natural expression of the part of a map represented by a given chain. This can be used for composition.

Results

Mixture modelling applied to refinement

We have developed a method based on a GMM (using one Gaussian per atom and a uniform background term) to represent the estimated contribution of various parts of a model to the experimentally observed intensity. The Gaussians are fitted to the model in a self-consistent way, such that their summed contributions represent a (locally) optimal fit to the density. The intensity attributed to a Gaussian, or a sum of Gaussians, can be used to estimate their importance in representing a specific part of the map density. For example, by summing the Gaussians for atoms from a given protein chain, it is possible to determine which part of the map is best represented by this chain, or other chains, or are part of the general background noise in this map. Those weighted contributions (termed responsibilities in GMM literature) allow us to perform a variety of tasks that are commonly performed on cryo-EM maps (described in Fig. 1): fitting an atomic model to the map, segmenting the map into several parts, each representing a distinct entity (for example, a distinct subunit in a protein complex), or combining focused maps into a single overall composite map, with optimal weights of the focused maps.

Although GMM approaches have been successfully employed before, this was usually a coarse-grained representation of the overall model and map^{7,8,31}. By describing each atom as a Gaussian point spread function, a link between map and model is directly established: the intensity of each voxel is a direct sum of the contribution of each atom, as a function of its position and B-factor. It is important to note that we define each atom's "B-factor" as the sigma of its respective Gaussian in the GMM. Additionally, the formalism used here does not require the use of Gaussian distributions, and alternative descriptions for the individual atomic contributions could be considered.

The responsibility calculation has several benefits: for regions of the map that are close to multiple parts of the structures, the mixture model allows for uncertainty in the assignment of the density. This soft-mixing improves the convergence of the refinement, by making it easier for structural elements to slide towards regions of density that are a better fit, even if they are currently fit to a high-density region of the map. The calculation is also self-consistent, as is empirically demonstrated below: changes in the initial position and B-factor assignment for the structure result in identical or similar fit for a wide range of initial values.

Ensemble generation based on B-factors

Our GMM representation models the local ambiguity within cryo-EM maps by tuning the B-factor of each atom. We reasoned that we could leverage this information to generate an ensemble of models that more accurately represents the variety of conformations that are compatible with the map. Models were randomly generated by perturbing the positions of atoms, based on their B-factors, followed by local L-BFGS³² minimisation (with OpenMM³³) to locate close-by structures that were compatible with the data³⁴. Ensemble maps were computed by averaging the simulated maps obtained for all sampled structures in the ensemble (Fig. 2).

We first assessed the accuracy of B-factor assignment in TEMPy-ReFF. While the B-factor optimisation is intended to be used together with position refinement, it is useful to test it independently by optimising the B-factors while keeping the atomic positions fixed. We found that for all cases we tested, the map-model CCC improved significantly when taking into account the refined B-factors for map simulation (Supplementary Table 1). The average B-factor convergence is shown in Supplementary Fig. 1, along with the corresponding change in the CCC using the examples of FabA bean necrotic stunt virus (FBNSV) (EMD-10097, 3.2 Å resolution, PDB ID: 6S44) and the SARS-Cov-2 RNA-dependent RNA polymerase (EMD-30127, 2.9 Å resolution, PDB ID: 6M71). The distribution of the B-factors is similar to that of B-factors obtained from the deposited models (Supplementary

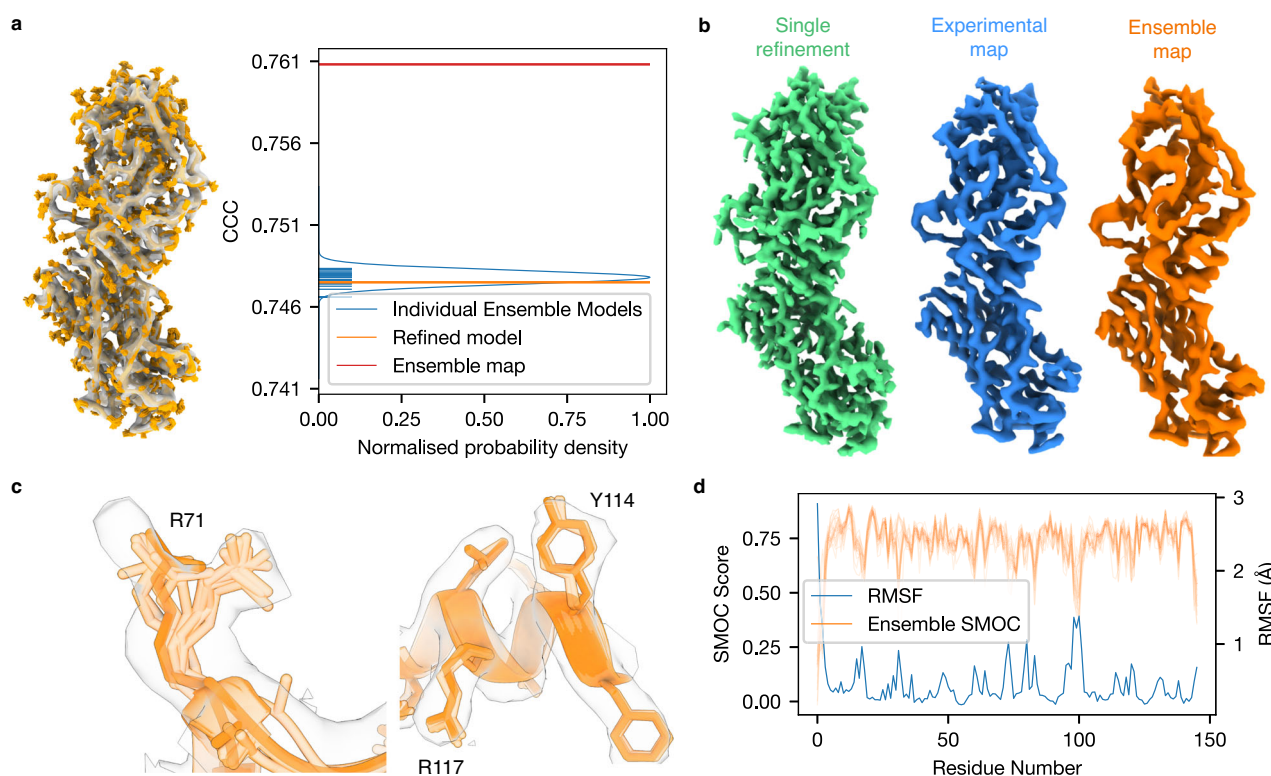


Fig. 2 | Ensemble representation of cryo-EM models. **a** Depiction of structure ensemble (orange), along with the map (transparent grey); a plot of the CCC of each individual model in the ensemble is shown (blue horizontal lines from y axis), as well as the ensemble map (red). **b** Depiction of a single-model map (green), experimental map, and our computed ensemble map at contour level 0.02. **c** Differences in the ensemble for different residues, for the ensemble of the Methionine Transporter (PDB ID: 7MCO, EMD ID: 23752): for residue R71 (left) the

ensemble is more widespread, and the side-chain density is more spread out into two peaks, each populated by parts of the ensemble. For high-resolution portions of the map shown on the right side, for example, R117 and Y114, the ensemble is highly constrained, and the side-chain density is well-defined. **d** SMOCF plot (shown in orange) and RMSF (shown in blue) for each structure in the ensemble for the FabA bean necrotic stunt virus (PDB ID: 6S44, EMD ID: 10097, map resolution 3.3 Å); the RMSF and SMOCF score are clearly anticorrelated.

Fig. 1b). Furthermore, the B-factor assignment is robust: we found that two refinements starting at initial values that differed by a factor of 5 converged to a similar solution (Supplementary Fig. 2). Finally, when we updated the atomic positions, we observed changes in the B-factors, this is a feature of the change in coordinates (as the two are not independent) (Supplementary Fig. 3).

Next, we investigated the use of our calculated B-factors in ensemble generation (Fig. 2). The best-fitted model for rotavirus VP6 (EMD-6272) at 2.6 Å appears as only one solution among many in the generated ensemble (Fig. 2a). On the other hand, the ensemble average map exhibits a much higher quality-of-fit to the experimental map than any single model (Fig. 2a, b, Supplementary Fig. 4). Intriguingly, the ensemble map resembles more closely the experimental map (Fig. 2a). We determine the optimal number of models in an ensemble by calculating the CCC with the ensemble map generated from an increasing number of models (Supplementary Fig. 5). A visual comparison between the single TEMPy-ReFF refined model and the ensemble is shown in Fig. 2c where insets of residue fit show the source of improvements: the density for an arginine (R71 from chain A) could be explained by positioning the side chain in two alternate conformations. The structures in the ensemble populate both possible conformations (Fig. 2c, left inset). In contrast, the ensemble of models is much more tightly clustered in well-resolved portions of the map, for example, residues R117 and Y114 from chain A (Fig. 2c, right inset). We also found, using the capsid protein from the FabA bean necrotic stunt virus (PDB ID: 6S44, EMD ID: 10097, map resolution 3.3 Å), that the per-residue SMOCF³⁵ score (averaged between all ensemble members) showed a strong anti-correlation with the RMSF between the ensemble measures (Pearson's coefficient -0.81 , Fig. 2d).

Benchmarking structure refinement

We assessed the quality of TEMPy-ReFF model refinement using a large dataset of 229 models taken from the PDB (see Methods) with corresponding maps at resolutions between 1.8 and 5 Å. We compared the CCC, MolProbity³⁶, and CaBLAM³⁷ scores before and after refinement. We benchmarked our method against the deposited PDB models as well as CERES³⁸ (see Methods), which is an automated Phenix³⁰ model re-refinement programme for cryo-EM maps at resolution ≤ 5 Å.

We observed, overall, similar performance between TEMPy-ReFF and CERES based on map-model similarity (CCC) and geometric model quality scores (MolProbity, CaBLAM, clash score) (Fig. 3, Supplementary Table 2). The average CCC scores for refined models from maps with a resolution range of 3–4 Å from TEMPy-ReFF (median: 0.633, mean \pm std: 0.627 ± 0.101) and CERES (median: 0.636, mean \pm std: 0.637 ± 0.087) were very similar (Fig. 3a). We only observed improved average CCC scores from TEMPy-ReFF refinements for models refined in maps at 4–5 Å resolution (mean CCC \pm std from TEMPy-ReFF: 0.672 ± 0.148 , CERES: 0.651 ± 0.147). However, we observed improved (lower) average MolProbity scores in many TEMPy-ReFF refined models. Specifically, the MolProbity scores for TEMPy-ReFF refined models from the highest resolution maps (< 3 Å), outperformed both CERES and models obtained from the PDB. Additionally, we noted a smaller improvement in MolProbity scores for models in the 3–4 Å resolution range. This was largely due to the almost total absence of clashes in TEMPy-ReFF refined models (Supplementary Table 2). However, we noted more CaBLAM outliers in TEMPy-ReFF refined models. Further, we observed a higher correlation between MolProbity score and map resolution (i.e., increasing MolProbity score as map resolution worsens) for TEMPy-ReFF refined models compared to

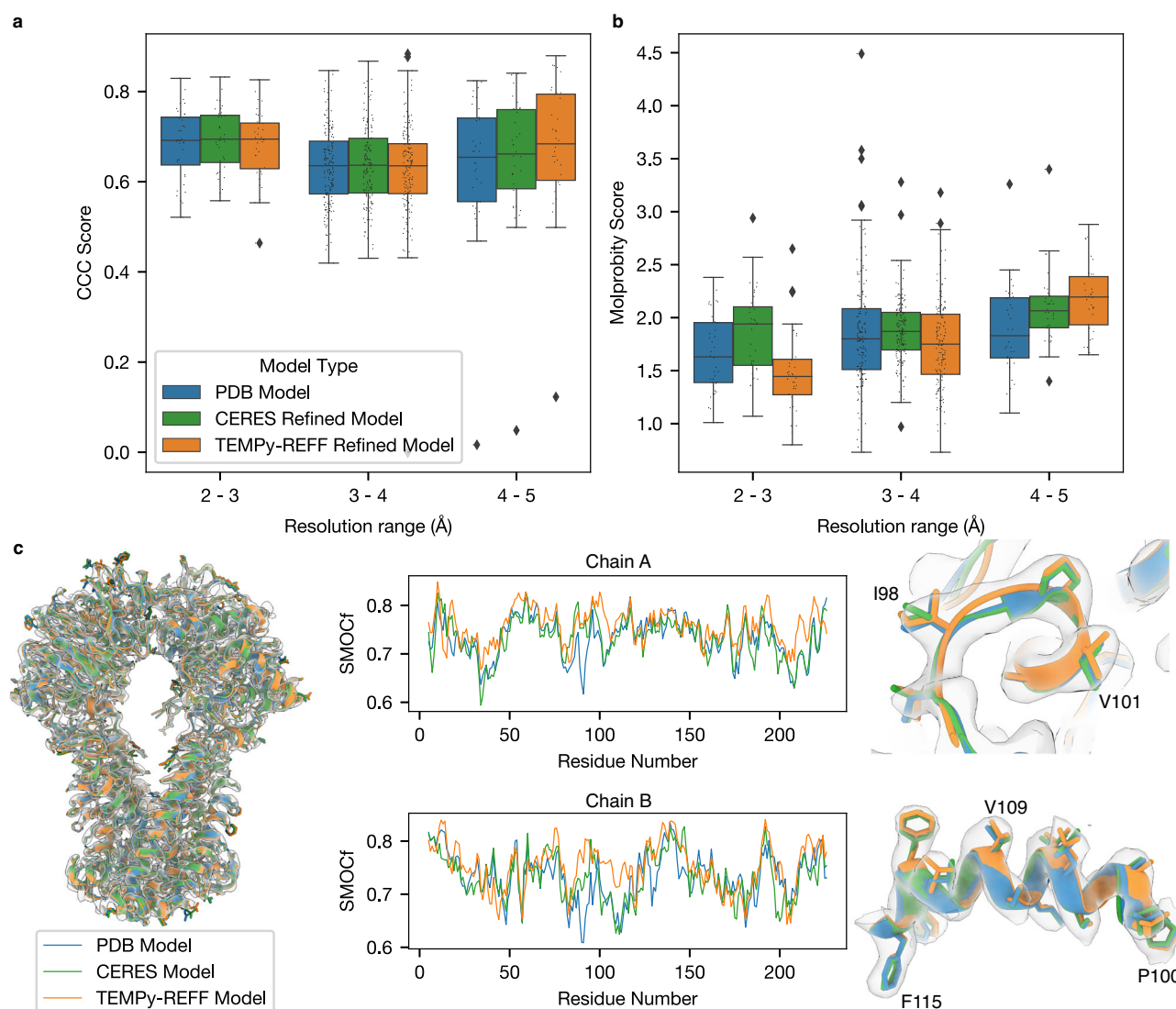


Fig. 3 | Refinement of the CERES benchmark. a Benchmark comparison using CCC, between the initial (PDB-deposited) models (blue), the CERES re-refined models (green), and TEMPy-ReFF refinement-based model (orange), separated by resolution bands of 1 Å. We evaluated $n = 229$ individual models. The central line in each boxplot defines the median value, the bounds of each box define the upper and lower quartiles and the whiskers define 1.5 times the interquartile range (IQR). Outliers (points outside this 1.5*IQR range) are marked with rhombus symbols. The individual score for each model is marked with a black point. **b** Benchmark comparison of the same 229 models using MolProbity score, the colouring and layout of

the boxplot is the same as in **a**. **c** Comparison between the refinement of the ABC methionine transporter (PDB ID: 7MCO, EMD ID: 23572, resolution 3.3 Å) with TEMPy-ReFF and the corresponding models from PDB and CERES. For all subpanels the colouring matches that used in **a**. The left panel shows the overlaid models within the cryo-EM map, which is rendered as a transparent surface. The central panels show the SMOCF scores for residues from chains A (upper panel) and B (lower panel). The left-hand panels show zoomed-in views of sections of chain A (upper panel) and B (lower panel) as highlighted in the respective SMOCF plots with black outlined boxes.

those obtained from the PDB and CERES (Supplementary Fig. 6). This might be due to geometric restraints that are commonly applied in other refinement software, including in CERES³⁸, but not in TEMPy-ReFF, where the geometry of the model is derived from the energy function and the MD force field.

We examined the local fit quality using SMOCF for one example from our benchmark: the ABC methionine transporter, solved at 3.3 Å (PDB ID: 7MCO, EMD ID: 23572). This example showed that local model fit for the TEMPy-ReFF refined model was similar overall, relative to those from the PDB and CERES (Fig. 3c). Some parts of the TEMPy-ReFF models showed better fit, and others poorer. This was perhaps unsurprising, given the overall similar performance across our benchmark at this resolution range (Fig. 2a). In areas where we did observe better local fit for TEMPy-ReFF refined models, this was apparently due to subtle changes in the positioning of the backbone and the orientation of side chains (Fig. 3c).

We next investigated the degree of structural rearrangement that was possible during TEMPy-ReFF refinement. We identified structures deposited in the EMD/PDB of which two separate conformations were identified. First, we analysed two structures of the Atm1 ABC transporter, in an open and closed conformation (EMDB IDs: 13613, 13614 at 3.3 and 3.2 Å resolution, respectively and corresponding PDB IDs: 7PSL, 7PSM, respectively)³⁹. We observed that large structural rearrangements (e.g., rotation of whole domains) would be required to refine the structure of closed conformation into the cryo-EM map of the open conformation (i.e., to refine the 7PSM into EMD-13613). Despite an increase in CCC from 0.15 to 0.31, refinement with TEMPy-ReFF was not able to reproduce the structure of the open conformation, presumably because the model became stuck in local minima (Supplementary Fig. 7a). In a previous study, we developed a method that combined density-guided-refinement (which is similar to MDF⁴⁰), with the hierarchical application of rigid-body restraints calculated

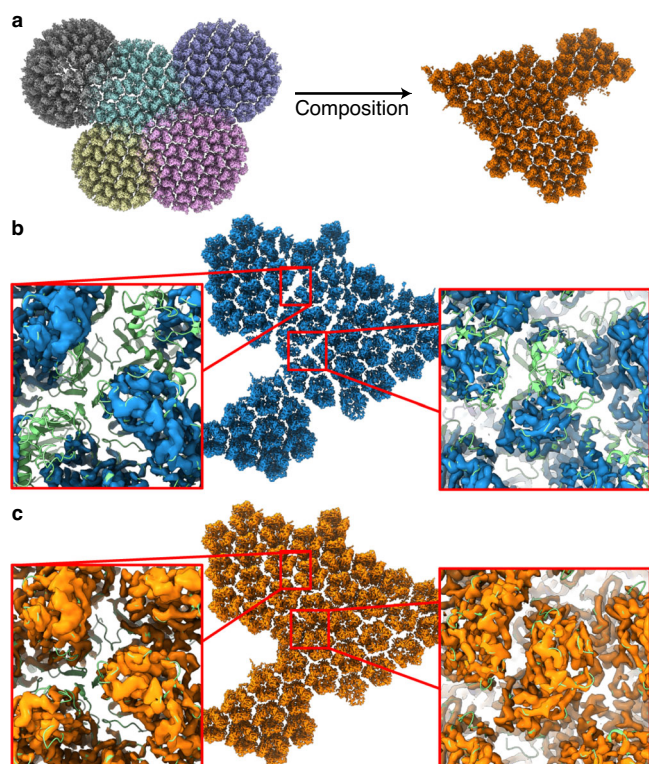


Fig. 4 | Using TEMPy-ReFF for map composition. **a** Composition of 5 component maps (EMD-34227, 34229, 34230, 34235, 34236) shown in their overlapping position on the left, combined to create the composite map shown on the right. **b** Composite map of the Singapore grouper iridovirus capsid (EMD-34815), shown as a blue surface rendering. In order to simplify visual comparison, we masked the original map such that only density around the fitted model (PDB ID: 8HIF) is shown. The deposited composite map retains some artefacts at the borders between the, approximately circular, component maps, where the map density is less intense. This is highlighted in the insets, which also show the fitted model, coloured green. **c** Composite map, shown as an orange surface rendering, produced using the responsibilities computed by TEMPy-ReFF as weights for each component map. The insets show the map density with the model, again shown in green, at the same location as shown per **b**. Clearly, the artefacts are no longer present.

using RIBFIND2 (version 2.0). This method was able to correct large structural changes in RNA complexes^{41–43}. We applied this method to the refinements of Atm1 and were able to successfully refine the model from the closed conformation into the open cryo-EM map (Supplementary Fig. 7a). The CCC was 0.34 after rigid-body refinement. We noticed some errors remained in the model, such as slightly incorrect placing of α -helices and amino acid side chains. To fix these issues, we ran an extra round of refinement using TEMPy-ReFF, which further improved the model to a final CCC of 0.54. We observed a similar outcome for refinement of the open conformation CGT ABC transporter⁴⁴ (EMDB ID: 14843, PDB ID: 7zo8) into the cryo-EM map for the closed conformation (EMDB ID: 14844, PDB ID 7zo9): refinement was only successful when combined with the application of hierarchical RIBFIND2 restraints (Supplementary Fig. 7b). Thus, we conclude that TEMPy-ReFF refinement, without additional rigid-body restraints, is best suited for refinement that requires local changes in the model, for example, arrangement of secondary structure elements and positioning of side-chains.

B-factor weighted composite maps

We hypothesised that our GMM approach for model representation could be applied to generating composite maps, where one combines multiple, potentially overlapping, reconstructions of the same

complex into a single map. This can be viewed as an inverse of the mixture modelling problem, where the intensity contributions of each component map must be correctly mixed together to produce an accurate composite map. We achieved this using our GMM representation to calculate responsibilities for every voxel in each component map (Eq. 9), such that portions of component maps that corresponded to atoms with lower B-factors were assigned the highest responsibilities. These responsibilities acted as weights for combining the component maps (Eq. 10). Our approach has several advantages: because the responsibility decays smoothly, there are no seams within composite maps and areas where the assignment would be uncertain are treated as such, and the density will not be arbitrarily assigned to a specific model or submap.

We evaluated our approach on a composite map of the Singapore grouper iridovirus capsid (EMD-34815) (Fig. 4). This map is composed of 5 component maps, with each overlapping significantly with at least 2 other component maps (Fig. 4a), using Chimera^{45,46}. Circular artefacts were visible in the deposited map, which occurred at the edges of the individual component maps, including at areas of the map containing a fitted model (Fig. 4b, Supplementary Fig. 8a). After generating a composite map using our responsibility-weighted approach, we found no visually distinguishable artefacts at equivalent locations in our composite map (Fig. 4c). This was reflected in a general increase in correlation between Fourier components of the TEMPy-ReFF composite map and the deposited model, compared to between the deposited map and model (Supplementary Fig. 8a). Additionally, the CCC between the model and TEMPy-ReFF composite map improved to 0.79, compared to 0.71 for the deposited map.

We extended our evaluation to composite maps which did not include visually obvious reconstruction artefacts by reproducing the composite map of RNA polymerase II (EMD-12969), composed from 3 separate maps (EMD-12966, EMD-12967, EMD-12968)⁴⁷. Here, we again see a general increase in correlation between Fourier components in the TEMPy-ReFF composite map and the model, as well as an increase in the model CCC score to 0.61, from 0.51 for the deposited map (Supplementary Fig. 8b).

Case study 1: yeast RNA polymerase III elongation complex

We explored the effectiveness of the TEMPy-ReFF approach in more detail by refining the model of yeast RNA polymerase III elongation complex (PDB ID: 5FJ8). The corresponding cryo-EM map (EMD-3178) was resolved at a global resolution of 3.9 Å⁴⁸. A brief observation of the deposited model suggests that it is well-fitted to the cryo-EM map: we computed the CCC, using ChimeraX, as 0.58. The validation statistics presented in the PDB are reasonable; clash score of 14, Ramachandran outliers 1.1% and side-chain outliers 2.1%, with an overall MolProbity score of 2.8.

The TEMPy-ReFF refined model had an improved correlation with the map, with a single-model final CCC of 0.62, whilst the ensemble map had a CCC of 0.70. The MolProbity score remained essentially unchanged at 2.7. A representation of the model, as well as the quality-of-fit for multiple chains, is shown in Fig. 5.

We next applied the TEMPy LoQFit score (see Methods) to locally assess the improvement of our TEMPy-ReFF refined model, versus the deposited model. Here, we only use the single-refined model from TEMPy-ReFF to ensure fair comparisons. We visualise the LoQFit score at each residue in both models using 2D plots (Fig. 5). The average LoQFit score for the deposited model was 5.1 Å, and model agreement was particularly high in chains A and B at the central regions of the model and map, where the average LoQFit score was 4.6 and 4.5 Å, respectively. However, even in these regions we observe peaks in the LoQFit score, consistent with poorer model fit, such as those seen around residues 192–210 and 745–759 in chain A (Fig. 5d), as reflected in the higher B-factors in this region (Fig. 5c). In addition to this, we identify extended regions of poorer model fit, generally occurring

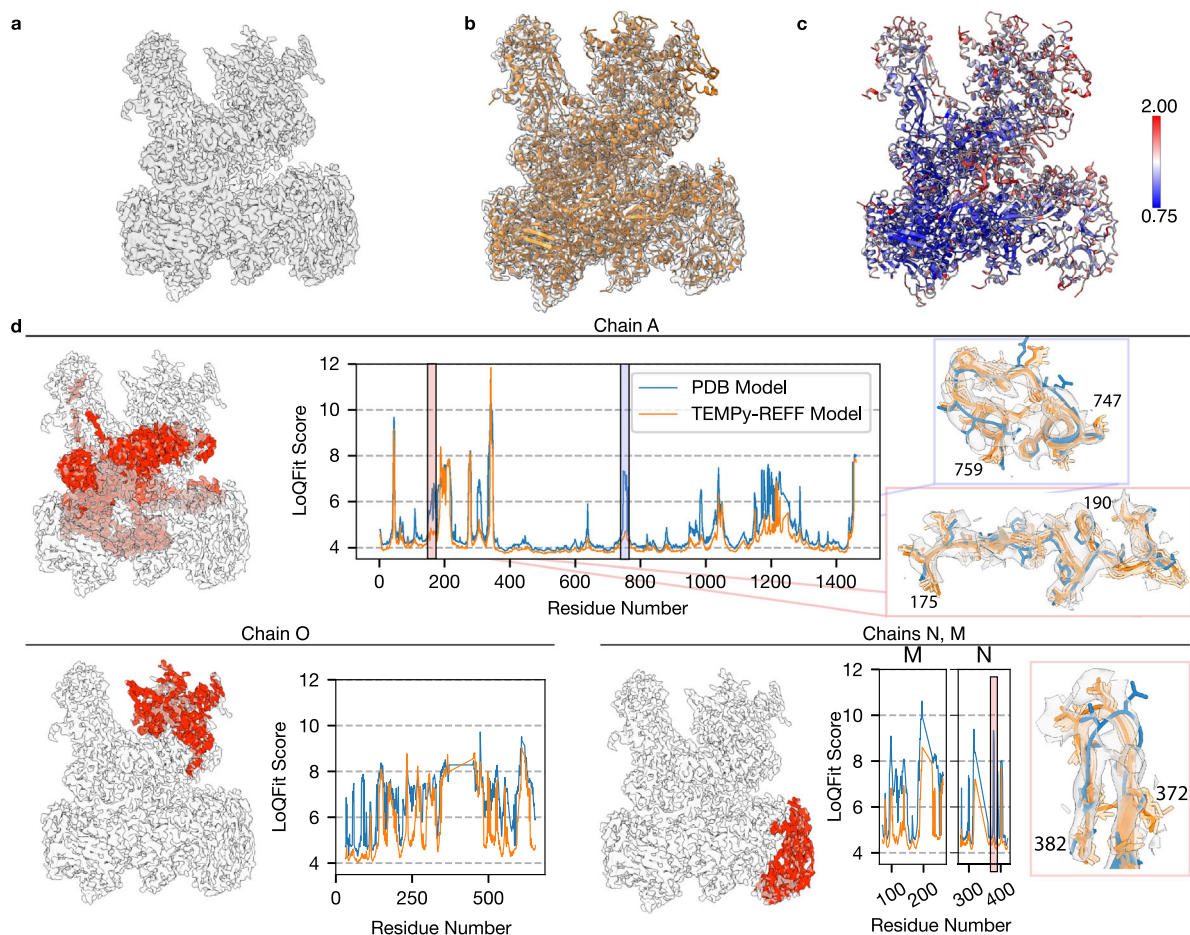


Fig. 5 | Case study of RNA polymerase III elongation complex. **a** The deposited 3.9 Å cryo-EM map of the RNA polymerase III elongation complex (EMD-3178). **b** The TEMPy-ReFF refined model of the RNA polymerase III complex deposited structure (PDB ID: 5FJ8) shown within the cryo-EM density. **c** The TEMPy-ReFF refined model (right) coloured according to the refined B-factors. **d** LoQFit scoring of individual chains from the RNA polymerase III complex, with the scores for the

starting model (obtained from the PDB) shown in blue, and for the TEMPy-ReFF refined model shown in orange. The position of these chains within the original cryo-EM map are highlighted in red. Insets show several regions before and after refinement coloured as per the LoQFit plots, with the ensemble of models shown in transparent orange.

within chains that lay at the edge of the complex in solvent-exposed regions with poorer resolution, including chains M and N (Fig. 5d). In these chains the average LoQFit score was 6.6 and 5.4 Å, respectively, reflecting the lower map resolution (and correlating with high B-factors), as well as poorer model fit in the deposited model. Refinement with TEMPy-ReFF resolved many of these poorer fitting regions: the average LoQFit score for the refined model improved to 4.6 Å, and we observed significantly better model fit at lower resolution regions of the map. The average LoQFit score for chain O improved to 5.7 Å in the refined model (from 6.8 Å, Fig. 5d), and in chains M and N the average LoQFit score improved to 5.1 and 4.5 Å after refinement. We investigated the significance of these changes in the LoQFit score. Firstly, we observed a close correlation between the LoQFit score and the local resolution at the equivalent position within a cryo-EM map (Supplementary Fig. 9). Secondly, we benchmarked LoQFit against other common local scoring functions, Q-score and SMOC, as well as our B-factor refinement. For Q-score and B-factors, we used the residue average (Q-score_{avg}), for comparison. To do this benchmarking, we measured the LoQFit, Q-score_{avg}, SMOCf and B-factors for 50 models refined by TEMPy-ReFF, and investigated the correlation between LoQFit and the other scoring functions via Pearson's correlation. This revealed a significant, inverse, correlation between LoQFit and Q-score_{avg} (−0.62 Pearson's correlation across all examples), and a significant correlation between LoQFit and the residue average B-factor (0.64 Pearson's correlation across all examples). We observed

a much less significant correlation with the SMOCf score (0.32), which varied much more significantly across the examples we tested, compared to the correlation between LoQFit and Q-score_{avg} and average B-factor (Supplementary Fig. 10). This was unsurprising, given the previously reported lack of correlation between the Q-score and SMOCf⁴⁹.

Case study II: nucleosome-CHD4 complex structure

The nucleosome is a large nucleoprotein present in the nucleus, which is the primary effector in the compaction of DNA. High-quality reconstructions have been obtained, but its dynamic nature and strained DNA strands wound around the histone proteins make it a challenging system to obtain a good structural model. We apply TEMPy-ReFF to refine the model associated with map EMD-10058⁵⁰ (PDB ID: 6RYR) (Fig. 6a–d). The deposited cryo-EM map clearly suffers from very variable resolution (range: 3–10 Å, see Supplementary Fig. 9), which affected the quality-of-fit of the deposited model (Fig. 6a). Following refinement, the local details of the map are well respected, especially showing improvement in the DNA structure, as reflected by the SMOCf score (chain I and J, Fig. 6c). Nucleic acids are often present in biomolecular complexes resolved by Cryo-EM, and refining their geometries with respect to the map is an important part of model refinement. In the deposited model, local deformations pull the bases slightly away from the density, and from the expected geometries to allow hydrogen bond formation. Our automated refinement

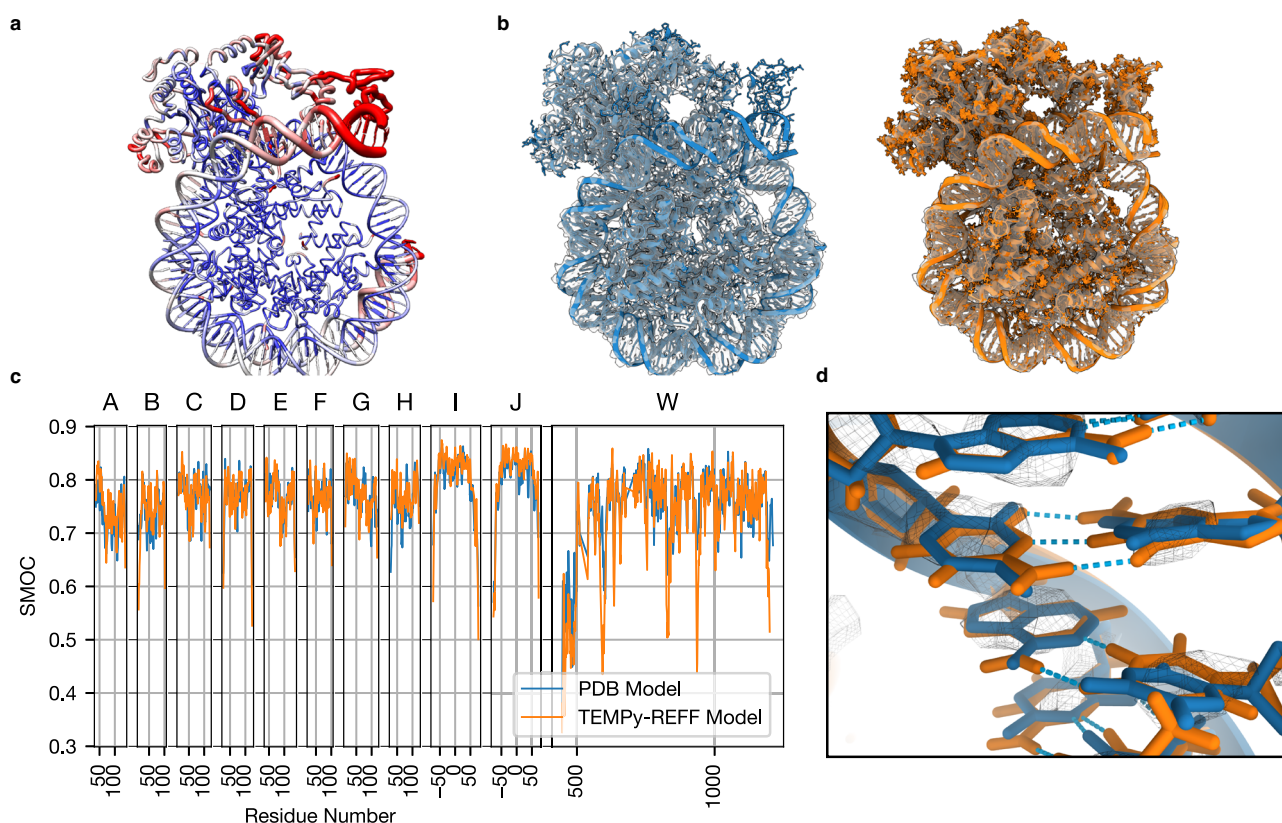


Fig. 6 | Case studies of Nucleosome-CHD4 complex. **a** A nucleosome structure in complex with chromatin remodelling enzyme CHD4 (EMD-10058, PDB ID: 6RYR) is shown (worm representation), with the width proportional to the TEMPy-ReFF refined B-factor, and colour based on local resolution (computed with ResMap). **b** Deposited model (left, blue) and the ensemble of models and ensemble map calculated with TEMPy-ReFF (right, orange), shown inside the cryo-EM map

(transparent grey). **c** SMOCf plot for each chain. The deposited model is shown in blue, and the TEMPy-ReFF model is shown in orange. **d** Zoom-in on some of the DNA base pairs (chain I/J, base pair 54) fitted in the map (mesh representation). The deposited model is shown in blue, TEMPy-ReFF model in orange and hydrogen bonds are indicated in cyan.

pulls them back, forming hydrogen bonds in the process (Fig. 6d). After refinement, the LoQFit and the local resolution follow similar trends (Supplementary Fig. 9), indicating the model is well fit to the map. This case study also further demonstrates how the ensemble map calculated with TEMPy-ReFF has greater similarity with the experimental map than a single model (either the deposited model or a single-refined model).

Case study III: SARS-CoV2 RNA polymerase and AlphaFold2

To refine a model into an experimental cryo-EM map, an initial model is needed. Although building a reliable model directly from the map is sometimes possible, in most cases, this cannot be done reliably as the resolution is not sufficient to allow a reliable assignment of every atomic position. In such cases, a starting model can be obtained using deep-learning-based *ab initio* tools, such as AlphaFold2⁵¹ or RosettaFold⁵². These programmes are frequently able to create very high-quality protein models⁵³. The predicted IDDT score⁵¹ (pIDDT) is also an excellent tool to decide which part of the model can be reliably kept, and which may not be correctly predicted, due to flexibility or lack of known homologous sequences and structures.

To assess the capability of our method to refine such a model, we used AlphaFold2-Multimer⁵⁴ to create a model of the SARS-Cov-2 polymerase. We used the polymerase sequence (UNIPROT ID: PODTD1, residues 4393–5324), with non-structural proteins 7 (UNIPROT ID: PODTD1, residues 3860–3942) and 8 (UNIPROT ID: PODTD1, residues 3943–4140). We only used templates present in the PDB at least a year earlier than the deposition date of the deposited model (PDB ID: 6M71)⁵⁵. The predicted model was refined into the SARS-Cov-2 polymerase cryo-EM map at 2.9 Å resolution (EMD-30127) (Fig. 7). The

resulting model (Fig. 7d) is highly similar to the deposited model (Fig. 7c) at most residue positions, which was modelled using Chimera⁴⁶, Coot¹⁴, and Phenix³⁰. However, more intriguingly, using a SMOCf plot, we show that some residues that were not present in the deposited structure⁵⁵ can actually be placed into the map, with fitting scores much greater than chance (Fig. 7c, d).

Discussion

We have presented TEMPy-ReFF, an MD-based atomic structure refinement method, which is driven by the local features of a cryo-EM map using a mixture model with an error term, to account for the noise in the map. Our approach naturally incorporates both position and B-factor estimations in the same framework. This information is essential to represent the local variability around atomic positions. We conducted comprehensive testing on a substantial dataset comprising 229 cryo-EM maps sourced from EMDB, spanning resolutions from 2.1–4.9 Å and their respective PDB and CERES atomic models. On a single-model level, TEMPy-ReFF achieves performance similar to the CERES re-refinement protocol, and in some instances, outperforms it by providing a more accurate fit to the map.

Currently one of the greatest challenges in model building into cryo-EM maps is evaluating the quality-of-fit in a system not described by a single resolution value, but rather varying local resolution. We address this challenge using B-factor estimation. We find, as previously shown^{21–23,25,26}, that an ensemble of equally well-fitted models represents this local variability better than a single model. However, we go one step further, by showing that an ensemble map calculated from these models, provides a better representation of the experimental map, in comparison to a traditional simulated map (which is typically

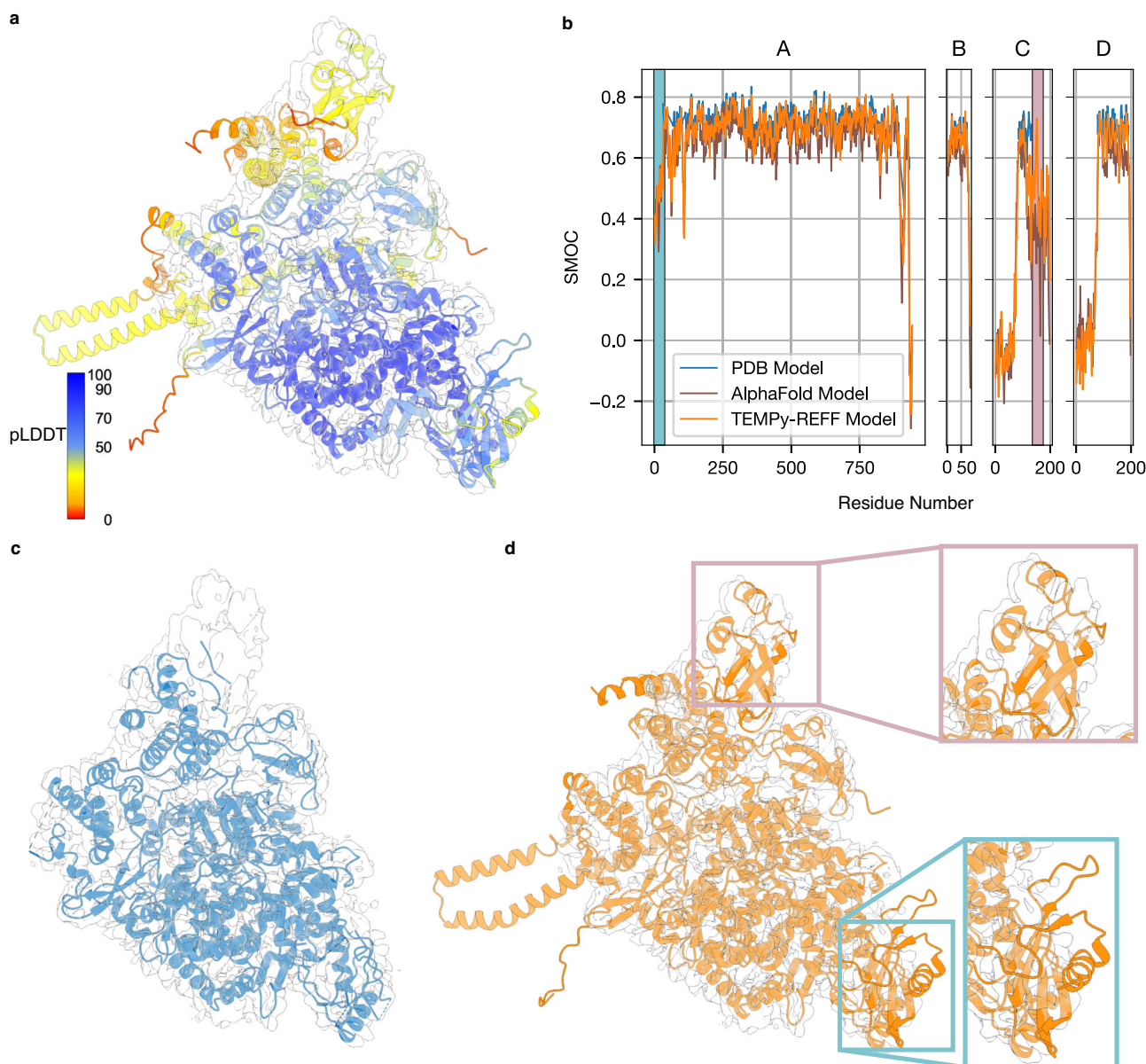


Fig. 7 | Case studies of SARS-CoV-2 RNA polymerase (AlphaFold2 model refinement). **a** AlphaFold2 predicted structure, with the colouring indicating the pLDDT confidence measure (blue means higher confidence, red means lower confidence), fitted in the deposited map (EMD-30127, grey) **b** SMOCf plot of the AlphaFold2 (shown in blue) and TEMPy-ReFF refined model (shown in orange). The regions highlighted in grey and pink (correspond to inset regions in Fig. 7d) contain residues that are not present in the deposited model but are present in the

AlphaFold2 model and are well-fitted to the map. **c** Deposited model for the SARS-CoV-2 RNA polymerase (PDB ID: 6M71, blue) fitted in the deposited map (transparent grey). Unassigned regions are visible, at the top and bottom right of the map. **d** TEMPy-ReFF model (orange) obtained by refining the AlphaFold2 prediction in the deposited map (transparent grey). Newly modelled regions that fit in the density (as in Fig. 6h) are shown with coloured squares.

generated from a single Gaussian function per atom) (Fig. 2a). This is showcased in Fig. 2c, where a potential double occupancy site for an arginine necessarily requires more than one model to be correctly represented. The improvement is also evident in regions of lower local resolution (Supplementary Fig. 4), which may indicate an inherent local flexibility of the structure, although this cannot be easily deconvolved from the blurring due to optical factors⁵⁶, or image processing approaches.

Ensemble methods have been common practice in the NMR community and have been suggested as a way of dealing with the uncertainty in the data^{22,23,34}. This has also been demonstrated previously for X-ray crystallographic data⁵⁷, and we similarly observe a plateau as more models are added to the ensemble (Supplementary Fig. 5). Furthermore, when analysing the differences on a local level

(for example at the residue level) using a distance measure (such as the RMSF), we observe that the local-fit-quality (using SMOCf) correlates well with those differences (Fig. 2d).

Overall, our automated refinement procedure is computationally efficient: computation time scales approximately linearly with map and model size (Supplementary Fig. 11). The resultant models are well-fitted to the cryo-EM map, based on the CCC. Without the ensemble representation of the fitted models, the local and global model-map fit score is comparable with those from Phenix (as represented by our comparison with CERES results). We also observed that TEMPy-ReFF refined models have typically as good, or better, MolProbity scores, compared to those from CERES and the PDB, across our benchmark (Fig. 4b). However, the correlation between resolution and MolProbity score was stronger for TEMPy-ReFF refined models, compared to those

from CERES (Supplementary Fig. 6). This difference is likely due to a different application of explicit structural restraints in CERES, compared to TEMPy-ReFF. Our refinement procedure does not include any specific restraints, for example, to reduce Ramachandran or rotamer outliers. Rather, models refined by TEMPy-ReFF are implicitly restrained by the balances of forces applied to the atoms by the force field. This should produce models with appropriate geometry, assuming the fitting force from the GMM is appropriately balanced within the MD force field. Indeed, the generally good MolProbity scores obtained in our benchmark (Fig. 4b) show this to be an appropriate approach. In particular, we noted that TEMPy-ReFF refined models virtually never contained significant clashes (Supplementary Table 2). However, many refinement programmes, including those used for CERES models, do apply geometric restraints (e.g., to eliminate phi/psi outliers). Based on our results, it seems that, broadly, these restraints favour reduced CaBLAM outliers, which are typically better for PDB/CERES models, at the expense of clash scores, which were consistently worse in PDB/CERES models compared to those from TEMPy-ReFF (Supplementary Table 2). We also show that TEMPy-ReFF refinements of nucleic acids can simultaneously improve the fit to the cryo-EM data and the chain geometry (Fig. 6a–d).

Since 2018, deposition of composite maps has been increasing significantly due to a growing number of macromolecular assemblies for which focused maps for different assembly subunits are obtained (often due to conformational flexibility). Some methods have been proposed to compose such maps²⁰, however, there is currently no systematic way to evaluate this. Here, we provide a self-consistent way to perform this procedure. Our approach has the advantage that the responsibility decays smoothly, i.e., there are no seams between segmented maps, or within composite maps: areas where the assignment would be uncertain are treated as such. However, the method also has some drawbacks, the clearest of which is that errors in modelling will result in errors in composition, and that the maps must be aligned manually, or using another software, prior to composite map generation with TEMPy-ReFF.

Finally, we show that our refinement protocol can take advantage of recent developments in the field of structure prediction^{51,52}. Starting refinements from AlphaFold2^{51,52} models is not only possible, it gives results on par with manual refinement (despite using an automated procedure) and highlights that better and more complete models can be obtained by using our automated refinement approach, including more residues that are sustained by the map information (Fig. 6e–h). However, we note that models that contained large errors required the application of rigid-body restraints for effective refinement (Supplementary Fig. 7). For these refinements, the TEMPy-ReFF GMM-based (unrestrained) refinement still played an important role in correcting minor errors that existed after rough refinement with rigid bodies. It is difficult to define an exact transition point at which rigid-body refinement, instead of unrestrained, is required for a given model, and this currently requires user intervention. However, we envisage a flexible and automated combination of these approaches could pave the way for more reliable, and reproducible model building, where alterations in refinement protocols can be objectively and continuously assessed^{53,58}.

Further work will be needed to understand the impact of ensemble model representation, and how to use such an approach in assessing model-map fit quality, especially for inherently flexible protein assemblies observed by cryo-EM. In this work, we explore how ensembles can be derived from local resolution information using our GMM interpretation of the experimental data. Although we are able to derive ensembles that improve the overall correlation with cryo-EM map, the model is admittedly simplistic. Assumptions that the Gaussians are isotropic and that resolution fluctuations are a result of conformational heterogeneity are approximations.

Indeed, future work needs to be able to disentangle resolution heterogeneity due to reconstruction and imaging artefacts from that caused by atomic displacements and structural variation. It is foreseeable that this will require an end-to-end approach where more information from reconstruction and the underlying 2D micrographs are used to address these challenges. Despite these limitations, we see this work as an important step, particularly in the field of drug discovery, where, the docking of candidate compounds is dependent on the local environment, and local errors or variability can significantly alter the results. Providing multiple models of cryo-EM maps from near-atomic to medium-resolution will allow more reliable predictions of ligand poses, thereby opening a window to many potential drug targets in medium-resolution cryo-EM maps.

Methods

Refinement algorithm

Given an atomic model, which can be described as a set of atoms each possessing a coordinate x , a B-factor B and an atomic numbers Z , the aim is to optimise these positions and B-factors to best model the experimental data. The refinement algorithm is inspired by the EM approach for GMMs⁵⁹. Here, atoms are represented as Gaussians with the centre of mass and B-factor represented by the mean of the Gaussian and sigma, respectively. Per the standard EM algorithm, we first compute the expected (simulated) map given the estimated atomic properties. A maximisation step is then performed to optimise the atomic properties. Traditionally, the maximised properties would be fed back to the expectation step and the EM process would be repeated until convergence. In order to incorporate stereochemical and physical information, we deviate from the standard EM algorithm: Rather than feed the maximised atomic properties back into the next expectation step we compute a force that biases atoms towards the optimised coordinates in an MD simulation. The algorithm is summarised below:

- Perform maximisation step
 - Generate the expected (simulated) map given a set of initial atomic positions, B-factors, and background error.
- Perform expectation step
 - For each atom determine a new desired position and B-factor.
 - Update the background noise term.
- Update the biasing force to encourage atoms towards the new positions.
- Repeat until convergence criteria are satisfied.

Expectation

The intensity ' P ' due to a given atom ' i ' at a coordinate v can be modelled as a Gaussian where \bar{x}_i , B_i and Z_i are the atoms positions, B-factor and atomic number, respectively:

$$P(\vec{v}, \bar{x}_i, B_i, Z_i) = Z_i e^{-\frac{|\vec{v} - \bar{x}_i|^2}{B_i}} \quad (1)$$

For brevity, we abbreviate the above equation for a given atom:

$$P_i(\vec{v}) = P(\vec{v}, \bar{x}_i, B_i, Z_i) \quad (2)$$

Now, the expected intensity of a given voxel in a cryo-EM map M_s (referred to as the simulated map) is given by the contributions of all N atoms with an additional error term E which will be introduced later:

$$M_s(\vec{v}) = \sum_i^N P_i(\vec{v}) + E \quad (3)$$

Maximisation

The maximisation step attempts to determine updated parameters that improve the simulated map in the next ‘expectation’ round. To perform the maximisation step for each atom a responsibility-weighted experimental map $W_i(\vec{v})$ is calculated for each atom. The responsibility for a given atom (γ_i) is given by:

$$\gamma_i(\vec{v}) = \frac{P_i(\vec{v})}{M_s(\vec{v})} \quad (4)$$

Next, the experimental map M_e is weighted by this responsibility:

$$W_i(\vec{v}) = M_e(\vec{v})\gamma_i(\vec{v}) \quad (5)$$

The new position x_i' of the i 'th atom is given by the weighted real-space average of the voxels, where \vec{v} is the real-space position of the voxel.

$$x_i' = \frac{1}{\text{tot mass}} \sum_{v \in V} W_i(\vec{v}) R(\vec{v}) \quad (6)$$

The new B-factor B_i' is given by the weighted variance.

$$B_i' = \frac{1}{\text{tot mass}} \sum_{v \in V} W_i(\vec{v}) |\vec{v} - \bar{x}_i|^2 \quad (7)$$

Due to experimental noise, atomic B-factors are often restrained^{10,60}. Here, we apply a simple weighting scheme, where the average B-factor of all atoms in a residue is used to weight the atoms.

The new estimate of the background noise E' is also calculated as the mean of the experimental map weighted by the responsibility of the error, where $|V|$ is the total number of voxels. Here, only voxels within 4σ of the atoms are included in the calculation. This ensures that the noise term isn't biased by density values that are not near the refined atoms.

$$W_{err}(\vec{v}) = M_e(\vec{v}) \frac{E}{M_s(\vec{v})} \quad (8)$$

$$E' = \frac{1}{|V|} \sum_{v \in V} W_{err}(\vec{v}) \quad (9)$$

Defining the fitting potential

After determining improved parameters for the atoms, the force field used to steer them is updated. We consider two methods to improve the fit quality: MD, where the system's coordinates are integrated over time, taking into account the forces atoms exert on each other; and energy minimisation, where the coordinates of the system are changed to minimise the energy function.

To combine our description of the map with the energy terms that are usually present in force fields, we compute a fictitious force representing the direction of the change in position induced by the Gaussian fitting (for MD). The energy term (E_{gmm}) is defined as:

$$E_{gmm} = k_{gmm} \left(1 - e^{-\frac{|\vec{x}_i - \vec{x}_i'|^2}{2B_i^3}} \right) \quad (10)$$

where k_{gmm} is a user-defined constant (we used 10^5 for all refinements in this manuscript), \vec{x}_i is an atom's current position, \vec{x}_i' is the updated position suggested by the GMM and B_i is the atomic B-factor.

Creating composite maps

Given an aligned set of experimental maps with fitted models, we use the mixture modelling formulation we provide to generate a composite map. The responsibilities attributed to each chain of a model can be used to weight their intensities when they are combined into the composite map. Adding the signal from all these maps together typically leads to artefacts at the seams (Fig. 4, Supplementary Fig. 8). To deal with this, the experimental maps are reweighted by the responsibility of the components (rather than the atoms) as per Eq. 4 and then summed together (Supplementary Fig. 12).

The input for the algorithm is a consensus model and multiple pre-aligned composite maps. Given C components each with a corresponding atomic model and an experimental map $M_{e,c}$, we create a simulated map M_c for the component. Here, we use the equation for simulating a map (Eq. 3), but only consider the contributions of the atoms of component C :

$$M_c(\vec{v}) = \sum_{v \in V} (\vec{v}) + E \quad (11)$$

Similarly, the responsibility for a component is determined by normalising it against the simulated map of all components. We retain only the high-resolution regions of these component maps by setting the atomic number to 0 when computing the simulated map for atoms in a given model, provided that the corresponding atom in another component map has a lower B-factor. The responsibility map for a given component, γ_c , is computed as follows:

$$\gamma_c = \frac{M_c(\vec{v})}{\sum_c M_c(\vec{v})} \quad (12)$$

Now, the final composite map, M_C , is defined as the sum of all the responsibility-weighted experimental maps.

$$M_C(\vec{v}) = \sum_c \gamma_c(\vec{v}) M_{e,c}(\vec{v}) \quad (13)$$

Conformation-based force calculation and MD

OpenMM is used for the conformation-based force calculation and MD³³. We tested CHARMM36 and AMBER14 in OpenMM (Supplementary Table 3), and they show slight differences in the preferred backbone dihedrals (Supplementary Fig. 13). Although other force fields were available, we used AMBER14 for our runs. We used a GB-Neck2 implicit solvent model⁶¹ and Langevin integrator with a 0.1 femtosecond timestep to calculate atomic trajectories.

Running the refinements

Before any positional refinement of a given model, the B-factors for all atoms were refined for 25 iterations. B-factors were capped to a maximum value of 1.5 for membrane proteins and 2.5 for all other models. At each refinement iteration, the simulation was run for 2000-time steps. The CCC was calculated for the updated model, using a global B-factor (set to be equivalent to the global resolution of the cryo-EM map) for map simulation (Eq. 3), and if the CCC did not improve for 5 iterations the refinement was stopped. If this convergence criterion was not met after 300 iterations, the refinement was stopped.

Local quality of fit (LoQFit)

We implemented a local-fit quality score as part of the TEMPY2 python package. The score – LoQFit – uses an approach similar to a local FSC score for cryo-EM maps⁶² in order to assess the fit quality of a protein model. This local FSC score is calculated for regions defined by a soft-edged spherical mask, centred at the C_α atom for each residue in the fitted model and applied to both M_s and M_e . The diameter of this mask is five times the global resolution of the experimental map. We use an

FSC threshold of 0.5 to determine the LoQFit score for each residue. To improve the smoothness of the final LoQFit plot, we include an option to estimate the exact frequency at 0.5 correlation between the two maps, using linear interpolation.

We also use SMOcf to estimate the local quality of fit³⁵. Briefly, SMOcf uses a local window around each residue, and then computes the Manders overlap coefficient between the simulated observed maps in this region.

Ensemble algorithm

To compute an ensemble of atomic models that fit the cryo-EM map, we create an ensemble of locally perturbed conformations. This is achieved by sampling the coordinates of each atom from a multivariate Gaussian. The mean value of this Gaussian is set to initial position of each atom, and the covariance matrix is constructed from the shifted B-factors (which are the original B-factors adjusted such that the minimum B-factor is fixed at 0.25). We then locally minimise each model in the ensemble, to keep acceptable stereochemistry.

Following this, we apply an ensemble fitting force and a density-guided force. The ensemble energy term E_{ens} is defined per atom as:

$$E_{ens} = \frac{k_{ens}}{\sqrt{2\pi^3 * B_i^3}} * \left(1 - e^{-\frac{|\vec{x}_i - \vec{x}'_i|^2}{B_i^3}} \right) \quad (14)$$

where E_{ens} is a constant (1000 is used for all examples shown in this manuscript), B_i is the atomic B-factor, \vec{x}'_i are the coordinates of the atom after resampling, and \vec{x}_i are the coordinates prior to sampling. The energy for the density-guided force is defined as the negative (interpolated) cryo-EM density value at the position of each atom, scaled by a constant k_{dens} , which typically needs to be optimised for each map (values used range between 5 and 200). With these forces applied, we run a short simulation (2000 steps of 0.1 femtoseconds) and minimise using L-BGFS in openMM³³.

We then generate blurred maps for each conformation in the ensemble, and compute a voxel-based average. To determine the number of models in an ensemble we increase the number of models until there is no increase in CCC. This average blurred map represents the final ensemble average map we use throughout the text.

RMSF

To compute the RMSF value for our generated ensemble, we first compute the mean structure, and then compute the RMSF using the normal formula. For an ensemble of structures, the residue fluctuation profiles for an ensemble with N models are calculated according to the formula:

$$RMSF = \sqrt{\frac{1}{N} \sum_j^N (x_{i(j)} - \langle x_i \rangle)^2} \quad (15)$$

where $x_{i(j)}$ denotes the position (coordinates) of the i -th C α atom in the structure of the j -th ensemble model and $\langle x_i \rangle$ denotes the averaged position of the i -th C α atom in all models in the ensemble.

Local resolution calculations

We used the ResMap method to compute local resolution estimates⁶³. ResMap uses local windows of varying size, and statistical tests to determine the most likely resolution for each voxel in the map.

Generation of benchmark and assessment

Our benchmark is based on the CERES database³⁸. We took the corresponding deposited maps and structures from EMDB⁶⁴ and PDB⁶⁵, and the re-refined structures from CERES. Because of the CERES database setup, our benchmark contains maps resolved from 2.1–4.9 Å

resolution. We did not include any CERES models that contained stretches of 3 or more consecutive residues with no modelled side chain atoms.

In almost all cases, we assess the goodness-of-fit of models using the CCC with ChimeraX 1.3, using the command *measure correlation*⁶⁶. The exception to this is the results presented in Fig. 3a, and in Fig. S4, in which the CCC was calculated using TEMPy⁶⁷. Simulated maps were generated using TEMPy with a uniform B-factor set to be equivalent to the global resolution value for the cryo-EM map, which was obtained from the EMDB. MolProbity and clash scores were calculated using *phenix.molprobity*⁶⁸, and CaBLAM using *phenix.cablam*³⁷.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data that support this study are available from the corresponding authors upon request. We obtained atomic models for refinement from the PDB and CERES, and the corresponding cryo-EM maps from the EMDB. All TEMPy-ReFF refined models described in this paper, alongside the corresponding models from the PDB and CERES, where appropriate, are deposited at the following Zenodo repository: [<https://doi.org/10.5281/zenodo.8395613>]. The AlphaFold2-Multimer predicted model shown in Fig. 7 is also deposited in the same Zenodo repository. The numerical data underlying the plots shown in Figs. 2a, 3a–c, 5d, 6c, 7b are provided as a Source Data file.

Code availability

TEMPy-ReFF is available at <https://www.topf-group.com/tempy-reff>.

References

- van Zundert, GydoC. P. Bijvoet Center for Biomolecular Research, Faculty of Science-Chemistry, Utrecht University, Utrecht, the Netherlands & Bonvin, AlexandreM. J. J. Fast and sensitive rigid-body fitting into cryo-EM density maps with PowerFit. *AIMS Biophys.* **2**, 73–87 (2015).
- Nicholls, R. A., Tykac, M., Kovalevskiy, O. & Murshudov, G. N. Current approaches for the fitting and refinement of atomic models into cryo-EM maps using CCP-EM. *Acta Crystallogr. D Struct. Biol.* **74**, 492–505 (2018).
- Ahmed, A., Whitford, P. C., Sanbonmatsu, K. Y. & Tama, F. Consensus among flexible fitting approaches improves the interpretation of cryo-EM data. *J. Struct. Biol.* **177**, 561–570 (2012).
- Singharoy A. et al. Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *Elife.* **5**, e16105 (2016).
- Chen, J. Z., Fürst, J., Chapman, M. S. & Grigorieff, N. Low-resolution structure refinement in electron microscopy. *J. Struct. Biol.* **144**, 144–151 (2003).
- Topf, M. et al. Protein structure fitting and refinement guided by cryo-EM density. *Structure* **16**, 295–307 (2008).
- Kawabata, T. Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model. *Biophys. J.* **95**, 4643–4658 (2008).
- Kawabata, T. Gaussian-input Gaussian mixture model for representing density maps and atomic models. *J. Struct. Biol.* **203**, 1–16 (2018).
- Igaev, M., Kutzner, C., Bock, L. V., Vaiana, A. C. & Grubmüller, H. Automated cryo-EM structure refinement using correlation-driven molecular dynamics. *Elife* **8**, <https://doi.org/10.7554/eLife.43542> (2019).
- Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D Struct. Biol.* **74**, 531–544 (2018).

11. Lopéz-Blanco, J. R. & Chacón, P. iMODFIT: efficient and robust flexible fitting based on vibrational analysis in internal coordinates. *J. Struct. Biol.* **184**, 261–270 (2013).
12. Tama, F., Miyashita, O. & Brooks, C. L. 3rd Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.* **147**, 315–326 (2004).
13. Wang R. Y. R. et al. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. *Elife*. **5**, e17219 (2016).
14. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
15. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. D Struct. Biol.* **74**, 519–530 (2018).
16. Best, R. B. et al. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).
17. Maier, J. A. et al. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
18. Klaholz, B. P. Deriving and refining atomic models in crystallography and cryo-EM: the latest Phenix tools to facilitate structure analysis. *Acta Crystallogr. D Biol. Crystallogr.* **75**, 878–881 (2019).
19. Nakane, T., Kimanius, D., Lindahl, E. & Scheres, S. H. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *Elife* **7**, <https://doi.org/10.7554/eLife.36861> (2018).
20. Farrell, D. P. et al. Deep learning enables the atomic structure determination of the Fanconi Anemia core complex from cryoEM. *IUCr J* **7**, 881–892 (2020).
21. Lukoyanova, N. et al. Conformational changes during pore formation by the perforin-related protein pleurotolysin. *PLoS Biol.* **13**, e1002049 (2015).
22. Farabella, I. et al. TEMPy: a Python library for assessment of three-dimensional electron microscopy density fits. *J. Appl. Crystallogr.* **48**, 1314–1323 (2015).
23. Sachse, C. et al. High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus. *J. Mol. Biol.* **371**, 812–835 (2007).
24. Mendez, J. H. & Stagg, S. M. Assessing the quality of single particle reconstructions by atomic model building. *J. Struct. Biol.* **204**, 276–282 (2018).
25. Herzik, M. A. Jr, Fraser, J. S. & Lander, G. C. A multi-model approach to assessing local and global Cryo-EM map quality. *Structure* **27**, 344–358.e3 (2019).
26. Pintilie, G., Chen, D. H., Haase-Pettingell, C. A., King, J. A. & Chiu, W. Resolution and probabilistic models of components in CryoEM maps of mature P22 bacteriophage. *Biophys. J.* **110**, 827–839 (2016).
27. Nierzwicki, Ł. & Palermo, G. Molecular dynamics to predict Cryo-EM: capturing transitions and short-lived conformational states of biomolecules. *Front. Mol. Biosci.* **8**, 641208 (2021).
28. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
29. Lawson, C. L. et al. EMDatabank unified data resource for 3DEM. *Nucleic Acids Res.* **44**, D396–D403 (2016).
30. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D Biol. Crystallogr.* **75**, 861–877 (2019).
31. Bonomi, M. et al. Bayesian weighing of electron cryo-microscopy data for integrative structural modeling. *Structure* **27**, 175–188.e6 (2019).
32. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program* **45**, 503–528 (1989).
33. Eastman, P. et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
34. Rieping, W., Habeck, M. & Nilges, M. Inferential structure determination. *Science* **309**, 303–306 (2005).
35. Joseph, A. P. et al. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. *Methods* **100**, 42–49 (2016).
36. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
37. Prisant, M. G., Williams, C. J., Chen, V. B., Richardson, J. S. & Richardson, D. C. New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink “waters,” and NGL Viewer to recapture online 3D graphics. *Protein Sci.* **29**, 315–329 (2020).
38. Liebschner, D. et al. CERES: a cryo-EM re-refinement system for continuous improvement of deposited models. *Acta Crystallogr. D Biol. Crystallogr.* **77**, 48–61 (2021).
39. Ellinghaus, T. L., Marcellino, T., Srinivasan, V., Lill, R. & Kühlbrandt, W. Conformational changes in the yeast mitochondrial ABC transporter Atm1 during the transport cycle. *Sci. Adv.* **7**, eabk2392 (2021).
40. McGreevy, R., Teo, I., Singharoy, A. & Schulten, K. Advances in the molecular dynamics flexible fitting method for cryo-EM modeling. *Methods* **100**, 50–60 (2016).
41. Malhotra S. et al. RIBFIND2: identifying rigid bodies in protein and nucleic acid structures. *Nucleic Acids Res.* **51**, gkad721 (2023).
42. Pandurangan, A. P. & Topf, M. Finding rigid bodies in protein structures: application to flexible fitting into cryoEM maps. *J. Struct. Biol.* **177**, 520–531 (2012).
43. Mulvaney, T. et al. CASP15 cryo-EM protein and RNA targets: refinement and analysis using experimental maps. *Proteins* **91**, 1935–1951 (2023).
44. Sedzicki, J. et al. Mechanism of cyclic β -glucan export by ABC transporter Cgt of Brucella. *Nat. Struct. Mol. Biol.* **29**, 1170–1177 (2022).
45. Zhao, Z. et al. Near-atomic architecture of Singapore grouper iridovirus and implications for giant virus assembly. *Nat. Commun.* **14**, 2050 (2023).
46. Pettersen, E. F. et al. UCSF chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
47. Chen, Y. et al. Allosteric transcription stimulation by RNA polymerase II super elongation complex. *Mol. Cell* **81**, 3386–3399.e10 (2021).
48. Hoffmann, N. A. et al. Molecular structures of unbound and transcribing RNA polymerase III. *Nature* **528**, 231–236 (2015).
49. Lawson, C. L. et al. Cryo-EM model validation recommendations based on outcomes of the 2019 EMDatabank challenge. *Nat. Methods* **18**, 156–164 (2021).
50. Farnung, L., Ochmann, M. & Cramer, P. Nucleosome-CHD4 chromatin remodeler structure maps human disease mutations. *Elife* **9**, <https://doi.org/10.7554/eLife.56178> (2020).
51. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
52. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
53. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moutl, J. Critical assessment of methods of protein structure prediction (CASP)-round XIV. *Proteins* **89**, 1607–1617 (2021).
54. Evans R. et al. Protein complex prediction with AlphaFold-multimer. *bioRxiv*. <https://doi.org/10.1101/2021.10.04.463034>.

55. Gao, Y. et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **368**, 779–782 (2020).
56. Heymann, J. B. Single-particle reconstruction statistics: a diagnostic tool in solving biomolecular structures by cryo-EM. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **75**, 33–44 (2019).
57. Chen, Z. & Chapman, M. S. Conformational disorder of proteins assessed by real-space molecular dynamics refinement. *Biophys. J.* **80**, 1466–1472 (2001).
58. Robin, X. et al. Continuous Automated Model Evaluation (CAMEO)-perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins* **89**, 1977–1986 (2021).
59. Bishop C. M. Pattern Recognition and Machine Learning. Springer New York. Accessed September 26, 2023. <https://link.springer.com/book/9780387310732>.
60. Murshudov, G. N. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).
61. Nguyen, H., Roe, D. R. & Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **9**, 2020–2034 (2013).
62. Cardone, G., Heymann, J. B. & Steven, A. C. One number does not fit all: Mapping local variations in resolution in cryo-EM reconstructions. *J. Struct. Biol.* **184**, 226–236 (2013).
63. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
64. Lawson, C. L. et al. Emdatabank.org: unified data resource for CryoEM. *Nucleic Acids Res.* **39**, D456–D464 (2011).
65. Burley, S. K. et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **49**, D437–D451 (2021).
66. Pettersen, E. F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70–82 (2021).
67. Cragolini, T. et al. TEMPY2: a Python library with improved 3D electron microscopy density-fitting and validation workflows. *Acta Crystallogr. D Biol. Crystallogr.* **77**, 41–47 (2021).
68. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
- Multimer. We thank Dr. Aaron Sweeney, Dr. Sony Malhotra, and Dr. Agnel Praveen Joseph for their helpful discussions.

Author contributions

J.G.B., T.M., T.C., and M.T. conceived the study. J.G.B., T.M., and M.T. designed the experiments. J.G.B., T.M., and T.C. developed the method. J.G.B. performed the benchmarking experiments. J.G.B., T.M., and M.T. analyzed the results. All authors contributed to writing the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at

<https://doi.org/10.1038/s41467-023-44593-1>.

Correspondence and requests for materials should be addressed to Maya Topf.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, 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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

Acknowledgements

This work was supported by the Leibniz Institute of Virology as part of Leibniz ScienceCampus InterACT (funded by the BWFGB Hamburg and the Leibniz Association) and a Wellcome Collaborative Award in Science (209250/Z/17/Z). We thank Dr. Sanjana Nair for her help with AlphaFold-

3.2 Publication in Nucleic Acid Research

In the previous paper we refined models which had been deposited in the PDB and were thus already reasonably well fit to the experimental cryo-EM maps. Such refinement involves only small changes in atomic positions. In practical model building scenarios, well fitting models are not the norm. Instead, models derived from other experiments or computational approaches such as AlphaFold are used as starting points. Often starting from such models poses a challenge as the flexible fitting method on its own is prone to becoming trapped in local minima or distorting the geometry (section 1.2.5).

Rigid bodies generated by the RIBFIND method (Pandurangan et al. 2012a) have been applied to the Flex-EM tool (Pandurangan et al. 2012b) to aid convergence and preserve desirable geometry. However, the approach has some limitations. RIBFIND only handles structures of proteins and the underlying implementation is extremely naive, making finding rigid bodies in the large complexes being solved today a timely endeavour. In the RIBFIND2 paper which follows, I share first authorship with Dr. Sony Malhotra. In this new implementation, I used an Iterative Strongly Connected Components style algorithm allowing timely execution of method. I developed a ChimeraX plugin and a web server to improve accessibility. Finally, I implemented an automated hierarchical refinement routine in TEMPY-ReFF which takes advantage of the rigid bodies defined by RIBFIND2 to improve the radius of convergence. In theory, Flex-EM users will now also be able to better fit divergent RNA conformations into cryo-EM maps.

RIBFIND2: Identifying rigid bodies in protein and nucleic acid structures

Sony Malhotra^{1,†}, Thomas Mulvaney^{2,3,4,†}, Tristan Cragolini^{2,5}, Haneesh Sidhu⁵, Agnel P. Joseph¹, Joseph G. Beton^{2,3} and Maya Topf^{2,3,4,*}

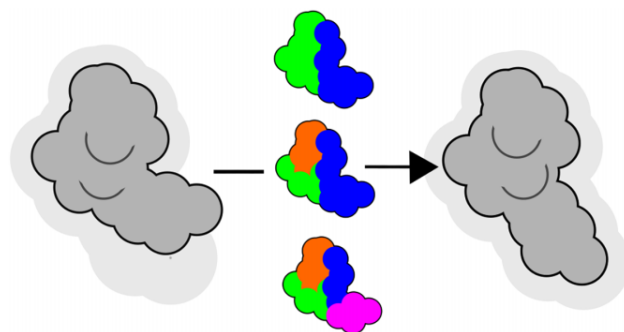
¹Science and Technology Facilities Council, Scientific Computing, Research Complex at Harwell, Didcot OX11 0FA, UK, ²Leibniz Institute of Virology, Hamburg 20251, Germany, ³Centre for Structural Systems Biology, Hamburg D-22607, Germany, ⁴Universitätsklinikum Hamburg Eppendorf (UKE), Hamburg 20246, Germany and ⁵Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, University of London, London WC1E 7HX, UK

Received December 06, 2022; Revised August 10, 2023; Editorial Decision August 11, 2023; Accepted August 21, 2023

ABSTRACT

Molecular structures are often fitted into cryo-EM maps by flexible fitting. When this requires large conformational changes, identifying rigid bodies can help optimize the model-map fit. Tools for identifying rigid bodies in protein structures exist, however an equivalent for nucleic acid structures is lacking. With the increase in cryo-EM maps containing RNA and progress in RNA structure prediction, there is a need for such tools. We previously developed RIBFIND, a program for clustering protein secondary structures into rigid bodies. In RIBFIND2, this approach is extended to nucleic acid structures. RIBFIND2 can identify biologically relevant rigid bodies in important groups of complex RNA structures, capturing a wide range of dynamics, including large rigid-body movements. The usefulness of RIBFIND2-assigned rigid bodies in cryo-EM model refinement was demonstrated on three examples, with two conformations each: Group II Intron complexed IEP, Internal Ribosome Entry Site and the Processome, using cryo-EM maps at 2.7–5 Å resolution. A hierarchical refinement approach, performed on progressively smaller sets of RIBFIND2 rigid bodies, was clearly shown to have an advantage over classical all-atom refinement. RIBFIND2 is available via a web server with structure visualization and as a standalone tool.

GRAPHICAL ABSTRACT



INTRODUCTION

Cryo-electron microscopy (cryo-EM) is the method of choice for elucidating structures of large macromolecular assemblies at high (better than ~4 Å) to medium resolutions (~4–10 Å). Already ~20% of cryo-EM structures in the Electron Microscopy Data Bank (EMDB) (1) contain RNA components. A large portion of the genome encodes for non-coding RNA (ncRNA) (2) and the Nucleic Acid Knowledge Base (NAKB) (3), the successor to the Nucleic Acid Database (NDB) (4,5), currently holds 16473 structures (as of August 2023). In the last year, 56% of new entries were derived from cryo-EM experiments. In total, 22% of all structures in the NAKB are from cryo-EM techniques at various resolutions, some of which prohibit clear determination of the atomic positions. These could be combined with RNA structure prediction and refinement algorithms, which are continuously improving (6,7). These changes in the field could lead to more insights into biological processes and experiments, such as CAS9-CRISPR gRNA generation and ribonucleoprotein assemblies.

*To whom correspondence should be addressed. Tel: +49 40 8998 87660; Email: maya.topf@cssb-hamburg.de

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

To derive an atomic model of an assembly, usually, atomic structures of assembly components are fitted into the cryo-EM map and then further refined within the map. Especially, but not exclusively, at medium resolutions, the latter process (also called ‘flexible fitting’) can be assisted and sped up by using rigid bodies (RBs) linked by flexible linkers in the fitted components. It can also improve the accuracy of the final refined model (8). At present, methods to identify RBs are mostly designed for protein structures (8).

The RIBFIND algorithm (9) was originally designed to detect RBs in protein structures via the clustering of secondary structural elements (SSEs), primarily to aid the fitting of structures into cryo-EM maps. RIBFIND was made available both as a web server and a standalone program (6). Here, we have developed a new algorithm, RIBFIND2, which identifies RBs in ncRNA structures by clustering SSEs assigned using the RNAView program (10). We have also optimized the original RIBFIND algorithm parameters for clustering protein structures. The algorithm was tested on structures containing proteins and RNA in different conformations.

We have implemented RIBFIND2 in a web server (Figure 1) with no login requirements—<https://ribfind.topf-group.com/>, which also supports a molecular JavaScript viewer—NGL viewer (11)—as it is more interactive, faster and scalable than our previous Java-based viewer (Jmol, <http://jmol.sourceforge.net/>). We also provide the software as a standalone package which can be downloaded from a link provided in the web server.

MATERIALS AND METHODS

Secondary structure determination for RNA

The secondary structure of RNA molecules was determined with the RNAView (10) program, which is also used by the NDB to assign secondary structures to nucleic acids. RNAView calculates base-pairing interactions in a molecule based on distance and angle restraints. From those base pairs, it divides the molecule into double-stranded helical segments and single-stranded loop segments. Although the secondary structure classification of RNA is significantly more complex than these two categories, for the purposes of clustering this binary division contains enough information. Because single-stranded segments are an important part of RNA tertiary structure and are involved in intramolecular interactions, they were treated as secondary structure elements (SSEs) in their own right, rather than merely as connecting elements (as loops generally are in proteins). The RNAView secondary structure predictions (which are in XML format) were used for further calculations.

Clustering protein and RNA structures

The clustering algorithm is partially based on the original RIBFIND algorithm (neighborhood-based clustering) developed for defining RBs in proteins (8,9). The algorithm groups SSEs together into RBs based on the ‘strength’ of their interaction. For proteins, ‘cutoff distance’ (previously called ‘contact distance’) is defined as the distance between the average atomic position of side-chain atoms, except for

glycine where the C α is used. For RNA it is the average atomic position of nucleotide atoms excluding the phosphate groups. The strength of the interaction between an SSE (A) and a partner SSE (B) is defined in terms of the fraction of ‘allowed’ residues (see below) in A which are within the cutoff distance of the allowed residue in B. For proteins and RNA, the default cutoff distance is 6.5 Å (12) and 7.5 Å (13,14), respectively. These cutoff values can be changed to user-defined values. For RNA, this default was selected based on the analysis of base-base interactions in ellipsoidal shells (13).

The interaction strength is defined in terms of the fraction of residues within the cutoff distance of one another, where $|X|$ denotes the number of elements in the set X :

$$\text{frac}(A, B) = \frac{|\text{cutoff}(\text{allowed}(A), \text{allowed}(B))|}{|\text{allowed}(A)|} \quad (1)$$

Because $\text{frac}(A, B)$ does not necessarily equal $\text{frac}(B, A)$, the interaction for the pair is instead defined as the maximum of the two:

$$\text{interaction}(A, B) = \max(\text{frac}(A, B), \text{frac}(B, A)) \quad (2)$$

The ‘allowed’ residues of an SSE enable finer control of interaction calculations. These are computed for each type of SSE. For β -sheets, only strands longer than three residues are allowed in interaction calculations (9). For unpaired RNA strands, a similar rule is applied. For α -helix to α -helix interactions, the ratio of the helix lengths in residues must be >0.4 (9).

Given the interaction function, a graph is constructed where nodes are SSEs and edges are the computed interactions. By choosing an *interaction threshold* (originally termed ‘cluster cutoff’) and removing edges from the graph that fall below this, the set of RBs (strongly connected components) changes. The algorithm, thus, produces unique sets of RBs and their respective interaction thresholds by iteratively removing edges in order of strength.

A ‘unique’ cluster number (UCN) for a given *interaction threshold* is defined as:

$$\text{UCN} = \frac{|\text{SSEs} \in \text{RBs}|}{|\text{SSEs}|} + |\text{RBs}| \quad (3)$$

where $|\text{SSE} \in \text{RigidBodies}|$ denotes the number of SSEs which are within RBs in the cluster of interest. We have previously demonstrated in detail the usefulness of the highest UCN in the refinement of three protein cases (9), where flexible fitting using clustered RBs resulted in a model that better fit the experimental map. The highest UCN has previously been chosen for refinement as it tends to have most of SSEs clustered into a large number of RBs. However, the highest UCN cluster may not always be the best for this purpose. We therefore compare it against a more costly ‘hierarchical’ approach in this paper.

Benchmark dataset for protein-nucleic acid complexes

The NDB (5) was searched for RNA structures with tertiary interactions to test the algorithm. A series of group IIC intron structures in different states of catalysis was first used to test the algorithm (PDB IDs: 3eog, 3eoh, 3bwp,

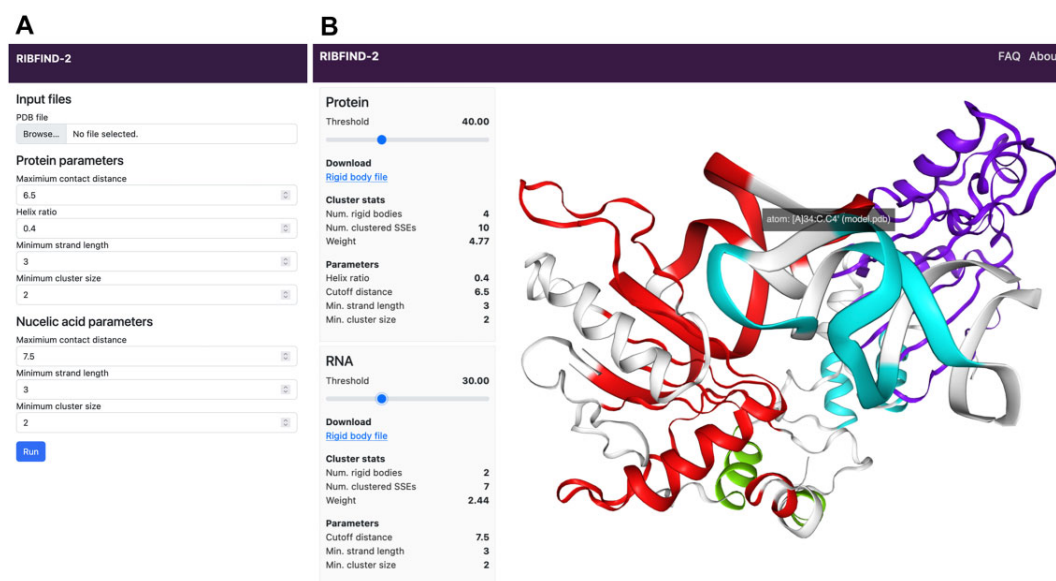


Figure 1. Snapshot of the RIBFIND2 server. (A) The input parameters required to submit the job to RIBFIND2 web server. (B) The results page with colored rigid bodies for both protein and RNA components in an input PDB ID (1mji). The user can turn on/off the protein, RNA or the cryo-EM map. The slider lets the user control the *interaction threshold* for both protein and RNA components.

Table 1. Dataset used to assess the performance of RIBFIND2 rigid bodies during TEMPy-REFF refinement

Type	Model	Map	Res. [†] (Å)	Description
Processome	7MQ9	23937	3.9	Cryo-EM structure of the human SSU processome, state pre-A1*
	7MQA	23938	2.7	Cryo-EM structure of the human SSU processome, state post-A1
Group II Intron complexed with intron-encoded protein (IEP)	7D0F	30532	5.0	Cryo-EM structure of a precatalytic group II intron RNP
	7D0G	30533	5.0	Cryo-EM structure of a precatalytic group II intron
IRES	7SYR	25538	3.6	Structure of the wt IRES eIF2-containing 48S initiation complex, closed conformation. Structure 12(wt)
	7SYQ	25537	3.8	Structure of the wt IRES and 40S ribosome ternary complex, open conformation. Structure 11(wt)

[†]Res. refers to the resolution of the cryo-EM map.

4ds6, 5j01, 5j02). Additionally, structures of the 80S ribosome in different states of Internal Ribosome Entry Site (IRES) translocation were then used, in which the small and large subunits were run separately and with protein chains removed (PDB IDs: 5juo, 5jus, 5jut, 5juu). The clustering of structures were viewed and analyzed using UCSF Chimera (15).

Application to cryo-EM refinement

We selected three cases of RNA structures, in two conformations each, to test the usefulness of RIBFIND2 in refining those structures in cryo-EM maps. These were: Group II Intron complexes with intron-encoded protein (IEP), Internal Ribosome Entry Site (IRES) and the Processome, with cryo-EM maps between 2.7 and 5 Å resolution.

We refined each atomic model into the cryo-EM map corresponding to the other conformation (Table 1). For the Group II intron models, both the RNA and protein chains (chains A and C respectively) were refined. Due to the large

size of the processome and IRES models, we refined only two of the major RNA chains from these models, which corresponded to chains ‘L1’ and ‘L2’ and chains ‘2’ and ‘z’, respectively. We compared two approaches of applying these restraints, the first based on the decomposition of RBs (‘hierarchical’) and the second based on choosing a single cluster with the highest UCN. In the hierarchical approach, RIBFIND2 clusters are selected in order of increasing interaction threshold, which leads to progressively smaller clusters and thus more and more flexibility. As a control, we performed an unrestrained refinement.

The refinement protocol included three steps (Supplementary Figure S1): (i) the model was first aligned to the target to produce a rough fit then locally optimized using the ‘fitmap’ tool in ChimeraX to produce the initial starting model; (ii) TEMPy-REFF (16) density-guided fitting was used in conjunction with progressively smaller RIBFIND2 RBs (hierarchical), the highest UCN set of RBs (UCN), or all-atom (unrestrained) and (iii) TEMPy-REFF all-atom Gaussian-mixture model refinement.

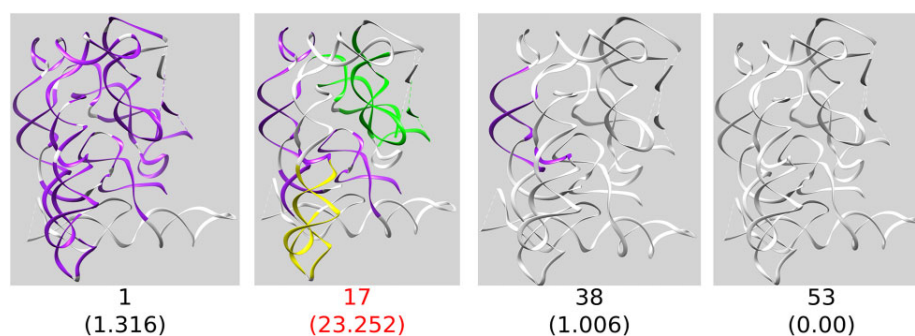


Figure 2. Clustering for group IIC intron (PDB ID: 3bwp) using RIBFIND2, labeled with *interaction threshold* and the UCN (below in parentheses). The *interaction threshold* for the highest UCN is highlighted in red.

Each set of RBs were refined using TEMPy-REFF until convergence, i.e. the variance in CCC score of the last five runs was $<10^{-9}$. The TEMPy-REFF density-guided force field was set to a strength of 20 in all experiments. The TEMPy-REFF GMM strength for the refinement step was 10^3 .

RESULTS

Clustering RNA in group II introns

The algorithm was first tested on Group IIC introns. The SSE clustering using a 1% threshold to no clusters (at threshold $\geq 17\%$) is shown in snapshots for an intron structure in Figure 2 (PDB ID: 3bwp). The intron starts as one cluster (comprising all secondary structures) initially and then breaks off into smaller clusters, with the number of clusters peaking at the highest UCN. The algorithm was also run on several other group IIC intron structures in various states of catalysis; The *interaction threshold* and highest UCN for each of these states is shown in Figure 2. Structures with PDB IDs: 3bwp, 3eog, 3eoh and 4ds6 are truncated (lacking domain 6) and consequently linear introns, as the branch point adenosine resides in domain 6 (17) while structures 5j01 and 5j02 are branched chimeric introns and are derived from *Oceanobacillus iheyensis*. A similar clustering pattern is observed across the different catalytic states with three key clusters emerging highlighted in purple, yellow and green (and for those intron structures without broken chains additionally a cluster in blue). Exceptions to this are the chimeras (5j01 and 5j02), which were created by replacing part of the *O. iheyensis* sequence with intron AVI.2 (17). In those cases, the yellow cluster does not break off and instead appears as part of larger purple clusters (Figure 3). 4ds6 also shows a red cluster at the top, which for all other states is non-clustered. This may reflect state-specific reduced flexibility in the pre-catalytic structure as well as for the chimeras.

Clustering RNA in the 80S eukaryotic ribosome

To test the algorithm on higher complexity RNA structures, a set of structures of the 80S ribosome bound to the Taura syndrome virus IRES were used (18). IRESs are RNA structures that carry out cap-independent translation of viral

mRNA via interacting with the 40S subunit (18). The ensemble of structures illustrates translocation and rearrangements of the IRES, coupled with 40S intra and inter-subunit rearrangements and therefore represents a good example of biologically relevant rigid body RNA movements.

The small subunit has been well characterized in terms of its dynamics and domains in many ribosome structures. The canonical small subunit is composed of head, beak, body and platform domains (Figure 4), based on transitions between different states during translocation (19,20). Snapshots of the trajectory of clustering from 1% threshold to no clusters (at interaction threshold $\geq 61\%$) for a single 40S conformation (PDB ID: 5juo) are shown in Figure 4. As observed with intron structures, larger clusters are present at lower thresholds, which eventually separate into smaller domains, but still include most of the SSEs, which then gradually localize to subsections or peripheries excluding most SSEs as the *interaction threshold* is increased. At thresholds of 15–25% there is a good separation of head, beak, platform, body and IRES domains. Moreover, the IRES initially starts as one cluster that breaks into two clusters approximately corresponding to its two known domains (the 5' region and PKI region) (18).

We further assessed the algorithm to generate functionally meaningful clusters using other conformations in addition to 5juo (PDB IDs: 5jut, 5juu, 5jup, 5jus). RIBFIND2 assignment with the highest UCN resulted in a similar clustering pattern into the classical domains, as well as of the IRES (which also adopts a different conformation in eac16h structure) (Supplementary Figure S2). As well as the canonical 40S domains, a set of 3–4 clusters (coloured orange, red, yellow and cyan in Figure 4) are consistently found in the lower half of the 40S subunit. The clustering of these RBs changes the least for the different states. Comparing the proportion of SSEs in clusters vs. the *interaction threshold* shows a drop around the threshold that corresponds to the highest UCN in all conformations (Supplementary Figure S3).

During the transition between the different conformations the head domain rotates by $\sim 40^\circ$ (19,20), a phenomenon also reported for bacterial and mammalian systems (21–23). Thus, the similar clustering pattern observed along the trajectory of different states suggests the clusters identified by the algorithm represent biologically relevant RBs.

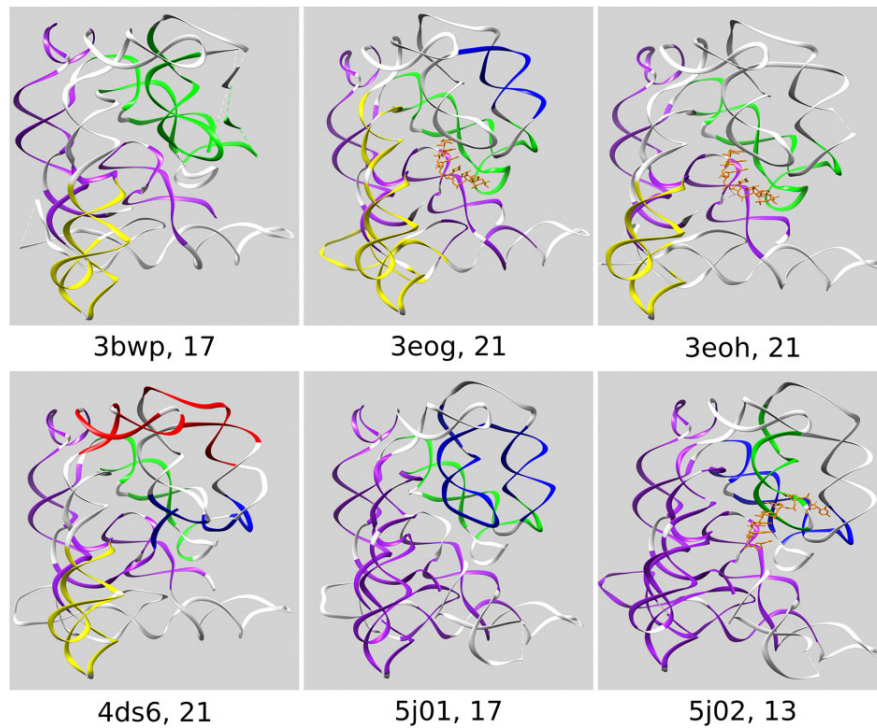


Figure 3. Clustering for group IIC introns in different catalytic states and conformations. For each structure, the clustering based on the highest UCN is indicated next to PDB ID.

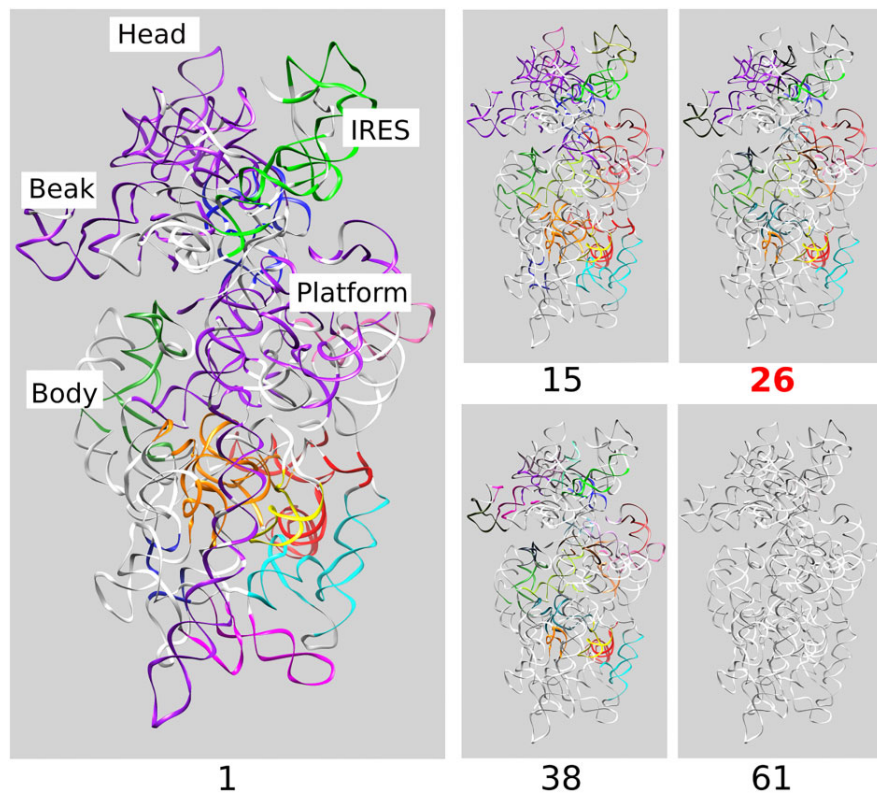


Figure 4. Clustering for small ribosomal subunit (PDB ID: 5juo). Canonical domains are marked in the first panel and clustering patterns are shown in ascending order labelled with corresponding interaction threshold. The interaction threshold of the clustering with the highest UCN is highlighted in red.

Table 2. Assessment of refined models using different refinement approaches. Best CCC and RMSD are highlighted in bold

Model	Target	Method	CCC	RMSD (Å)
7mq9	7mqa	Target	0.64	0.0
		Initial	0.49	15.1
		Hierarchical	0.62	7.3
		UCN (25.53)	0.6	8.5
		Unrestrained	0.57	12.6
7mqa	7mq9	Target	0.64	0.0
		Initial	0.47	14.9
		Hierarchical	0.56	7.6
		UCN (33.62)	0.53	11.24
		Unrestrained	0.54	12.3
7d0f	7d0g	Target	0.85	0.0
		Initial	0.69	6.2
		Hierarchical	0.81	3.2
		UCN (14.66)	0.81	3.2
		Unrestrained	0.81	3.1
7d0g	7d0f	Target	0.85	0.0
		Initial	0.69	5.9
		Hierarchical	0.82	3.1
		UCN (15.84)	0.82	3.1
		Unrestrained	0.81	2.9
7syr	7syq	Target	0.75	0.0
		Initial	0.70	6.9
		Hierarchical	0.78	2.8
		UCN (47.77)	0.78	2.7
		Unrestrained	0.78	2.9
7syq	7syr	Target	0.73	0.0
		Initial	0.68	6.9
		Hierarchical	0.78	2.2
		UCN (43.77)	0.77	2.4
		Unrestrained	0.76	3.3

Clustering in the large subunit

Compared to the small subunit the large subunit is less dynamic during translocation and more compact with the RNA not classically seen as separated into distinct domains (24,25). The clustering of a single large subunit (PDB ID: 5juo) from 1% threshold to no clusters (at interaction threshold $\geq 61\%$) is shown in snapshots in Supplementary Figure S3. The central core of the 60S subunit largely stays as a large cluster (orange, Supplementary Figure S2) with peripheral SSEs gradually breaking off into different clusters. Overall, the core of 60S is conserved between the different states, particularly at the lower end of the *interaction threshold* range shown. The surfaces break into small clusters for all states representing flexibility compared to the core globular domain.

Using clusters of RIBFIND2 for cryo-EM structure refinement

The examples from Table 1 were used to perform refinement in two ways: hierarchical and UCN-based (see Materials and Methods). The hierarchical approach combines the advantages of using both the highest UCN and unrestrained approaches: large cluster sizes, used at the start of refinement, enable large conformational changes during fitting and hence prevents the model from getting stuck in small pockets of density, whilst small cluster sizes facilitate the

small adjustments required for accurate refinement once the model is placed in an approximately correct position.

In total, we performed refinements of six structures for the three cases (Materials and Methods and Table 1). We assessed the performance using a density-based metric, cross correlation (CCC), and a density-independent metric, root mean square deviation (RMSD). The latter was calculated over C4' RNA atoms (26,27) of the refined model from the target structure (Table 2). CCC scores for hierarchical refinements were generally higher or comparable to the UCN and unrestrained approaches. For the processome, we excluded residue ranges 1256–1516 and 1839–1860 from RMSD calculations. The former is in a low resolution part of the map, the latter is a small modelled fragment, which is disconnected from the rest of the model.

The combination of CCC and RMSD scores was best for the hierarchical approach, suggesting better fit was obtained whilst minimizing overfitting (lower RMSD values, Table 2). The geometry of the refined models were assessed using RNAValidate which is part of the PHENIX software (28) (Supplementary Table S1). There were no obvious differences between the restrained and unrestrained refinements. However, all refinements had a decrease in suite outliers and an increase in bond-angle outliers.

A comparison of the CCC score trajectories during refinement for the hierarchical, UCN and unrestrained approaches is presented in Figure 5A. A close-up of the IRES differences from the target model demonstrates the advantages of using RIBFIND2-defined RBs over the unrestrained refinement where no RBs are used (Figure 5B). Generally, both the hierarchical and UCN approaches enabled the flexible-fitting to converge on a conformation closer to the target structure (Figure 5, Tables 2 and S1). The local fit-to-map of the IRES model 7syq in map EMD-25538 was assessed using SMOC scores (Supplementary Figure S4A) which are part of the TEMPy (29). Compared to the hierarchical-based fitting (blue), UCN-based (orange) and unrestrained (green) refinements produced models with lower SMOC scores in the beak domain which is marked by a box. A close-up of the refined models in this region shows that the hierarchical model was closer to the target model (red) (Supplementary Figure S4B).

RIBFIND2 web server

To make the program user-friendly, RIBFIND2 has been implemented as a web server (<https://ribfind.topf-group.com/>). By default, the server accepts a single PDB file. However, the advanced form allows the previously described distance thresholds and interaction parameters to be adjusted from their defaults.

For proteins, the following parameters are user-definable:

1. The protein residue cutoff distance (default 6.5 Å).
2. The minimum ratio of lengths between helices for them to interact (default 0.4).

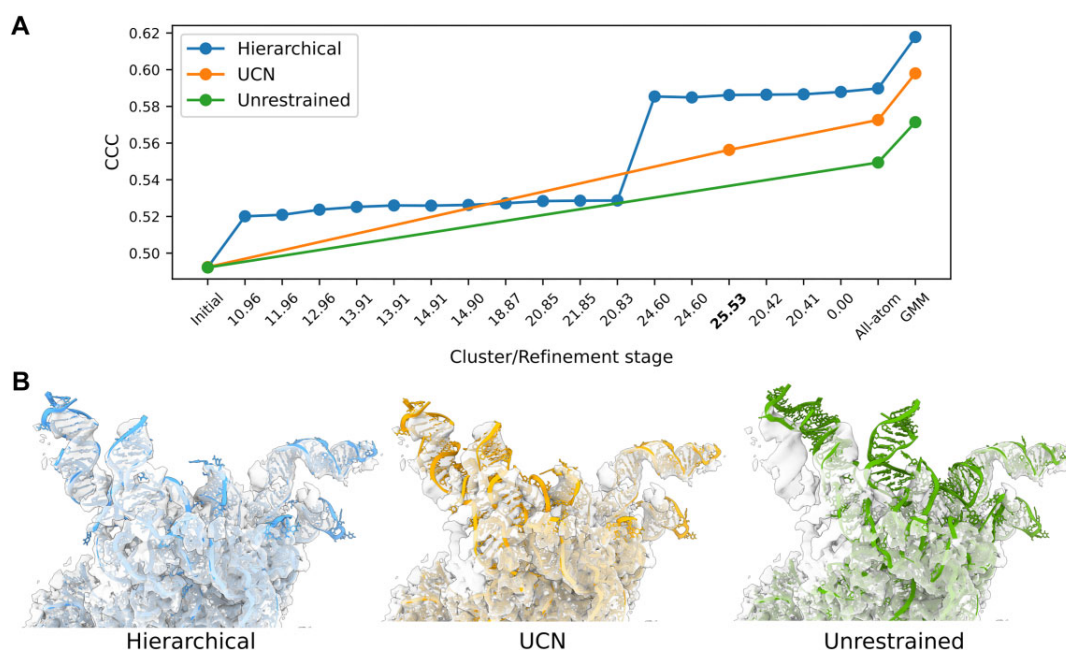


Figure 5. Refinement results for the processome model PDB ID: 7mq9 in cryo-EM map EMD-23938. **(A)** The CCC scores after each step for the hierarchical (blue), UCN (orange) and unrestrained (green) refinement procedures. Both restrained approaches yielded models with higher CCC scores. **(B)** Visualization of nucleotides 150–350 in the final model within the cryo-EM map (transparent grey). For the unrestrained model (green) and UCN model (orange) it is clear that some helices were unable to move to the correct density.

3. The minimum beta-strand length determines if a strand can be involved in interactions (default 4).
4. The cluster size, which determines the minimum number of clustered SSEs which can be called a RB.

For nucleic acids, the following parameters are user definable:

1. The nucleic acid cutoff distance (default 7.5 Å).
2. The minimum strand length of an RNA loop which can interact (default 4).
3. The minimum cluster size which is considered a RB (default 2).

The PDB file is read and validated internally before it is submitted for execution. After successful completion, the clusters are displayed on the web page with the results of the job. Failure to run the job caused by DSSP or RNAView processing their input are reported to the user so they may correct these issues.

Using the slider control on the result page, the user can view different sets of RB clusters generated for each *interaction threshold* and save the corresponding RB file in a text format (which can be used, e.g. by TEMPY-REFF or FlexEM (30)). The run-time for some examples is listed in Supplementary Table S2.

NGL and visualization. It is useful to provide an efficient visualization of biomolecular structures, and their separation in structural elements or a group of structural elements. To this end, we have implemented a NGLview JavaScript molecular viewer, where each RB in the user uploaded PDB file is colored uniquely (8). All SSEs and loops that do not

form part of any cluster are colored white. The clustering with the greatest UCN identified by the program is displayed by default. Directly embedded in the results page, it allows for quick and responsive visualization of very large structures (e.g. viral capsids). The viewer allows intuitive interaction using the mouse to quickly and easily change the display and consistent coloring of the structural blocks identified by RIBFIND2.

DISCUSSION

Rigid-body identification in biomolecular structures is a highly useful step to analyse structural models and to refine them against experimental data. Yet, few methods exist to do so in an automated fashion for nucleic acids. We have shown here that RIBFIND2 can be used to provide RBs that correspond to biologically relevant units in important RNA and protein/RNA structures, using group II introns and ribosome subunits as examples. We have also demonstrated that combining RBs identified by RIBFIND2 with a cryo-EM refinement method enhances the final quality of the model. This is particularly relevant for cryo-EM RNA structures, which are on average characterized by lower resolution compared to protein structures. Further, the optimal number of RBs is also dependent on the resolution of the map. We have previously shown using a simulated benchmark that the improvement in CCC drops as the resolution drops and at resolutions worse than 10 Å, it is hard to obtain an accurate refined model. At lower resolution, multiple structures tend to have similar fitting scores and hence it is more difficult to refine them. Previously, clus-

tering protein secondary structure elements into larger RBs was shown to improve the flexible fitting process (9).

Here we have demonstrated that RBs can also be used for the flexible fitting of RNA structures. To achieve this, we used a hierarchical approach (fitting each of the RIBFIND2 sets of RBs in order of 0–100%) which produced models which were closer to the target structure (lower RMSD) and were a better fit-to-map (higher CCC) compared to the model resulting from a standard unrestrained (all-atom) refinement or a refinement based on the highest UCN. Future work could include integrating the current method with deep-learning-based structure prediction methods due to its successful combination with cryo-EM model refinement. This could be done in an interactive manner, for example with molecular visualization and molecular dynamics tools, and could also aid in providing better model assessment and functional interpretation.

DATA AVAILABILITY

All data used in this study can be obtained from the PDB and EMDB repositories. The RIBFIND2 webserver is accessible at <http://ribfind.topf-group.com>. The refined models and associated RIBFIND clusters are deposited in the Zenodo repository, at <https://dx.doi.org/10.5281/zenodo.8221020>.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Jae Anne Bach Hardie for her contributions. We also thank Guendalina Marini and Aaron Sweeney for their helpful feedback on the manuscript.

FUNDING

Wellcome Trust grant [209250/Z/17/Z]; cooperation of Leibniz Institute of Virology (LIV) and Universitätsklinikum Hamburg Eppendorf (UKE) (as part of Leibniz ScienceCampus InterACT, funded by the BWFGB Hamburg and the Leibniz Association) and the Landesforschungsförderung Hamburg (HamburgX). Funding for open access charge: Leibniz-Institute for Virology. *Conflict of interest statement.* None declared.

REFERENCES

- Patwardhan, A. (2017) Trends in the Electron Microscopy Data Bank (EMDB). *Acta Crystallogr D Struct Biol*, **73**, 503–508.
- ENCODE Project Consortium, Snyder, M.P., Gingeras, T.R., Moore, J.E., Weng, Z., Gerstein, M.B., Ren, B., Hardison, R.C., Stamatoyannopoulos, J.A., Graveley, B.R. *et al.* (2020) Perspectives on ENCODE. *Nature*, **583**, 693–698.
- Berman, H.M., Lawson, C.L. and Schneider, B. (2022) Developing community resources for nucleic acid structures. *Life*, **12**, 540.
- Coimbatore Narayanan, B., Westbrook, J., Ghosh, S., Petrov, A.I., Sweeney, B., Zirbel, C.L., Leontis, N.B. and Berman, H.M. (2014) The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res.*, **42**, D114–D22.
- Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.H., Srinivasan, A.R. and Schneider, B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
- Miao, Z., Adamiak, R.W., Antczak, M., Boniecki, M.J., Bujnicki, J., Chen, S.-J., Cheng, C.Y., Cheng, Y., Chou, F.-C., Das, R. *et al.* (2020) RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*, **26**, 982–995.
- Kretsch, R.C., Andersen, E.S., Bujnicki, J.M., Chiu, W., Das, R., Luo, B., Masquida, B., McRae, E.K.S., Schroeder, G.M., Su, Z. *et al.* (2023) RNA target highlights in CASP15: Evaluation of predicted models by structure providers. *Proteins*, <https://doi.org/10.1002/prot.26550>.
- Pandurangan, A.P. and Topf, M. (2012) RIBFIND: a web server for identifying rigid bodies in protein structures and to aid flexible fitting into cryo EM maps. *Bioinformatics*, **28**, 2391–2393.
- Pandurangan, A.P. and Topf, M. (2012) Finding rigid bodies in protein structures: application to flexible fitting into cryoEM maps. *J. Struct. Biol.*, **177**, 520–531.
- Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H. and Westhof, E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
- Rose, A.S. and Hildebrand, P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
- Miyazawa, S. and Jernigan, R.L. (1996) Residue – residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, **256**, 623–644.
- Bottaro, S., Di Palma, F. and Bussi, G. (2014) The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res.*, **42**, 13306–13314.
- Bernauer, J., Huang, X., Sim, A.Y.L. and Levitt, M. (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA*, **17**, 1066–1075.
- Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
- Cragolin, T., Kryshchak, A. and Topf, M. (2021) Cryo-EM targets in CASP14. *Proteins*, **89**, 1949–1958.
- Costa, M., Walbott, H., Monachello, D., Westhof, E. and Michel, F. (2016) Crystal structures of a group II intron lariat primed for reverse splicing. *Science*, **354**, aaf9258.
- Abeyrathne, P.D., Koh, C.S., Grant, T., Grigorieff, N. and Korostelev, A.A. (2016) Ensemble cryo-EM uncovers inchworm-like translocation of a viral IRES through the ribosome. *Elife*, **5**, e14874.
- Rodnina, M.V., Fischer, N., Maracci, C. and Stark, H. (2017) Ribosome dynamics during decoding. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **372**, 20160182.
- Frank, J. (2017) The translation elongation cycle—capturing multiple states by cryo-electron microscopy. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **372**, 20160180.
- Fischer, N., Konevega, A.L., Wintermeyer, W., Rodnina, M.V. and Stark, H. (2010) Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature*, **466**, 329–333.
- Agirrezabala, X., Lei, J., Brunelle, J.L., Ortiz-Meoz, R.F., Green, R. and Frank, J. (2008) Visualization of the hybrid state of tRNA binding promoted by spontaneous ratcheting of the ribosome. *Mol. Cell*, **32**, 190–197.
- Budkevich, T., Giesebrecht, J., Altman, R.B., Munro, J.B., Mielke, T., Nierhaus, K.H., Blanchard, S.C. and Spahn, C.M.T. (2011) Structure and dynamics of the mammalian ribosomal pretranslocation complex. *Mol. Cell*, **44**, 214–224.
- Agirrezabala, X., Liao, H.Y., Schreiner, E., Fu, J., Ortiz-Meoz, R.F., Schulten, K., Green, R. and Frank, J. (2012) Structural characterization of mRNA-tRNA translocation intermediates. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 6094–6099.
- Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.

26. Das,R. and Baker,D. (2007) Automated *de novo* prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 14664–14669.
27. Yan,Y. and Huang,S.-Y. (2018) RRDB: a comprehensive and non-redundant benchmark for RNA-RNA docking and scoring. *Bioinformatics*, **34**, 453–458.
28. Adams,P.D., Afonine,P.V., Bunkóczi,G., Chen,V.B., Davis,I.W., Echols,N., Headd,J.J., Hung,L.-W., Kapral,G.J., Grosse-Kunstleve,R.W. *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 213–221.
29. Cragolini,T., Sahota,H., Joseph,A.P., Sweeney,A., Malhotra,S., Vasishtan,D. and Topf,M. (2021) TEMPy2: a Python library with improved 3D electron microscopy density-fitting and validation workflows. *Acta Crystallogr. D Struct. Biol.*, **77**, 41–47.
30. Topf,M., Lasker,K., Webb,B., Wolfson,H., Chiu,W. and Sali,A. (2008) Protein structure fitting and refinement guided by cryo-EM density. *Structure*, **16**, 295–307.

3.3 Publication in Proteins: Structure, Function & Bioinformatics

The Critical Assessment of Structure Predictions (CASP) happens on a biennial basis since 1994. Here the state of the art in structure prediction methods are assessed. Participating groups (Predictors) must predict the structure of proteins or protein assemblies of various complexity using the provided amino acid sequence alone. In CASP15, a large improvement was observed in complex prediction (Ozden et al. 2023) consistent with the introduction of AlphaFold-Multimer (Evans et al. 2022). The increasing relevance of cryo-EM in the structural biology field was reflected in the large number of targets which were solved using this method. As previously described, structural predictions are central to many low resolution cryo-EM modelling approaches such as flexible fitting. This makes CASP an ideal test-bed for understanding the quality of the current state-of-the art for flexible fitting purposes.

In CASP13 (Kryshtafovych, Malhotra, et al. 2019) and CASP14 (Cragolini, Kryshtafovych, et al. 2021), individual chains or domains of predictions were flexibly fitted into cryo-EM maps. Given the advances in complex prediction, in CASP15 entire monomers and complexes were refined into the cryo-EM data provided by experimentalists.

I lead this collaborative effort, which involved Rachael Kretsch from the Das group at Stanford and Luc Elliott from the Rigden group at University of Liverpool. My main contributions were the development of the cryo-EM refinement and validation pipeline, as well as the assessment of the flexible fitting results. The RNA prediction category, which was a new category in CASP15 (Das 2023), offered a number of challenges. Many of these RNAs were small, flexible, and solved to between 3.5 and 7.6Å using cryo-EM. Here we refined only the best RNA predictions for each target, according to the expert assessment of Rachael Kretsch. Due to the low resolutions of the provided cryo-EM maps for RNA structures, and poor initial backbone geometry of predictions, good RNA geometry was not achievable using the TEMPy-ReFF tool alone. To this end, I implemented an RNA specific pipeline to handle these lower quality predictions, where ERRASER2 (Chou et al. 2013) was employed along side TEMPy-ReFF to produce high quality fitted models which were in some cases better than the target structures.

CASP15 cryo-EM protein and RNA targets: Refinement and analysis using experimental maps

Thomas Mulvaney^{1,2}  | Rachael C. Kretsch³  | Luc Elliott⁴  |
Joseph G. Beton¹  | Andriy Kryshatfovych⁵  | Daniel J. Rigden⁴  |
Rhiju Das^{3,6,7}  | Maya Topf^{1,2} 

¹Centre for Structural Systems Biology (CSSB), Leibniz-Institut für Virologie (LIV), Hamburg, Germany

²University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany

³Biophysics Program, Stanford University School of Medicine, California, USA

⁴Institute of Systems, Molecular & Integrative Biology, The University of Liverpool, Liverpool, UK

⁵Genome Center, University of California, Davis, California, USA

⁶Department of Biochemistry, Stanford University School of Medicine, California, USA

⁷Howard Hughes Medical Institute, Stanford University, California, USA

Correspondence

Maya Topf, Centre for Structural Systems Biology (CSSB), Leibniz-Institut für Virologie (LIV), Hamburg, Germany.
Email: maya.topf@cssb-hamburg.de

Funding information

Leibniz Institute of Virology; BWFGB Hamburg; Landesforschungsförderung Hamburg (Hamburg-X); Collaborative Computational Project for Electron cryo-Microscopy (CCP-EM); US National Institute of General Medical Sciences (NIGMS, National Institutes of Health), Grant/Award Numbers: R35 GM122579, R01 GM100482; Stanford Bio-X; Howard Hughes Medical Institute

Abstract

CASP assessments primarily rely on comparing predicted coordinates with experimental reference structures. However, experimental structures by their nature are only models themselves—their construction involves a certain degree of subjectivity in interpreting density maps and translating them to atomic coordinates. Here, we directly utilized density maps to evaluate the predictions by employing a method for ranking the quality of protein chain predictions based on their fit into the experimental density. The fit-based ranking was found to correlate well with the CASP assessment scores. Overall, the evaluation against the density map indicated that the models are of high accuracy, and occasionally even better than the reference structure in some regions of the model. Local assessment of predicted side chains in a 1.52 Å resolution map showed that side-chains are sometimes poorly positioned. Additionally, the top 118 predictions associated with 9 protein target reference structures were selected for automated refinement, in addition to the top 40 predictions for 11 RNA targets. For both proteins and RNA, the refinement of CASP15 predictions resulted in structures that are close to the reference target structure. This refinement was successful despite large conformational changes often being required, showing that predictions from CASP-assessed methods could serve as a good starting point for building atomic models in cryo-EM maps for both proteins and RNA. Loop modeling continued to pose a challenge for predictors, and together with the lack of consensus amongst models in these regions suggests that modeling, in combination with model-fit to the density, holds the potential for identifying more flexible regions within the structure.

KEYWORDS

3D structure prediction, AlphaFold, CASP, CASP15, cryoEM, protein structure, refinement, RNA, RNA structure

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Proteins: Structure, Function, and Bioinformatics* published by Wiley Periodicals LLC.

1 | INTRODUCTION

Assessment of models in CASP is traditionally based on comparing predicted coordinates with the coordinates of reference structures provided by experimentalists. For evaluation purposes, the experimental structures are considered the “gold standard”. However, experimental structures by their nature are only models themselves—their construction involves a certain degree of subjectivity in interpreting density maps and translating them to atomic coordinates. In several previous CASPs, in parallel to the coordinate-to-coordinate evaluation, we carried out an evaluation of models versus the experimental data for a subset of cryo-EM-derived structures.^{1,2} In this article, we continue this trend and check the fit of CASP15 models to cryo-EM density maps. We also study how the density-guided refinement of the best models improves their fit to map, and how the refined models fare with regards to the experimental structures. For the first time, besides the protein targets, we analyze RNA structures.

The number of structures newly solved by 3D-EM roughly doubles every 2 years and totals 14 500 as of March 2023, constituting more than 8% of protein structures in the whole PDB (<http://www.rcsb.org/>)³ (compared to around 4% only 2 years ago). Reflecting this growth, CASP also registered an uptick in the percentage of cryo-EM targets. In CASP14, 7 out of 54 evaluated targets (13%) were determined by cryo-EM, while in CASP15 the corresponding numbers were 27 out of 93 (29%), including 8 of the 12 (67%) RNA-containing structures.

While AlphaFold2 did not participate in the assembly category in the previous CASP experiment, it was noted that its predictions could have alleviated many interface modeling errors.⁴ Since then, AlphaFold-Multimer,⁵ RosettaFold⁶ and AF2Complex⁷ are a few examples of a growing number of deep-learning approaches to complex prediction. In CASP15, predictions of oligomeric targets were sufficiently good to directly refine whole proteins and complexes rather than smaller evaluation units. To test the applicability of the predictions in real-world cryo-EM structure determination tasks, we employed a method for ranking models. Additionally, given the improvement in the average cryo-EM map resolution, we decided to not only refine the best-predicted models into the corresponding maps but also assess higher resolution aspects of predicted models, such as their side-chain orientations.

For the RNA targets, predictions were ranked using their cryo-EM maps in another study of this special issue.⁸ Therefore, here we used the maps to refine the best-ranked RNA predictions. However, whilst cryo-EM for studying proteins can often achieve near-atomic resolution, for RNA-only structures this method generally has not yet been able to achieve the same levels of resolution. Additionally, structure prediction for RNA is far less mature than for proteins, making RNA refinement into cryo-EM maps even more challenging.

2 | MATERIALS AND METHODS

In CASP15, with the increased accuracy of modeling, we evaluated more targets, including multidomain and oligomeric ones (Figure 1,

Table 1). In this paper, we had two aims: (1) to assess how well each protein chain of the predictions fitted the density if it was docked individually in the map (i.e., in complexes, without the context of the fully predicted complex); and (2) to check whether the predicted models could be improved in the context of the experimental data. For the first aim, we ranked individual protein chains based on rigid cryo-EM docking (Section 2.1). For the second aim, we took the top-ranked model for each protein target and also used all the predictions for protein and RNA targets that passed minimum accuracy filters (see Section 2.2.1 for proteins and Section 2.2.2 for RNA). These were superposed on their corresponding reference structures and the fit of each model was then optimized with ChimeraX⁹ (Supp. Methods 1). This would show us that even when the prediction is not accurate enough, it can still serve as a good starting point for model building. For example, six targets shown in Figure 2 (T1154, H1158, T1158, T1170o, R1126, and R1156v3) were generally well modeled down to the secondary structure level; however the overall conformations only partly fitted the density. Below we describe the methodologies we used for the two approaches.

2.1 | Ranking of individual protein chains based on rigid cryo-EM docking

Instead of rigidly fitting the entire complex in the map, one can identify the optimal initial position for each of the protein components in the model using an exhaustive search or another heuristic. Predictions were re-ranked based on this global fitting approach using Cross-Correlation (CC).

The docking of models in this study was carried out using two automatic docking programs, Molrep^{10,11} and PowerFit.¹² Both programs use a six-dimensional search to maximize an overlap-correlation score between a given model and the map file. Molrep incorporates a Spherically Averaged Phased Translation Function (SAPTF), followed by a Rotation Function (RF) and Phased Translation Function (PTF), which achieves a suggested first fit and then improves the overlap score with a six-dimensional optimization search.^{10,11} On the other hand, PowerFit incorporates an exhaustive six-dimensional search, including rotation at a pre-set angle sampling density and translation across the map file. Input parameters for the docking included the input map file, model and resolution.¹² The top model was determined by the CC score calculated using ChimeraX.⁹

A group ranking was generated as follows using the complete *chain* submissions submitted by groups instead of the CASP-defined Evaluation Units (EUs).¹³ Predictors may submit five models for each target. To reduce the computational time required for the docking process, only the first submitted model for each target per group was considered. For each target, a score was assigned per group reflecting its position in the CC ranking for that target. The top model was given a score of 123 since this was the total number of groups. An automatic rank of 0 was given where a group did not submit a prediction for a given target. For an overall group ranking, a cumulative score for each group was tallied across all targets for which that group

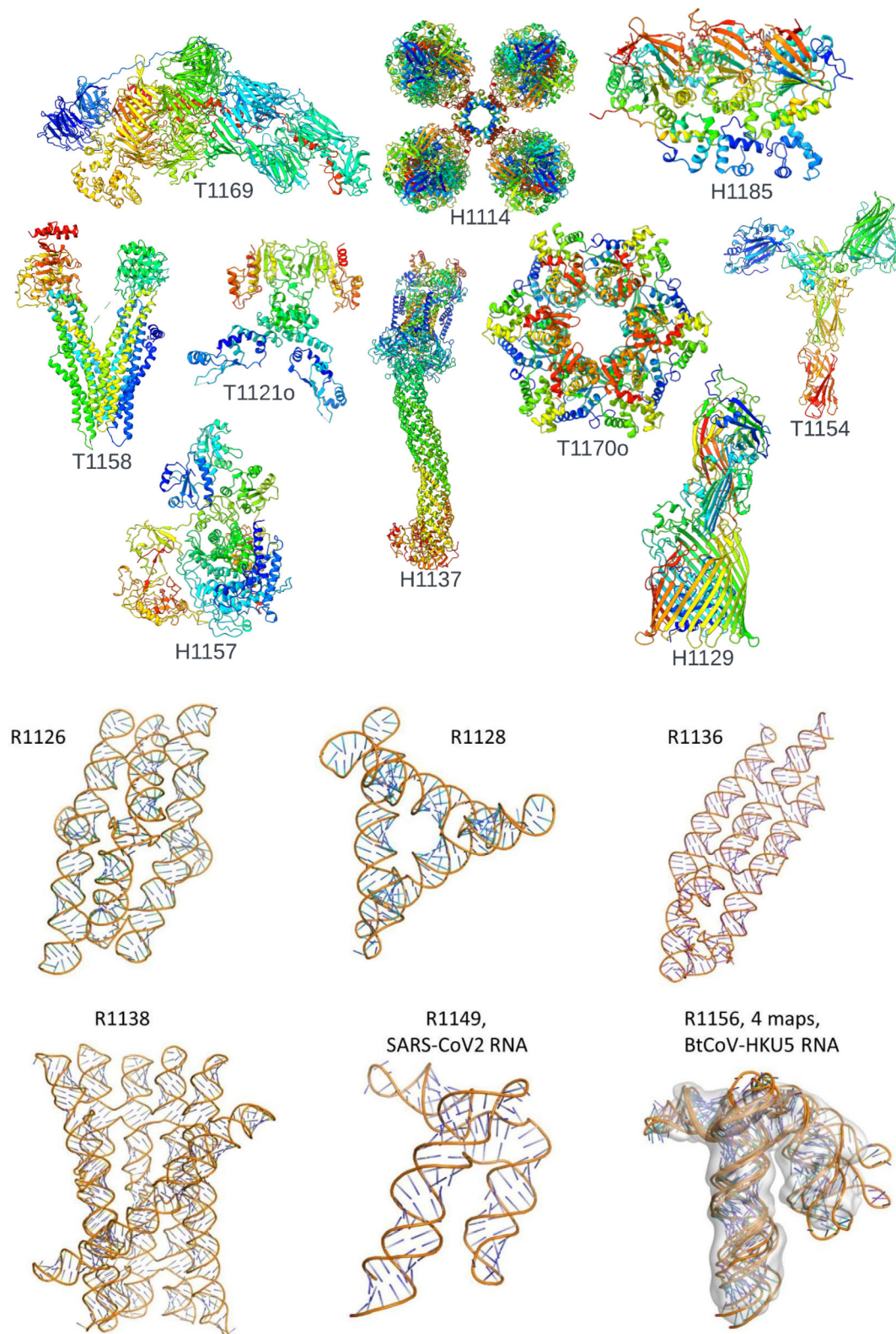


FIGURE 1 Overview of the cryo-EM targets used for refinement and analysis in CASP15: Reference structures for 10 protein targets (A) and 6 RNA targets (B) solved by cryo-EM in CASP15.

submitted a prediction. For comparison, similar rankings were done for each group and target using the composite S_{CASP15} score defined by Simpkin, et al.¹⁴

For single chain targets, the prediction from the top group was chosen as the starting candidate. For oligomeric targets (H1114, H1129, H1158, T1121o, T1170o, H1185), a cumulative score of the

TABLE 1 Overview of targets with refined predictions.

Target type	Target	Num. of predictions refined	Resolution (Å)	Num. residues/nucleotides
Protein	H1129	9	2.6	1387
	H1157	11	3.3	1524
	T1158	13	3.3	1340
	T1154	17	3.0	1424
	H1137	40	3.1	3939
	T1170o	11	3.0–3.3	1908
	H1185	13	3.4	1334
	T1121o	2	3.7	739
	T1169	2	3.3	3364
RNA	R1126	6	5.6	363
	R1128	7	5.3	238
	R1136v1	5	4.4	374
	R1136v2	5	3.5	374
	R1138v1	3	4.9	720
	R1138v2	3	5.2	720
	R11149	3	4.7	124
	R1156v1	1	5.8	135
	R1156v2	1	6.6	135
	R1156v3	4	7.6	135
	R1156v4	2	7.6	135

Note: Targets with predictions which met the minimum score criteria were refined. Note that for T1169 only two models were refined (see Section 3.2.1).

individual chains was tallied. The model from the highest scoring group across all chains for a target was selected for refinement. For these models, no attempt was made to recombine individual fitted chains: instead the originally submitted multi-chain assembly was re-docked so that this full assembly was the starting model for the refinement process.

2.2 | Model refinement

2.2.1 | Selection of models for refinement from proteins and protein complex-targets

Our rigid-body docking protocol is designed to test how well individual chains reflect the experimental density. However, we know from previous CASP competitions, that predictions, despite often modeling domains and SSEs to a high degree of accuracy, often fare less well when it comes to overall conformation. In previous papers of this series 1,2 we have performed flexible fitting and refinement on cryo-EM targets showing that, with the aid of the experimental data, models oftentimes can be as good as the reference structure. It is important to note that flexible fitting methods require the starting models to be quite accurate at the SSE and domain levels, as these features are not derived from the fitting process. Flexible fitting routines such as the one used in this paper may not converge if the

models are far from the global solutions. To select models which have both accurate SSEs and are not too far from the global optimum, we pick the highest ranking models based on the CASP assessment scores and the cryo-EM-based model assessment protocol (see Section 2.1). To qualify, predictions had to score above 0.7 on the IDDT (oligo-IDDT for oligomers) scale. Additionally, predictions for monomeric targets required a GDT_TS score greater than 0.7. In the case of oligomeric targets, predictions with QS, TM, and F1 scores^{4,15,16} all greater than 0.7, 0.8, and 0.6 respectively were eligible for refinement.

2.2.2 | Selection of models for refinement from RNA targets

All RNA-containing cryo-EM targets were considered for refinement. If there were multiple experimental maps, predicted models were selected separately for each map. The predictors were not asked to predict these conformations separately and hence, in some cases, the same predicted model was refined against multiple maps. Due to the low prediction accuracy all models submitted by each team were considered.⁸ The best models were selected as the top ranked structures across all submitted models based on the previously described map-to-model Z-score, Z_{EM} .⁸ Due to the fit qualities an automatic threshold would result in few models per target, so manual visual inspection was additionally used to select models that, even without

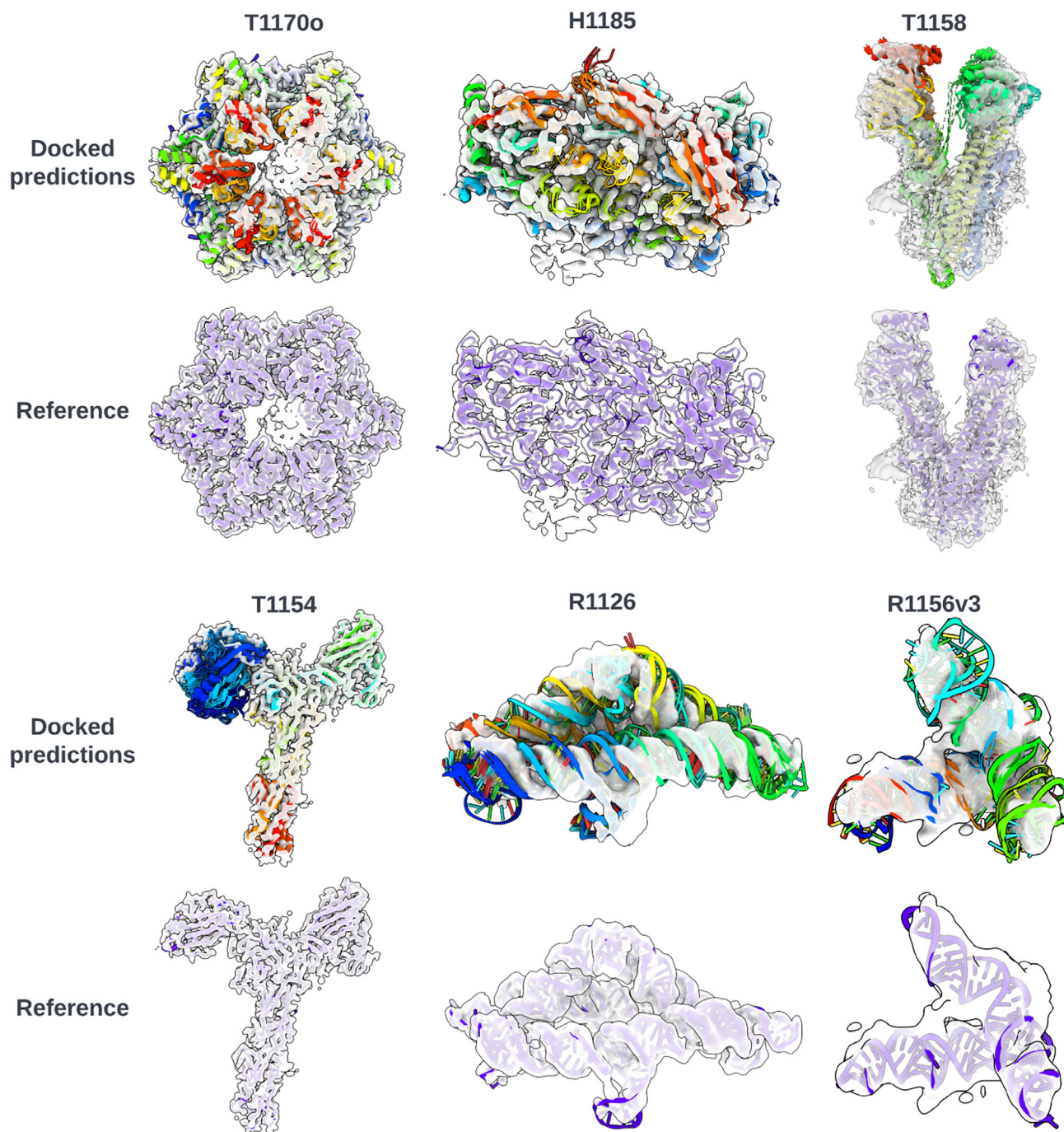


FIGURE 2 Docked predictions vs. the reference model for 6 CASP15 targets. The reference models are displayed in blue within the corresponding cryo-EM maps. The ensembles of docked predictions are shown in rainbow colors.

good fits, were the most promising for refinement. Based on these rankings and visual inspection of fit of the top 10 models by an expert, a final set of models for each target was selected.

2.2.3 | Model fitting and refinement

The refinement protocol was an incremental improvement on what was used in prior CASP challenges.^{1,2} In this CASP we additionally

incorporated an updated version of RIBFIND (RIBFIND2),¹⁷ which can help to improve the refinement process by clustering secondary structure elements (in both proteins and nucleic acid structures). Combined with ERRASER2, a yet to be published successor to Erraser¹⁸ for correcting geometry in RNA structures, this allowed the refinement of both protein and RNA predictions, even when significant conformational changes were required. A more in-depth description of the pipeline is available in the Supplemental Methods along with a graphical overview (Figure S1).

2.2.4 | Model assessment measures for proteins

The protein predictions for cryo-EM protein targets and the subsequent refined models were evaluated for their goodness-of-fit to the experimental cryo-EM density map (model-to-map goodness-of-fit) using the following metrics: The local (per-residue) goodness-of-fit was evaluated with the TEMPY2 Segmented Manders' Overlap Coefficient (SMOC) score¹⁹ and global goodness-of-fit using the ChimeraX cross-correlation measurement. The SMOC score represents the Manders' overlap coefficient for overlapping residue fragments: it is computed on local spherical regions around the seven residues in the current window. Overlapping windows are used, producing one numerical value per residue. SMOC scores can be calculated for the whole structure by averaging the per-residue scores. In order to compare the quality of fit to the density of side-chain vs. backbone, we have implemented two new "localized" SMOC scores in TEMPY: SMOCs and SMOCb. These scores assess the voxels around the side-chain atoms (SMOCs) and around the backbone atoms (SMOCb), respectively. To compute the SMOCs and SMOCb scores, each residue from the predictions was locally aligned to the target using the C-alpha atoms of the residue and its immediate neighbors. Because side-chains are a high-resolution feature, we did not use sliding windows in this case, that is, SMOCs and SMOCb scores were computed on the aligned residues. The geometry of the targets, the predictions, and the refined models were all assessed using MolProbity.²⁰

2.2.5 | Model assessment measures for RNA

For CASP15, we have implemented a new SMOC score in TEMPY—SMOCn—to assess the fit of nucleic acid chains. SMOCn is calculated similarly to the original SMOC score, which was designed to assess the protein chain in the density, by sliding windows around nucleotides instead of amino acids.¹⁹ Due to resolution limitation, the "localized" SMOCb and SMOCs scores were not used for RNA. As the RNA experimental maps were generally of a lower resolution than their protein counterparts, assessing geometry was important to ensure models were not overfit to the maps. RNA Validate, which is part of the Phenix²¹ software package, was used to assess the geometry of the RNA targets, predictions and refinements. We focussed our geometry analysis on the "average suitability" scores produced by RNA Validate. "Suites" are defined by the pucker of two consecutive backbone sugars and the five torsion angles between them. Empirical studies have shown that these suites inhabit a number of characterized states in 7-dimensional space. "Average suitability" is a measure of how well the suites in an RNA model match the discrete conformers found in the empirical data.²²

3 | RESULTS

3.1 | Ranking of protein models using docking into cryo-EM maps

Our comparison of docking results from PowerFit¹² and Molrep¹⁰ showed that PowerFit usually produced better fitting models

(Supp. Methods 2). We therefore carried out the ranking using PowerFit.

There was a significant, strong positive correlation between the cumulative S_{CASP15} rankings and the cryo-EM-based docking rankings (Figure 2). The top five groups from the docking rankings, in order, were: Yang, BAKER, GuijunLab-Assembly, FoldEver and PEZYFoldings. Each of these groups submitted predictions for all targets. The Yang group ranked consistently high on all targets and had the most (three) top ranking models (Table 2). Each of the top groups incorporated AlphaFold 2 style networks into their methods, with the exception of BAKER who used RosettaFold. For making performance comparisons, control representations of AlphaFold 2 are annotated (Figure 3) with group names NBIS-af2-multimer, NBIS-af2-standard, Colabfold, and Colabfold_human. Colabfold and Colabfold_human submitted predictions for every target but their results, while confirming the value of these readily available predictions for cryo-EM map fitting, were not amongst the very best. The best ranked prediction for each target was selected for refinement if it was not already selected based on the CASP criteria (see Section 3.2.1). These models are listed in Table 3. Target H1137 was excluded since, unlike other targets, there was no single group that had consistently suitable docked models across all interfaces.

3.2 | Protein targets—Refinement of top predictions

3.2.1 | Selection of protein targets for refinement using CASP criteria

We refined the 118 predictions for multi-domain proteins and protein complexes (Table 1) that either passed our filter based on CASP score (Section 2.2.2) or ranked first based on the fit of individual chains (Section 2.1). For 6 targets (Table S1), the top-ranked model based on docking of chains was not included in the list of models which passed the CASP filter. However, a comparison between the poses of the top-ranked docked models and the ones determined by superposition and optimization in ChimeraX shows high similarity (Figure S3). Except for these 6 top-ranked best models, we used the superposed ones as a starting point for refinement.

The only listed target which did not have models that passed the CASP-based selection criteria was T1169 (Table 1). Predictions of individual domains in T1169 were good but the full protein models were not accurate enough to pass the threshold due to partially inaccurate domain organization. This protein was the largest single chain model in CASP history with 5 domains and over 3000 residues. Here we chose the model with the highest GDT_TS score (GDT_TS = 57.7, IDDT = 0.63) which was from Yang-server (group 229). Finally, we did not refine predictions for target H1114 for which the corresponding cryo-EM map is at 1.52 Å resolution. Given the high resolution of the map and the high quality of the predictions for this target (the best model had a TM-score = 0.97, oligo-IDDT = 0.86, QS-score = 0.79, F1-score = 84.13), we decided to use it for a fine-tuned side-chain analysis instead.

TABLE 2 Group ranking based on docking.

Target	Top group by docking rankings	Top group by S_{CASP15} ranking	Groups selected for refinement
T1114s1	Gonglab-THU	SHT	FoldEver-Hybrid
T1114s2	Panlab	trComplex	
T1114s3	Yang	B11L	
T1121	GuijunLab-RocketX	GuijunLab-Threader	GuijunLab-RocketX
T1129	Venclovas	N/A	Venclovas
T1137s1	BhageerathH-Pro	PEZYFoldings	Venclovas ^a
T1137s2	SHORTLE	Yang	
T1137s3	RostlabUeFOFold	UM-TBM	
T1137s4	ACOMPMOD	N/A	
T1137s5	DELCLAB	UM-TBM	
T1137s6	RostlabUeFOFold	UM-TBM	
T1137s7	Shennong	DMP	
T1137s8	McGuffin	McGuffin	
T1137s9	Yang	PEZYFoldings	
T1154	Venclovas	Elofsson	Venclovas
T1157s1	Yang-Multimer	N/A	Yang-Multimer
T1157s2	Yang	N/A	
T1158	MULTICOM	Asclepius	MULTICOM
T1169	Shennong	Shennong	Shennong
T1170	FTBiot0119	MUFold_H	FTBiot0119
T1185s1	BhageerathH-Pro	BAKER	Yang-Multimer
T1185s2	Yang-Multimer	OpenFold-SingleSeq	
T1185s4	BAKER	Manifold-E	

Note: The top-scoring CC model for each target. Also indicated are the top-scoring groups for the same targets, in the general CASP assessment using the CASP15 score.¹⁴ Some chain models did not receive a CASP15 score because certain elements used in the CASP15 score formula were not calculated since the chain in question was split into multiple AUs. These were given an N/A classification.

^aThese targets were selected in a different way—see Section 2.1.

3.2.2 | Overall model analysis

Average SMOC scores of predictions prior to refinement were poor with a large degree of variation amongst the predictions for each target (Figure 4A). After refinement, average SMOC scores were closer to those of the respective targets, typically with significantly reduced variance. For example, the top models refined from the predictions of target T1154 had a SMOC curve very similar to that of the target SMOC curve (Figure 4B,C). Interestingly, all top predictions for this target based on the CASP criteria could be refined in the N-terminal part of the structure, despite its initial wrong orientation. This is likely to be attributed to the hierarchical refinement protocol, where the N-terminal is first pulled into the density as one rigid body. On the other hand, in the regions of residues 810–814 (Figure 5B), there is a sharp drop in the SMOC plot due to the “loopy” characteristics of the region (see below). In fact, most targets had some loops which did not reach the high SMOC scores seen in the rest of the structure after refinement, suggesting these regions were poorly modeled thus bringing down the average SMOC scores. We explore specific cases where loops were poorly modeled in Section 3.2.2. Overall MolProbity scores, which are a log-weighted

combination of the clash score, percentage of unfavoured Ramachandran dihedrals, and unfavorable side-chain rotamers, generally improved after refinement with scores less than 2.0 being common (typically, MolProbity scores below 3.0 are considered good). However, for a number of targets, the MolProbity scores were worse. In these cases (H1129, H1185, T1154), the provided maps had been processed using DeepEMhancer.²³

The six models that did not pass the initial CASP scoring criteria but ranked high based on PowerFit docking (Section 3.2.1, Table S1) were improved after refinement, generally exceeding the cutoffs for “accurate” models (Section 2.2.1). However, some scores for T1154 and T1121o were worse after refinement due to distortions. In the case of T1154, an incorrect interaction at the N-terminus caused a poor set of rigid-bodies to be generated during refinement. In the case of T1121o, a domain was misoriented and could not be optimized.

3.2.3 | Analysis of loop predictions

Given that overall the predictions were very accurate for proteins and that the top predictions required very little refinement in order to fit

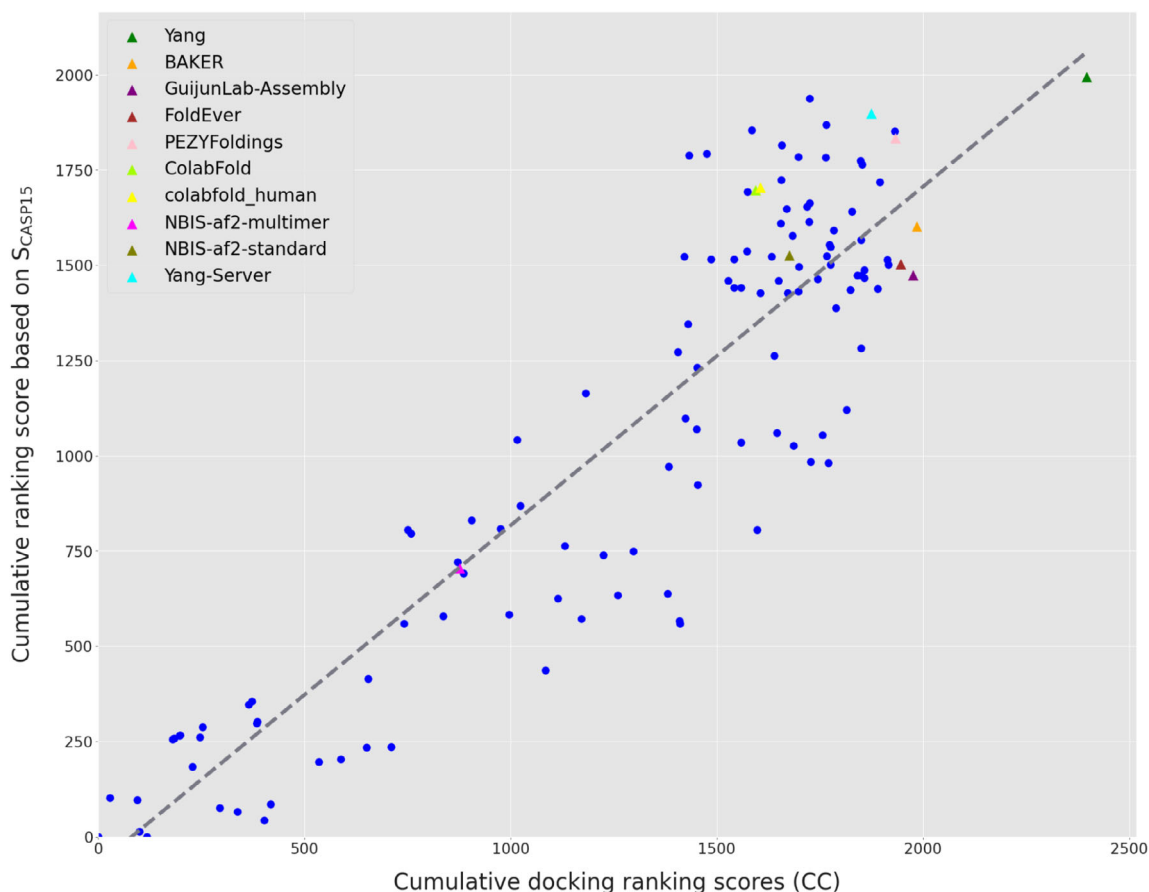


FIGURE 3 Group ranking for cryo-EM targets. Cumulative per-group docking ranking scores plotted against S_{CASP15} rankings across docking targets where S_{CASP15} scores were available (oligomeric reference structures were split into individual chains—see also Table 2). The gray line indicates the line of best fit with a strong positive correlation between the two rankings ($r = .827$, $p < .0001$). The top five performing docking ranking groups are labeled, as are the “control” AlphaFold 2 submissions. These groups are shown as triangles, others as blue circles.

well into their corresponding target cryo-EM maps, we decided to focus next on examining how well the loops in the top predictions were refined. Below are specific targets where the accuracy of loops was examined in detail.

H1157—Complex of CtEDEM and CtPDI1P at 3.3 Å resolution

This target consists of two proteins, each with multiple domains. These were modeled in a challenging experimental map with varying resolutions. These varying resolutions are clearly reflected by B-factor estimates produced by TEMPy-ReFF (Figure S2). Initial inspection of the target revealed minor modeling issues: some aromatic side-chains were not well-fitted to the density and a number of loops were in regions of the map that had resolution too low to be modeled with confidence. Interestingly, many of the predictions managed to produce side-chains which better fit the density than the reference. More intriguingly, the best predictions modeled a loop in chain A between residues 210–230 much better than in the target. These were further improved upon refinement (Figure 5A). Despite the excellent performance in modeling this large loop, predictors struggled to model some other loops.

T1154—S-layer protein A (SlaA) at 3.0 Å resolution

Many bacteria and archaea have a protein-based barrier which encapsulates the cell known as an S-layer. A CASP target of the outer S-layer component of the archaea *Sulfolobus acidocaldarius*²⁴ was well predicted at the domain level, with the model fit to the experimental data improving after the refinement. Despite the overall high-resolution, a short loop between residues 810–814 had very poor density. Predictions were unable to produce loops close enough to the correct geometry to be refined into the map (Figure 5B). Although automated refinement starting from these models was not possible, the general lack of consensus amongst the predictions likely reflected some degree of disorder which was mirrored by the poor resolution seen in this region of the map.

H1129—The bacteriophage pb5 protein in complex with FhuA at 3.1 Å DeepEMhancer map

Much like the swift adoption of deep-learning methods in the structure prediction community, deep-learning has been transforming image processing and reconstruction methods in the cryo-EM scene. Here, a dimeric complex of the bacteriophage pb5 protein and its

TABLE 3 RNA predictions which were selected for refinement.

Target	Group	Prediction model numbers
R1126	232	1–5
	287	2
R1128	232	1–5
	287	1,3
R1136v1, R1136v2	232	1,3,5
	287	4
	325	1
R1138v1, R1138v2	232	3,4,5
R1149	054	1
	125	3
	416	3
R1156v1	128	5
R1156v2	128	5
R1156v3	128	1,5
	232	3
	287	1
R1156v4	232	3
	439	2

binding partner (the bacterial outer-membrane protein FhuA) is derived from a map which had been sharpened using the deep-learning tool DeepEMhancer.^{23,25} Despite the overall high resolution of this map, residues 190 and 191 of a short loop were not modeled in the target structure with density dropping out in this region. Similar to the short loop in T1154, none of the predictions gave a “refineable” or even visually plausible fit in this region (Figure 5C). However, the model provided by Wallner (group 037) was, by visual inspection, close and could potentially be locally fitted and refined using interactive tools such as Coot²⁶ or ISOLDE.²⁷ Despite often making visual interpretation easier, an unfortunate side effect of DeepEMhancer is that lower-resolution regions of the map tend to be removed. It is possible that the unprocessed map (which we did not have) may have offered better information about this likely disordered region. A number of poorly resolved loops had a higher atomic B-factor, as determined by TEMPy-ReFF (Supp. Methods 1), compared to the rest of the model. Interestingly, we observed a similar pattern in the root mean square fluctuation (RMSF) of the best predictions (Figure S2). We thus hypothesize that these poorly resolved portions of the map were caused by increased local mobility, which was also captured by the predictions which deviated from one another in these lower resolution regions.

3.2.4 | Analysis of side-chain predictions

To examine how well CASP predictors can predict side chains, we analyzed predictions for target H1114, which was determined based

on a high-resolution map (1.52 Å). The target is a hydrogenase isolated from *Mycobacterium smegmatis* that forms a large oligomeric complex of the HucS, HucL, and HucM proteins.²⁸ The SMOC scores for backbone and sidechain atoms of unrefined predictions compared against those of the target for each residue are shown in Figure 6. Sidechain SMOC scores (SMOCs) were clearly not predicted as well as the backbone scores (SMOCb), suggesting poor atom placement (Figure 6A). An example is model 1 from Yang (group 439). In this case, although the backbone was relatively well fitted (average SMOCb = 0.72), some side chains were incorrectly positioned, such as those of GLU15 and HIS166 (Figure 6B).

3.2.5 | Refinement of T1169—The mosquito salivary gland surface protein 1 at 3.3 Å resolution

Target T1169 is the mosquito salivary gland surface protein 1, a monomeric protein composed of more than 3000 residues involved in pathogen transmission from mosquitos. None of the predictions passed our CASP criteria for multidomain protein refinement (GDT-TS > 0.7 and LDDT > 0.7). This is potentially due to the existence of a domain in T1169 with a previously unidentified fold, and others with low sequence homology to known structures.²⁹ Therefore, we decided to compare between the top-fit prediction based on chain ranking which was from Shennong (group 466), against the prediction with the highest GDT-TS score (57.7) which was from Yang-server (group 229) (Figure 7A). The Shennong model was ranked third based on GDT-TS with a score of 54.1. Note that based on global fit-to-density using ChimeraX cross-correlation (CC) scores, the Yang-server model also had a better correlation with the experimental map (CC = 0.55 for Shennong and CC = 0.61 for Yang-server). The refined models of each of these predictions are shown in the 3.3 Å cryo-EM map (Figure S4A). SMOC scores of the predicted models show that each prediction has regions that are more accurate than the other. From the corresponding SMOC plot (Figure S4B), the CASP-criteria selected prediction produced a better refined model with a SMOC profile closer to that of the target. The poorer refinement of the Shennong group prediction (Table S1) is likely due to the incorrect placement of the N-terminal β-propeller towards the center of the molecule (residues 1–340), which could not be fixed during refinement (Figure S4B).

3.3 | RNA targets: Refinement of top predictions

3.3.1 | Selection criterion of RNA targets for refinement

Six of the eight RNA-containing targets were selected for refinement. The two RNA-protein complexes (RT1189, RT1190) were not selected as targets due to poor prediction accuracy (RMSD > 15.9 Å, GDT_TS < 27). A separate analysis of these predictions was performed instead.⁸ Furthermore, no predictions passed the CASP-scored

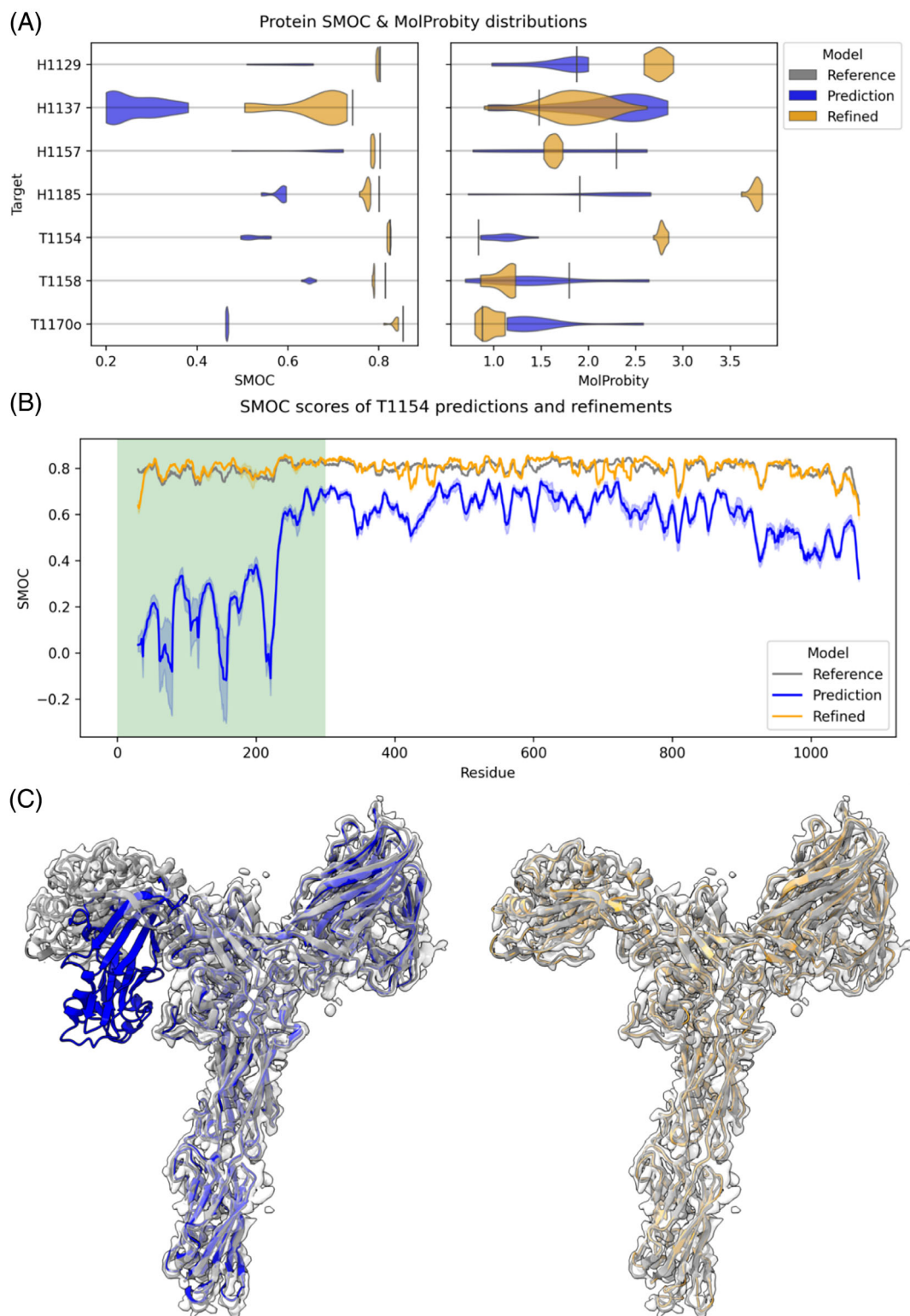


FIGURE 4 Overview of protein refinement results. In (A), the distribution of average SMOC scores for the qualified CASP predictions before and after refinement. Score for the experimental model is shown as a vertical line. Target T1169 and T1121o were not included as only two models for each were refined. In (B), the residue level SMOC plot is shown for T1154 and its predictions. The dark orange and blue lines are the mean refined and docked SMOC scores with the minimum and maximum values in light orange and blue. The N-terminal domain, which fitted poorly in all of the predictions (as indicated by the highlighted region), needed significant movement during refinement and is shown in (C) for model 1 from PEZYFoldings (group 278). Plots and 3D structures are in orange for refined models, in gray for reference structures and in blue for predictions.

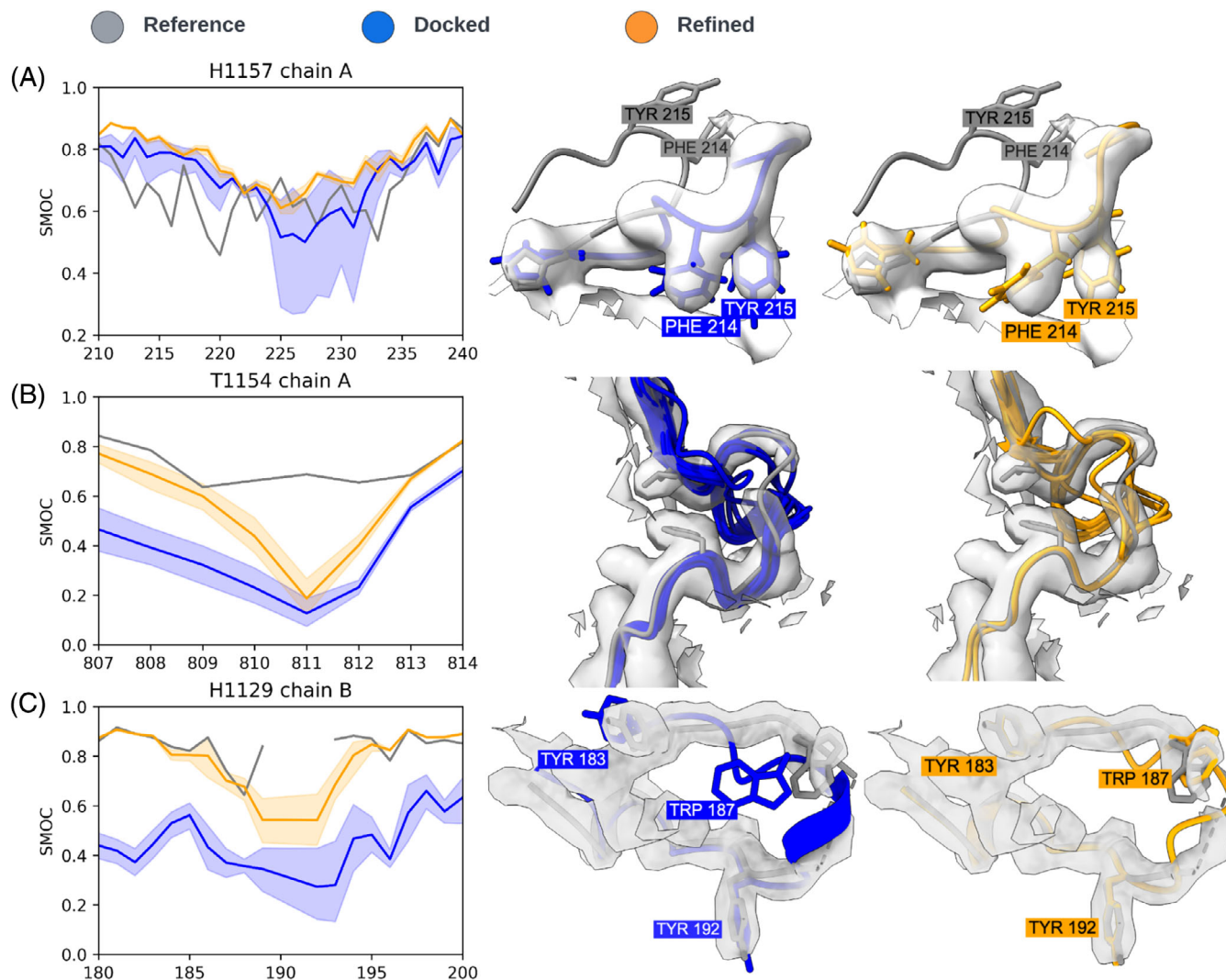


FIGURE 5 Protein loop case-studies. In all the visualizations, the target model is gray and the predictions are blue and orange before and after refinement, respectively. The dark orange line in the plot is the mean SMOC score, with the shaded region representing the minimum and maximum value for the set of predictions. (A) The reference model (for H1157) had a large poorly modeled loop in chain A as indicated by the low SMOC scores in 210–230 region. The best-refined predictions were a much better fit. In orange, a refined prediction from McGuffin (group 180). (B) This short loop, in T1154 was not modeled well enough by any predictions to be refined into the density. The low-intensity density may also be an indicator that this region is disordered. (C) Residues 190–191 of chain B were not modeled in the reference model for H1129 indicated by the dotted line. None of the predictions were able to produce a refinable loop that fitted the DeepEMhancer-sharpened map in this region. However, the model submitted by Wallner (group O37) which is depicted, was visually the best fitting before and after refinement.

selection for proteins ($GDT_TS > 0.7$, $IDDT > 0.7$) so we used an alternative selection process for RNA models. For each target, the Z_{EM} ranking was used to obtain a top 10 models which were then visually inspected to obtain a set of models we thought most likely to be refined by criteria such as limited geometric problems, and minimal chain distortions needed to move into map⁸ (Table 3). Targets R1126, R1128, and R1149 had a single experimental structure and thus their top models by Z_{EM} were selected and after manual fitting; 6, 7, and 3 models were refined, respectively. For the three remaining RNA-only cryo-EM targets, multiple experimental maps were used for refining the predicted structures.

For R1136, the two experimental maps, representing the ligand bound and unbound conformations, were topologically very similar, so the same models (5 total) were selected to refine into both maps.

R1136 included 15 submitted models with the same RNA structure – they differed in their ligand prediction—so only 325_1 was used for refinement. For R1138, all top predictors were closer to the “mature” state, with no predictions close to the “young” state according to global topological and fit-to-map metrics. The top models (3 total) for the “mature” state were thus refined to both maps. For R1156 each map was considered separately resulting in 8 total refinements.

3.3.2 | Overall RNA model analysis

The RNA predictions had average SMOC scores above 0.8 after refinement for all but the young conformation of R1138 discussed

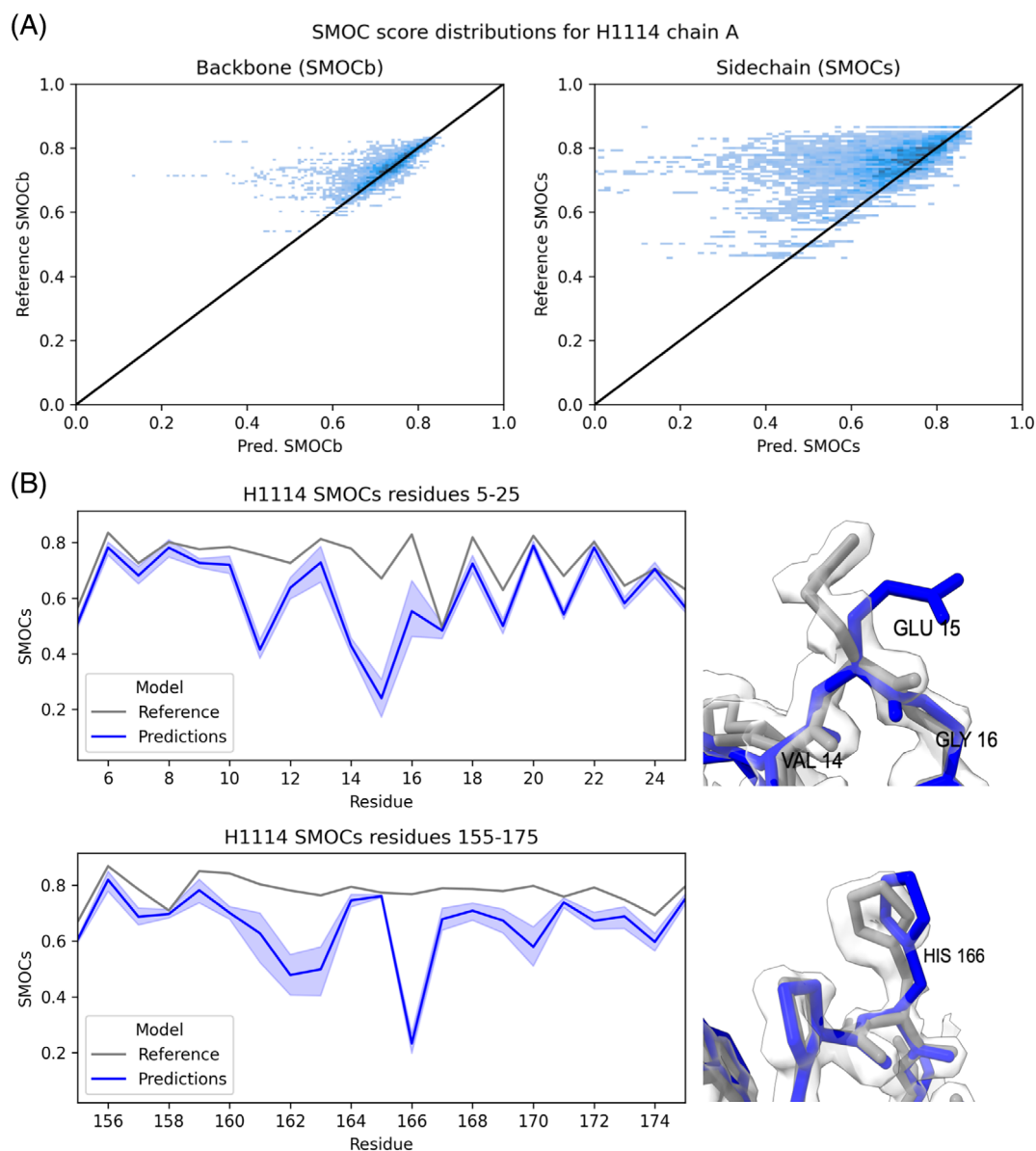


FIGURE 6 Side-chain analysis of H1114. SMOC scores for backbone and sidechain atoms of H1114 predictions compared against those of the target reference structure for each residue (A). Backbone SMOCb scores (left) and sidechain SMOCs scores (right) of the reference structure vs the predictions. In (B) incorrectly positioned side-chains of GLU15 and HIS166 from model 1 prediction by Yang (group 439) (blue) compared to the reference (gray). These residues were consistently poorly placed by predictors.

below, despite predicted models starting far from the reference structure (all GDT_TS<0.7) (Figure 7A). In fact, for R1128, R1138v2, R1149, and R1156v3 targets, refined predictions surpassed the SMOC values of models fitted into the same RNA cryo-EM maps as reference models (Figure 7A). Further, while prediction started with a spread of SMOC scores, the variance in SMOC score was reduced upon refinement. These results indicate that the refinement procedure was successful in fitting the models into the maps, moving all predictions to a similar solution, even in cases where large changes were needed. Compared to protein models where the fit of loops and side-chains could be assessed due to the higher resolution of the experimental maps, here the focus was on the overall fit of high level features.

R1138 a 6-helix bundle at 4.9 Å resolution

A particularly interesting example for cryo-EM refinement of RNA models was the predictions and refinement for R1138, a designed 6-helix bundle of RNA with a clasp (6HBC).³⁰ This target had reference structures and experimental maps for two alternative conformations, a short-lived “young” conformation, and a stable “mature” conformation. The refinements for the mature conformation gave a better fit to the experimental density than the target reference structure (Figure 7B) with the majority of residues having higher SMOC scores than those in the target reference structure. These predictions required significant conformational change as seen in Figure 7C and Video S2. The overall geometry, as assessed by the “average suiteness” score (see Methods), was also better in the

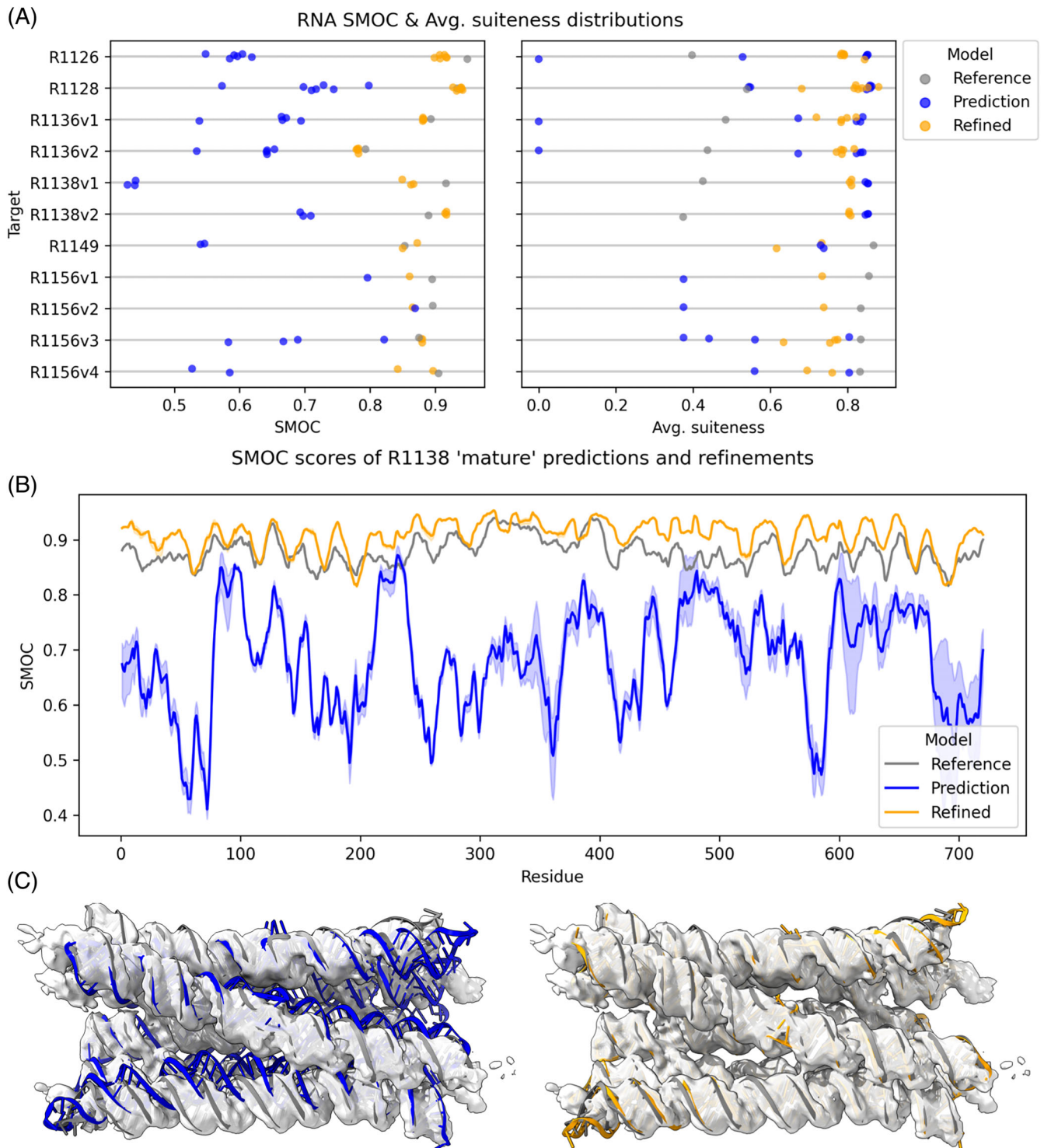


FIGURE 7 Overview of RNA refinement results (A) The average SMOC scores for the target, predictions, and refined predictions are shown alongside the RNA Validate “average suiteness.” (B) The residue level SMOC plots of R1138 in the mature conformation map and the predicted and refined models. The dark blue and orange lines are the average SMOC score for the predictions and refinements respectively, with the lightly shaded area representing the minimum and maximum values. (C) An R1138 prediction by Alchemy_RNA2 (group 232) in the “mature” conformation map. Depicted are the prediction (blue) and refined prediction (orange) with respect to the reference model (gray).

refined models than the reference structure (Figure 7A). However, CASP predictions for the “young” conformation failed to refine to the same extent (Figure 7A, Video S3). This poorer result might be

attributed to the greater degree of rearrangement of the helices and the breaking and reforming of hydrogen bonds in the kissing loop clasp required to convert from models resembling the mature

conformation to the early conformation. The breaking and forming of such hydrogen bonds can in principle occur, but is unlikely, in our refinement protocol.

R1126 a designed “Traptamer” at 5.6 Å resolution

The refined predictions of the designed RNA target R1126, a designed RNA origami scaffold for a Broccoli and Pepper aptamer pair,³¹ had lower average SMOC scores than the reference target structure. However, this result may be due to the reference structure being overfitted to the cryo-EM map at the expense of realistic RNA geometry, as reflected by the low suiteness scores of the target structure compared to the refined models (Figure 8A). Selected predictions for this target had a large degree of conformational diversity with models varying between 9 and 13 Å RMSD from the target. Despite our refinement protocol improving the overall fit-to-map and improving the geometry of some of these predictions, a number of predictions from Alchemy_RNA2 (group 232) exhibited an incorrect crossover between strands (Figure 8A). Fixing such issues would require breaking and rebuilding chains which is not allowed in our refinement protocol.

Both Alchemy_RNA2 (group 232) and Chen (group 287) provided a number of predictions which offered excellent refined models. All of these predictions required significant conformational changes to fit the experimental map. Often these movements involved breaking predicted interactions. One striking example is in the second prediction from Chen (Figure 8B, Video S1). In the prediction, a stem-loop was curled around and interacting with an upstream helix. In order to fit the density, the stem-loop interaction was broken allowing it to move into density.

R1156v3–BtCoV-HKU5 SL5 at 7.6 Å

Maps and reference structures for four alternative conformations of the SL5 domain from 5'UTR from the Bat coronavirus BtCoV-HKU5 were provided for assessment in this CASP. This domain is known to have a conserved secondary structure in many coronaviruses,^{32,33} which is thought to be important in the packaging of viral particles during infection.³⁴ Maps for this target varied in resolution from 5.6 to 7.6 Å. The four refined predictions for the third conformation (R1156v3) exhibited average SMOC values slightly higher than the reference structure. Although the suiteness scores for the refined predictions were lower than for the reference structure, in all but one case they were better than the unrefined predictions. In contrast to the Traptamer example above, where refinement involved the breaking of an interaction of a apical loop, the refinement of the second prediction from Alchemy_RNA2 involved the formation of an interaction between a apical loop and an internal loop in another part of the model (Figure 8C).

4 | DISCUSSION

In CASP15, 29% of the total targets, including 67% of the RNA-containing targets, were determined using cryo-EM. The accuracy of predictions for protein targets assessed in this paper and the overall

quality of experimental maps allowed many predictions to be further refined to near-native conformations. Compared to most CASP assessments, where a single reference model has been used as the ground truth, cryo-EM assessment finds itself in a privileged position. To aid the assessment, cryo-EM maps are typically available in conjunction with target reference models—which are after all just best attempts at model building using the experimental map, human knowledge, and current state of the art technology. This is particularly important, as cryo-EM data tends to have lower resolutions than crystallographic experiments. Because 3D reconstructions are built from averages of many particles, they may also capture continuous motions and flexibility of the visualized macromolecule, which can then manifest itself as lower resolution regions. There is thus an added degree of uncertainty in any static 3D structure that is derived from cryo-EM data.

One model, which particularly highlighted the importance of experimental data this year, was H1157. This model had an average resolution of 3.3 Å with many regions of the map having lower local resolution. Intriguingly, a large loop which was erroneously modeled in the target was much better modeled by the top predictions, with aromatic side chains well placed in the density. If, on the other hand, we only had the target model as ground truth (i.e., we did not use the experimental map for assessment), these better predictions would have not been noticed.

For the majority of targets, where the author's submitted model (target reference model) and experimental map were in good agreement, some parts of the predicted models resulted in better fit to map following refinement. At the same time, many targets had loops which were not well predicted. Typically, the geometry of these loops varied amongst predictions, with many failing to be refined because they were too distant from the target. The lack of consensus amongst some of these loops was often reflected by lower local resolutions in the experimental map (Figure S2). While we did not investigate the relationship between these two phenomena in this paper, in CASP14 cryo-EM assessment, we showed anticorrelation between the standard deviation of the SMOC scores of the predicted models (SMOC SD) and SMOC scores of the target structures.²

The strong correlation between the rankings based on the cryo-EM-based docking score and the composite S_{CASP15} score shows that high quality models can often be picked using experimental data alone. For model building practitioners, this is particularly relevant, as reference structures may not be available. Given the difficulty of building models into experimental maps and the fact that there is not a single prediction tool which excels across all targets, docking, and ranking offers an approach to screen for good starting models, potentially from multiple structural prediction tools. Some maps provided by the experimentalists had been sharpened with DeepEMhancer.²³ This caused a degradation in MolProbity scores, likely because the TEMPy-ReFF GMM³⁵ puts more weight on the sharpened map, overpowering the geometry restraints. Another unfortunate side-effect of DeepEMhancer maps was that low-resolution regions tended to disappear entirely in the sharpened maps. DeepEMhancer attempts to reproduce the sharpening produced by LocScale³⁶ but without the need for an atomic model. However, this deep-learning approach

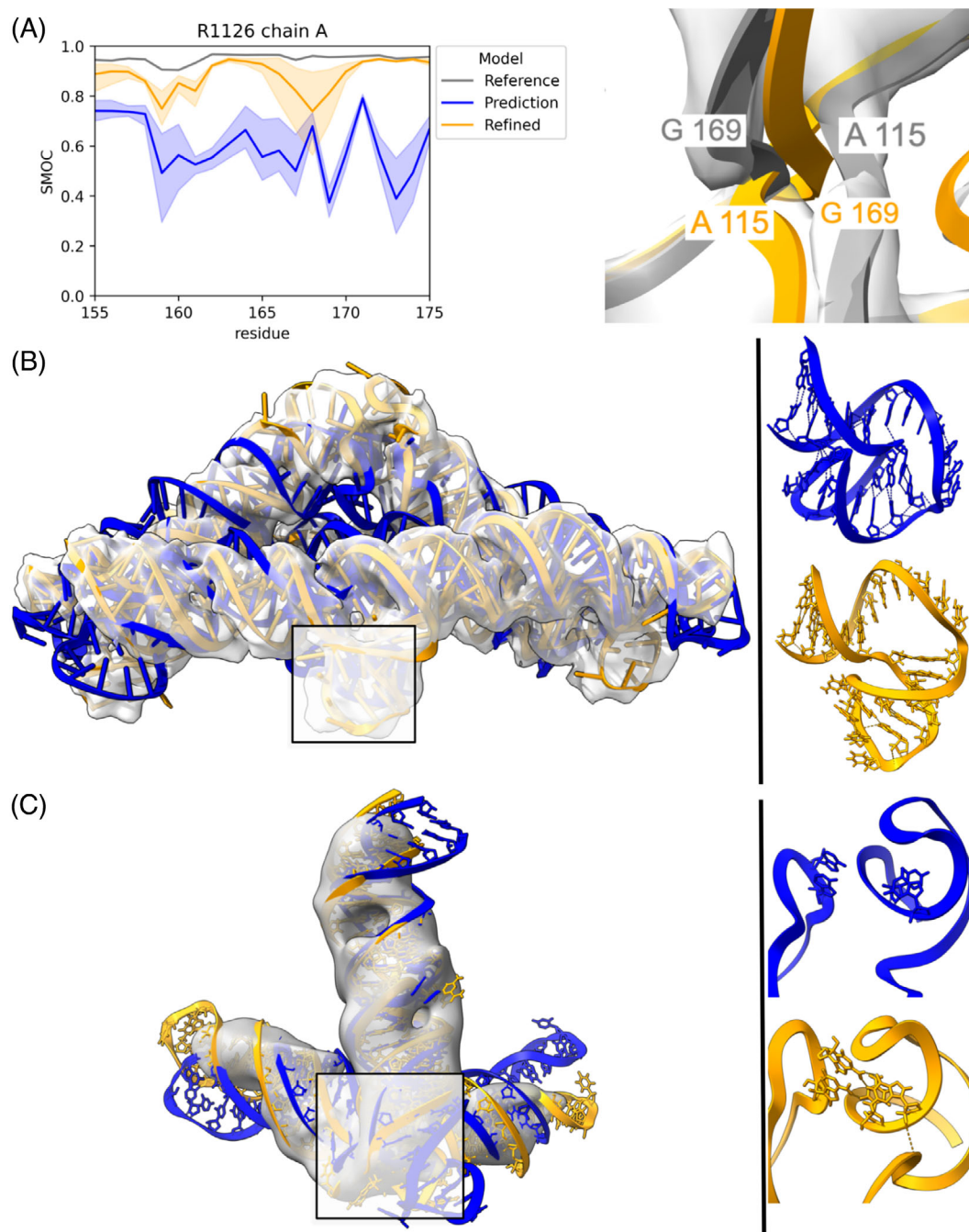


FIGURE 8 RNA refinement case studies. In all the visualizations, the target model is gray and the predictions are blue and orange before and after refinement, respectively. The dark orange line in the plot is the mean SMOC score, with the shaded region representing the minimum and maximum value for the set of predictions. (A) A SMOC plot of R1126 predictions and their refinements. Some refinements had residues between 155 and 175 with a variable SMOC score, large shaded region. This was due to strands crossing over, in some of the predictions, as shown in the right panel. (B) A model of R1126 from Chen (group 287) and its refinement. Overall, the R1126 predictions were refinable despite large conformational changes often being required. On the right, a close-up of the highlighted area showing the breaking of loop interaction during refinement. (C) A model of R1156 from Alchemy_RNA2 (group 232) and its refinement. After refining the model into the third conformation map, it better fitted the experimental density. On the right, a close-up of the highlighted area showing the formation of new interactions between an apical loop and an internal loop.

tends to remove low resolution regions entirely, creating maps that look like they have been tightly masked, and may also hallucinate density. The current consensus on this emerging technique amongst the cryo-EM community is that such maps should not be used for refinement or measuring map-model quality but rather as an intermediate

aid during model building. We would thus advocate for structure providers to offer raw maps with processed maps optional in future CASP challenges. Many of the predictions displayed a diverse set of loops in these regions. While sharpened maps may aid in model building, low-resolution regions can be an important indicator of flexibility and

disorder. In future CASP cryo-EM assessments it would be useful to encourage the authors to provide unsharpened maps, and even half maps for further assessments.

For the first time in CASP history, RNA structures were provided as targets and the majority of them had associated cryo-EM density maps. Compared with the proteins, these RNA maps had much lower resolutions. Indeed, in some maps such as those of R1156, pitches of helices were not always visible. Local fit-to-map scores, such as the newly developed SMOCr, can aid the assessment of RNA models in these challenging resolutions. Here, this local fit analysis indicated that many secondary structures and important geometric features can be accurately predicted. Furthermore, we showed that *in silico* models can, after further refinement, offer plausible models that better reflect the experimental maps even at low resolutions. However, at such low resolutions, it is possible for many alternative structures to fit the density with equal likelihood. Due to both the known flexibility of the RNA molecules and the heterogeneity of the experimental maps, ensembles of models are arguably a more accurate way to describe the underlying experimental data.^{8,37}

Despite the overall quality of predictions, some reorientation of domains and secondary structure elements was often required, particularly for RNA models. The multistage pipeline presented offers an approach to fitting and refinement of structural models into cryo-EM maps at a variety of resolutions. The use of progressively smaller rigid-bodies has been shown to aid the fitting of models that require large conformational changes.¹⁹ However, if the models contain topological errors or significant misplacements of elements even such a detailed approach will fail.

As mentioned above, in CASP15 there were two RNA-protein complexes (RT1189, RT1190). The predictions associated with these targets were not refined due to poor accuracy.⁸ Given the current progress in the structure prediction field, we expect further improvement on this front in future CASPs.

CryoEM has been an important method for elucidating large atomic structures, albeit often at a lower resolution than crystallographic experiments. This CASP15 for example, the largest monomeric structure in the history of CASPs, T1169, was a cryo-EM target. Moreover, cryo-EM experiments are now not just capturing large molecules but often achieving atomic levels of detail. In CASP15, focussed maps for the target H1114 reached an astonishing resolution of 1.52 Å. While at such resolutions, computational models are not required for model building, high-resolution data offers an opportunity to assess accuracy at an even finer level. Using the SMOC score separately for backbone (SMOCb) and side-chains (SMOCs), allowed us to show that while the overall backbone geometry of H1114 predictions was well modeled, side-chain orientations did not always agree with the experimental map. Given the progress in both protein structure prediction and cryo-EM fields, we foresee such analyses becoming more routine in the future.

AUTHOR CONTRIBUTIONS

Thomas Mulvaney: Conceptualization; investigation; software; methodology; data curation; writing – original draft; writing – review and

editing; visualization; project administration. **Rachael C. Kretsch:** Conceptualization; methodology; software; investigation; writing – original draft; writing – review and editing; visualization; data curation. **Luc Elliott:** Conceptualization; investigation; writing – original draft; writing – review and editing; methodology; software; data curation; visualization. **Joseph G. Beton:** Conceptualization; software; investigation; writing – original draft; writing – review and editing; visualization; methodology; data curation. **Andriy Kryshchak:** Resources; supervision; project administration; data curation; writing – original draft; writing – review and editing; funding acquisition; investigation. **Daniel J. Rigden:** Resources; supervision; project administration; writing – review and editing; writing – original draft; methodology; funding acquisition; investigation. **Rhiju Das:** Funding acquisition; writing – original draft; writing – review and editing; project administration; resources; supervision; methodology; investigation. **Maya Topf:** Writing – original draft; funding acquisition; writing – review and editing; project administration; resources; supervision; methodology; investigation.

ACKNOWLEDGMENTS

This research was supported by the cooperation of Leibniz Institute of Virology (as part of Leibniz ScienceCampus InterACT, funded by the BWFGB Hamburg and the Leibniz Association) and by the Strategic Incentive Program of LIV and the Landesforschungsförderung Hamburg (Hamburg-X) (to T.M, J.G.B, and M.T). The PhD studentship of L.E. is co-funded by the Collaborative Computational Project for Electron cryo-Microscopy (CCP-EM). This research was supported by the US National Institute of General Medical Sciences (NIGMS, National Institutes of Health) grants R01 GM100482 (A.K.) and R35 GM122579 (R.D.); Stanford Bio-X (to R.D. and R.C.K.); and the Howard Hughes Medical Institute (HHMI, to R.D.). This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Thomas Mulvaney  <https://orcid.org/0000-0002-4373-6160>

Rachael C. Kretsch  <https://orcid.org/0000-0002-6935-518X>

Luc Elliott  <https://orcid.org/0009-0002-0181-4041>

Joseph G. Beton  <https://orcid.org/0000-0001-7499-3867>

Andriy Kryshchak  <https://orcid.org/0000-0001-5066-7178>

Daniel J. Rigden  <https://orcid.org/0000-0002-7565-8937>

Rhiju Das  <https://orcid.org/0000-0001-7497-0972>

Maya Topf  <https://orcid.org/0000-0002-8185-1215>

REFERENCES

- Kryshtafovych A, Malhotra S, Monastyrskyy B, et al. Cryo-electron microscopy targets in CASP13: overview and evaluation of results. *Proteins*. 2019;87(12):1128-1140. doi:10.1002/prot.25817
- Cragolini T, Kryshtafovych A, Topf M. Cryo-EM targets in CASP14. *Proteins*. 2021;89(12):1949-1958. doi:10.1002/prot.26216
- Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235
- Ozden B, Kryshtafovych A, Karaca E. Assessment of the CASP14 assembly predictions. *Proteins*. 2021;89(12):1787-1799. doi:10.1002/prot.26199
- Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. Published online March 10, 2022. doi:10.1101/2021.10.04.463034
- Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 2021;373(6557):871-876. doi:10.1126/science.abj8754
- Gao M, Nakajima An D, Parks JM, Skolnick J. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat Commun*. 2022;13(1):1744. doi:10.1038/s41467-022-29394-2
- Das R, Kretsch RC, Simpkin AJ, et al. Assessment of three-dimensional RNA structure prediction in CASP15. *Protein*. 2023;91(12):1747-1770. doi:10.1002/prot.26602
- Pettersen EF, Goddard TD, Huang CC, et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci*. 2021;30(1):70-82. doi:10.1002/pro.3943
- Vagin A, Teplyakov A. MOLREP: an automated program for molecular replacement. *J Appl Cryst*. 1997;30(6):1022-1025. doi:10.1107/S0021889897006766
- Vagin AA, Isupov MN. Spherically averaged phased translation function and its application to the search for molecules and fragments in electron-density maps. *Acta Crystallogr D Biol Crystallogr*. 2001;57(Pt 10):1451-1456. doi:10.1107/s0907444901012409
- van Zundert G, Bonvin A. Fast and sensitive rigid-body fitting into cryo-EM density maps with PowerFit. *AIMS Biophys*. 2015;2(2):73-87. doi:10.3934/biophys.2015.2.73
- Kryshtafovych A, Rigden DJ. To split or not to split: CASP15 targets and their processing into tertiary structure evaluation units. *Proteins*. 2023;91(12):1558-1570. doi:10.1002/prot.26533
- Simpkin AJ, Mesdaghi S, Sánchez Rodríguez F, et al. Tertiary structure assessment at CASP15. *Proteins*. 2023;91(12):1616-1635. doi:10.1002/prot.26593
- Lafita A, Bliven S, Kryshtafovych A, et al. Assessment of protein assembly prediction in CASP12. *Proteins*. 2018;86(suppl 1):247-256. doi:10.1002/prot.25408
- Guzenko D, Lafita A, Monastyrskyy B, Kryshtafovych A, Duarte JM. Assessment of protein assembly prediction in CASP13. *Proteins*. 2019;87(12):1190-1199. doi:10.1002/prot.25795
- Malhotra S, Mulvaney T, Cragolini T, et al. RIBFIND2: identifying rigid bodies in protein and nucleic acid structures. *Nucleic Acids Res*. 2023;51(18):9567-9575. doi:10.1093/nar/gkad721
- Chou FC, Richardson JS, Das R. ERRASER, a powerful new system for correcting RNA models. *Computl Crystallogr Newslett*. 2013;10:74-76. doi:10.1038/nmeth.2262
- Joseph AP, Malhotra S, Burnley T, et al. Refinement of atomic models in high resolution EM reconstructions using flex-EM and local assessment. *Methods*. 2016;100:42-49. doi:10.1016/j.ymeth.2016.03.007
- Williams CJ, Headd JJ, Moriarty NW, et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci*. 2018;27(1):293-315. doi:10.1002/pro.3330
- Adams PD, Afonine PV, Bunkóczi G, et al. PHENIX: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr*. 2010;66(Pt 2):213-221. doi:10.1107/S0907444909052925
- Richardson JS, Schneider B, Murray LW, et al. RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA ontology consortium contribution). *RNA*. 2008;14(3):465-481. doi:10.1261/ma.657708
- Sanchez-Garcia R, Gomez-Blanco J, Cuervo A, Carazo JM, Sorzano COS, Vargas J. DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *Commun Biol*. 2021;4(1):874. doi:10.1038/s42003-021-02399-1
- Gambelli L, McLaren M, Connors R, et al. Structure of the two-component S-layer of the archaeon *Sulfolobus acidocaldarius*. *bioRxiv*. Published online October 7, 2022. doi:10.1101/2022.10.07.511299
- van den Berg B, Silale A, Baslé A, Brandner AF, Mader SL, Khalid S. Structural basis for host recognition and superinfection exclusion by bacteriophage T5. *Proc Natl Acad Sci U S A*. 2022;119(42):e2211672119. doi:10.1073/pnas.2211672119
- Casañal A, Lohkamp B, Emsley P. Current developments in coot for macromolecular model building of electron cryo-microscopy and crystallographic data. *Protein Sci*. 2020;29(4):1069-1078. doi:10.1002/pro.3791
- Croll TI. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr D Struct Biol*. 2018;74(Pt 6):519-530. doi:10.1107/S2059798318002425
- Grinter R, Kropp A, Venugopal H, et al. Structural basis for bacterial energy extraction from atmospheric hydrogen. *Nature*. 2023;615(7952):541-547. doi:10.1038/s41586-023-05781-7
- Liu S, Xia X, Calvo E, Zhou ZH. Native structure of mosquito salivary protein uncovers domains relevant to pathogen transmission. *Nat Commun*. 2023;14(1):899. doi:10.1038/s41467-023-36577-y
- McRae EKS, Rasmussen HØ, Liu J, et al. Structure, folding and flexibility of co-transcriptional RNA origami. *Nat Nanotechnol*. 2023;18(7):808-817. doi:10.1038/s41565-023-01321-6
- Sampedro Vallina N, McRae EKS, Hansen BK, Boussebayle A, Andersen ES. RNA origami scaffolds facilitate cryo-EM characterization of a broccoli-pepper aptamer FRET pair. *Nucleic Acids Res*. 2023;51(9):4613-4624. doi:10.1093/nar/gkad224
- Miao Z, Tidu A, Eriani G, Martin F. Secondary structure of the SARS-CoV-2 5'-UTR. *RNA Biol*. 2021;18(4):447-456. doi:10.1080/15476286.2020.1814556
- Yang D, Leibowitz JL. The structure and functions of coronavirus genomic 3' and 5' ends. *Virus Res*. 2015;206:120-133. doi:10.1016/j.virusres.2015.02.025
- Maclean B, Marco S, Brittany RM. Lessons learned and yet-to-be learned on the importance of RNA structure in SARS-CoV-2 replication. *Microbiol Mol Biol Rev*. 2022;86(3):21. doi:10.1128/mmb.00057-21
- Cragolini T, Beton J, Topf M. Cryo-EM structure and B-factor refinement with ensemble representation. *bioRxiv*. Published online June 9, 2022. doi:10.1101/2022.06.08.495259
- Jakobi AJ, Wilmanns M, Sachse C. Model-based local density sharpening of cryo-EM maps. *Elife*. 2017;6:6. doi:10.7554/eLife.27131
- Beton JG, Cragolini T, Kaleel M, Mulvaney T, Sweeney A, Topf M. Integrating model simulation tools and cryo-electron microscopy. *Wiley Interdiscip Rev Comput Mol Sci*. 2022;13:e1642. doi:10.1002/wcms.1642

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mulvaney T, Kretsch RC, Elliott L, et al. CASP15 cryo-EM protein and RNA targets: Refinement and analysis using experimental maps. *Proteins*. 2023;91(12):1935-1951. doi:10.1002/prot.26644

4

Discussion

The three papers presented in this thesis describe an approach for fitting atomic models to cryo-EM data and at the same time estimating the resolution of the atoms. However, there are a number of issues which were not well addressed.

4.1 Overfitting and force determination

In (Beton*, Mulvaney* et al. 2024) a new Gaussian Mixture Model approach to flexible fitting was developed and tested on a large set of deposited models (229 in total) with experimental maps reconstructed at 2.1 to 4.9Å. These were compared to the models obtained from the CERES database containing Phenix-refined models.

The results of the TEMPy-ReFF benchmark, when broken down by resolution, were somewhat surprising. At high resolutions, the TEMPy-ReFF refined models shared the same distribution of model-map cross-correlation scores as their CERES counterparts, but better (lower) MolProbity scores. At lower resolutions, the cross-correlation of TEMPy-ReFF models were better, but the MolProbity scores became worse. It was not expected that the GMM based methods would produce models with better geometry than Phenix with an equivalent fit to the experimental data at high resolutions. Rather, we expected it to be more accurate at lower resolutions where the GMM should better model the underlying experimental data and thus be less likely to induce distortions. The cross-correlation with the experimental data was better at lower resolutions, suggesting the GMM does a good job at fitting the experimental density but is possibly overfitting.

The strength of the TEMPy-ReFF GMM potential (which is user tuneable) was set to 10^5 across all resolution ranges in the benchmark. It is highly likely that the strength was not always appropriate, and should potentially be reduced at lower resolutions to prevent overfitting. The earlier discussed approaches to estimate biasing strength and avoid overfitting (see section 1.2.4) of (DiMaio, J. Zhang, et al. 2013; Igaev et al. 2019) are currently being assessed.

4.2 Membranes and other unmodelled regions

Cryo-EM is a particularly relevant technique when it comes to studying membrane proteins. To study these proteins, they are embedded in lipid nanodiscs which approximate cellular membranes. The membranes themselves produce a signal which is significantly higher than background. In GMMs, Gaussians try to fit all of the observed data. The downsides of this, were observed in some TEMPy-ReFF runs, where it was noted that membrane proximal regions would sometimes move into the membrane associated density. Currently, TEMPy-ReFF models the cryo-EM data with a Gaussian per atom, and a single background component. Establishing a second component, which takes into account this membrane density, would potentially help prevent atomic Gaussians from trying to expand into these regions.

It is likely that some of the lower than expected MolProbity scores which were observed in the benchmark of TEMPy-ReFF at lower resolutions, were due to membrane associated density. It is likely the lower resolution examples in the CERES benchmark contained more membrane proteins than the high resolution data set, as these frequently yield lower resolution reconstructions.

Given the importance of cryo-EM to membrane protein elucidation efforts, handling membrane density more robustly within the GMM paradigm described is something I am actively working on.

4.3 Interpreting TEMPy-ReFF “B-factors”

The local resolution as determined by TEMPy-ReFF is often equated to a Bfactor or atomic displacement factor. In general, it is probably more accurate to think of the value as a measure of modelling uncertainty. This is different from the true uncertainty in atomic positions, due to say molecular motion. The reason for this is a reconstruction captures not just the atomic motions but also corruptions due to spherical aberration. Incoherency in the electron source and alignment errors also lead to Gaussian blurring. Removing all of these global sources of image blurring (for example by sharpening) in theory will lead to a reconstruction where local heterogeneity due to actual atomic motion is all that remains. However, sharpening is never so accurate, in part because the global sources of blurring can only be approximately modelled.

4.4 Supported atomic models

TEMPy-ReFF is currently limited in the types of atomic models it can refine. In particular, chemical modification of amino acids and nucleic acids (a situation which is not uncommon and often pertinent to the underlying biology) is not currently supported. Instead, these additional groups are removed from structures before refinement. In some respects, the situation should be easy to rectify as many common chemical modifications such as glycosylation and methylation are handled by modern biomolecular force fields. In practice, these modifications need to be well defined in the PDB files of the initial structures in order for the force field to recognise and correctly parameterise them. Poorly defined chemical modifications, which for example might have atoms missing or have some ambiguity, are difficult

to handle automatically. It is unclear what the best solution is, but it will likely require some prompting of users for input.

4.5 Geometry refinement

A large assumption is that starting models have reasonable geometry and that this local geometry should be preserved or is close enough to an ideal geometry that MD force fields can further improve it. This is frequently the case for proteins. Where large conformational changes are required, the hierarchical protocol based on RIBFIND2 has been shown in thesis to work well. However, it is often the smaller details such as rotamers and backbone geometry which are difficult to correct as rotations and flips involve steep energy barriers. Simulated annealing (Adams et al. 2010; Topf et al. 2008) and search heuristics (Oldfield 2001) could be employed to better explore this conformational space.

The overall fit to the density and orientation of secondary structure elements of RNA predictions from CASP15 could be improved using the hierarchical refinement protocol of TEMPy-ReFF, but the finer details of RNA secondary structure could only be improved by employing specialised refinement software such as ERRASER (Chou et al. 2013). A simulated annealing based approach might help here, or additional restraints on backbone suites as per QRNA (Stasiewicz et al. 2019), but for the time being I recommend the protocol outlined in the (Mulvaney et al. 2023) paper.

A major limitation to RIBFIND2 in its current form is while it can produce rigid body clusters from atomic structures of protein-nucleic acid complexes, these rigid body clusters are generated separately: one for proteins, and one for nucleic acids. In the case of complexes, this can be sub-optimal, as it leads to independent fitting of structural regions which may need to be considered as a rigid whole. Future versions of RIBFIND2 should address this limitation by additionally defining a protein-nucleic acid interaction distance.

4.6 Modelling of atomic charges

The approach established in this paper is able to optimize the positions and Bfactors of atoms under a model where all atoms are effectively neutral. This is because we (and others) use an approximation (eq 1.6) of the more accurate scattering factors of (eq. 1.5). Many atoms, are known to possess partial charges and to thus deviate from neutral scattering curves. Indeed, this is observed in cryo-EM maps and explored in reviews (Bick et al. 2024; Marques et al. 2019).

While our GMM approach produces Bfactors which are generally in good agreement with other methods for determining local resolution, it is probable that they are less accurate for negative atoms. Here, I would hypothesise that, the Bfactors end up being higher to accommodate for the lack of density. Good computational models should be able to explain the experimental data. TEMPy-ReFF and other tools which take account of local resolution are a step in the right direction, but charge is something that also needs to be accounted for.

Considering that RNA molecules have negatively charged phosphate backbones and that their tertiary structure is dependent on positively charged metal ions, accurate simulations of these charge effects have been shown to be important (J. Wang, Z. Liu, et al. 2018) and not accounting for them has been the source of modelling errors (J. Wang, Natchiar, et al. 2021).

The more accurate the simulated cryo-EM map is the better we are able to determine discrepancies between model and observed data. Simulated maps, are the basis for many scores in the TEMPy package. Being able to detect discrepancies between atomic model and observed data due to unexpected charge distribution may also be important for understanding ligand binding sites.

Some attempts have been made to manually adjust the charge to better model the data in Phenix (Hryc et al. 2017). Here, as partial charges cannot be set, the atomic occupancy was reduced or made negative, to account for negative density.

It is clear that accounting for this in an automated fashion will lead to a better description of the observed density and a more nuanced interpretation of the atomic model and the underlying chemistry.

4.6.1 Is flexible fitting still the future?

During the course of this PhD, a method has emerged for biasing AlphaFold during the structure prediction stage itself, using restraints derived from cross-linking experiments (Stahl et al. 2023). It is foreseeable, that incorporating low resolution cryo-EM data into AlphaFold-like structure prediction methods could lead to an alternative, more streamlined approach to model building. Figuring out a general strategy for doing so would be useful for other low resolution experimental methods such as small angle X-ray scattering (SAXS).

4.7 Conclusion

In this thesis, a new biasing potential for flexible fitting of structural models in cryo-EM maps is described (Beton*, Mulvaney* et al. 2024). Its accuracy was compared against 229 models previously refined with Phenix. With the help of rigid bodies (Malhotra*, Mulvaney* et al), this culminated in the refinement of a diverse set of proteins and RNA structure predictions in the CASP15 modelling challenge (Mulvaney et al. 2023). Overall, when compared to Phenix, the method produced models with better geometry and fit to density at high resolutions, but performed slightly worse at lower resolutions.

RNA predictions from CASP15 often deviated far from the experimental model, yet these could still be successfully fit to the maps with good agreement with the experimental model.

Some of the methods reported by experimentalists for building the target RNA models for CASP15 involved complex protocols. One such example included: initial manual model building in ChimeraX starting from crystal structures, flexible fitting with MDFF, manual refinement using ISOLDE, Phenix real space refinement, and finally RNA geometry adjustments with QRNA (Sampedro Vallina et al. 2023).

In my CASP15 paper, I offer a glimpse into a more automated future, where RNA structure predictions can be flexibly fit to experimental data with similar accuracy to hand-tuned models using a simple pipeline.

Handling the diverse array of chemical modifications which are common amongst biological molecules is an important next step in making the software more applicable.

Finally, charge is an overlooked aspect of modelling including in the software presented in this thesis. An interesting direction for the future would be to either explicitly model any partial charges during map simulation, or to attempt to determine them along with local resolution during refinement.

References

- Adams, P. D., P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart (Feb. 2010). “PHENIX: a comprehensive Python-based system for macromolecular structure solution”. en. In: *Acta crystallographica. Section D, Biological crystallography* 66.Pt 2, pp. 213–221. ISSN: 0907-4449,1399-0047. DOI: 10.1107/S0907444909052925. URL: <http://dx.doi.org/10.1107/S0907444909052925>.
- Address, K. J., J. P. Basilion, R. D. Klausner, T. A. Rouault, and A. Pardi (Nov. 1997). “Structure and dynamics of the iron responsive element RNA: implications for binding of the RNA by iron regulatory binding proteins”. en. In: *Journal of molecular biology* 274.1, pp. 72–83. ISSN: 0022-2836,1089-8638. DOI: 10.1006/jmbi.1997.1377. URL: <http://dx.doi.org/10.1006/jmbi.1997.1377>.
- Adrian, M., J. Dubochet, J. Lepault, and A. W. McDowell (1984). “Cryo-electron microscopy of viruses”. en. In: *Nature* 308.5954, pp. 32–36. ISSN: 0028-0836,1476-4687. DOI: 10.1038/308032a0. URL: <http://dx.doi.org/10.1038/308032a0>.
- Andersen, H. C. (Feb. 1980). “Molecular dynamics simulations at constant pressure and/or temperature”. en. In: *The Journal of chemical physics* 72.4, pp. 2384–2393. ISSN: 0021-9606,1089-7690. DOI: 10.1063/1.439486. URL: <http://dx.doi.org/10.1063/1.439486>.
- Bekker, H., H. J. C. Berendsen, E. J. Dijkstra, S. Achterop, R. Van Drunen, D. Van Der Spoel, and H. Sijbers (1993). “Gromacs: parallel computer molecular dynamics simulations”. In: *Physics computing 92*. Edited R. de Groot J. Nadrchal. World Scientific, pp. 252–256.
- Beton, J. G., T. Cragolini, M. Kaleel, T. **Mulvaney**, A. Sweeney, and M. Topf (Nov. 2022). “Integrating model simulation tools and cryoelectron microscopy”. en. In: *Wiley interdisciplinary reviews. Computational molecular science*. ISSN: 1759-0876, 1759-0884. DOI: 10.1002/wcms.1642. URL: <https://onlinelibrary.wiley.com/doi/10.1002/wcms.1642>.
- Beton*, J. G., T. **Mulvaney***, T. Cragolini, and M. Topf (Jan. 2024). “Cryo-EM structure and B-factor refinement with ensemble representation”. en. In: *Nature communications* 15.1, p. 444. ISSN: 2041-1723. DOI: 10.1038/s41467-023-44593-1. URL: <http://dx.doi.org/10.1038/s41467-023-44593-1>.
- Bick, T., P. M. Dominiak, and P. Wendler (2024). “Exploiting the full potential of cryo-EM maps”. en. In: *BBA advances* 5.100113, p. 100113. ISSN: 2667-1603. DOI: 10.1016/j.bbadv.2024.100113. URL: <http://dx.doi.org/10.1016/j.bbadv.2024.100113>.
- Bonilla, S. L., Q. Vicens, and J. S. Kieft (Aug. 2022). “Cryo-EM reveals an entangled kinetic trap in the folding of a catalytic RNA”. en. In: *Science advances*

- 8.34, eabq4144. ISSN: 2375-2548. DOI: 10.1126/sciadv.abq4144. URL: <http://dx.doi.org/10.1126/sciadv.abq4144>.
- Böttcher, B., S. A. Wynne, and R. A. Crowther (Mar. 1997). “Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy”. en. In: *Nature* 386.6620, pp. 88–91. ISSN: 0028-0836,1476-4687. DOI: 10.1038/386088a0. URL: <http://dx.doi.org/10.1038/386088a0>.
- Brilot, A. F., J. Z. Chen, A. Cheng, J. Pan, S. C. Harrison, C. S. Potter, B. Carragher, R. Henderson, and N. Grigorieff (Mar. 2012). “Beam-induced motion of vitrified specimen on holey carbon film”. en. In: *Journal of structural biology* 177.3, pp. 630–637. ISSN: 1047-8477,1095-8657. DOI: 10.1016/j.jsb.2012.02.003. URL: <http://dx.doi.org/10.1016/j.jsb.2012.02.003>.
- Brooks, B. R., C. L. Brooks, D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, L. Caffisch, Q. Caves, R. Cui, M. Dinner, S. Feig, J. Fischer, M. Gao, W. Hodoscek, K. Im, T. Kuczera, J. Lazaridis, V. Ma, E. Ovchinnikov, R. W. Paci, C. B. Pastor, J. Z. Post, M. Pu, B. Schaefer, R. M. Tidor, H. L. Venable, X. Woodcock, W. Wu, D. M. Yang, and M. York (2009). “Karplus: CHARMM: Biomolecular simulation Program”. In: *J. Comp. Chem* 30, pp. 1545–1615.
- Brünger, A. T. (Jan. 1992). “Free R value: a novel statistical quantity for assessing the accuracy of crystal structures”. en. In: *Nature* 355.6359, pp. 472–475. ISSN: 0028-0836,1476-4687. DOI: 10.1038/355472a0. URL: <http://dx.doi.org/10.1038/355472a0>.
- Brünger, A. T., P. D. Adams, and L. M. Rice (Mar. 1997). “New applications of simulated annealing in X-ray crystallography and solution NMR”. en. In: *Structure (London, England: 1993)* 5.3, pp. 325–336. ISSN: 0969-2126,1878-4186. DOI: 10.1016/s0969-2126(97)00190-1. URL: [http://dx.doi.org/10.1016/s0969-2126\(97\)00190-1](http://dx.doi.org/10.1016/s0969-2126(97)00190-1).
- Brünger, A. T., G. M. Clore, A. M. Gronenborn, and M. Karplus (June 1986). “Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 83.11, pp. 3801–3805. ISSN: 0027-8424,1091-6490. DOI: 10.1073/pnas.83.11.3801. URL: <http://dx.doi.org/10.1073/pnas.83.11.3801>.
- Brünger, A. T., J. Kuriyan, and M. Karplus (Jan. 1987). “Crystallographic R factor refinement by molecular dynamics”. en. In: *Science (New York, N.Y.)* 235.4787, pp. 458–460. ISSN: 0036-8075,1095-9203. DOI: 10.1126/science.235.4787.458. URL: <http://dx.doi.org/10.1126/science.235.4787.458>.
- Burch, B. D., C. Garrido, and D. M. Margolis (Jan. 2017). “Detection of human immunodeficiency virus RNAs in living cells using Spinach RNA aptamers”. en. In: *Virus research* 228, pp. 141–146. ISSN: 0168-1702,1872-7492. DOI: 10.1016/j.virusres.2016.11.031. URL: <http://dx.doi.org/10.1016/j.virusres.2016.11.031>.
- Butcher, S. E. and A. M. Pyle (Dec. 2011). “The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks”. en. In: *Accounts of chemical research* 44.12, pp. 1302–1311. ISSN: 0001-4842,1520-4898. DOI: 10.1021/ar200098t. URL: <http://dx.doi.org/10.1021/ar200098t>.
- Cardone, G., J. B. Heymann, and A. C. Steven (Nov. 2013). “One number does not fit all: Mapping local variations in resolution in cryo-EM reconstructions”. en. In: *Journal of structural biology* 184.2, pp. 226–236. ISSN: 1047-8477,1095-8657.

- DOI: 10.1016/j.jsb.2013.08.002. URL: <http://dx.doi.org/10.1016/j.jsb.2013.08.002>.
- Casañal, A., B. Lohkamp, and P. Emsley (Apr. 2020). “Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data”. en. In: *Protein science: a publication of the Protein Society* 29.4, pp. 1069–1078. ISSN: 0961-8368,1469-896X. DOI: 10.1002/pro.3791. URL: <http://dx.doi.org/10.1002/pro.3791>.
- Case, D., H. M. Aktulga, K. Belfon, I. Y. Ben-Shalom, J. T. Berryman, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, G. Iii, V. W. D. Cisneros, T. Cruzeiro, N. Darden, M. Forouzesh, G. Ghazimirsaeed, T. Giambau, M. K. Giese, H. Gilson, W. Gohlke, J. Goetz, Z. Harris, S. Huang, S. Izadi, K. Izmailov, M. C. Kasavajhala, T. Kovalenko, T. S. Kurtzman, P. Lee, Z. Li, C. Li, J. Lin, T. Liu, R. Luchko, M. Luo, M. Machado, K. M. Manathunga, Y. Merz, O. Miao, G. Mikhailovskii, H. Monard, K. Nguyen, F. Onufriev, S. Pan, D. R. Rahnamoun, C. Roitberg, S. Sagui, J. Shajan, C. L. Shen, N. R. Simmerling, J. Skrynnikov, J. Smith, R. C. Swails, J. Walker, J. Wang, X. Wang, Y. Wu, Y. Wu, Y. Xiong, D. M. Xue, C. York, Q. Zhao, and P. Zhu (2024). “Amber 2024”. In.
- Chou, F.-C., P. Sripakdeevong, S. M. Dibrov, T. Hermann, and R. Das (Jan. 2013). “Correcting pervasive errors in RNA crystallography through enumerative structure prediction”. en. In: *Nature methods* 10.1, pp. 74–76. ISSN: 1548-7091,1548-7105. DOI: 10.1038/nmeth.2262. URL: <http://dx.doi.org/10.1038/nmeth.2262>.
- Chudakov, D. M., M. V. Matz, S. Lukyanov, and K. A. Lukyanov (July 2010). “Fluorescent proteins and their applications in imaging living cells and tissues”. en. In: *Physiological reviews* 90.3, pp. 1103–1163. ISSN: 0031-9333,1522-1210. DOI: 10.1152/physrev.00038.2009. URL: <http://dx.doi.org/10.1152/physrev.00038.2009>.
- Cragolini, T., H. Sahota, A. P. Joseph, et al. (2021). “TEMPy2: a Python library with improved 3D electron microscopy densityfitting and validation workflows”. In: Section D: Structural URL: <https://onlinelibrary.wiley.com/doi/abs/10.1107/S2059798320014928>.
- Cragolini, T., A. Kryshtafovych, and M. Topf (Dec. 2021). “Cryo-EM targets in CASP14”. en. In: *Proteins* 89.12, pp. 1949–1958. ISSN: 0887-3585,1097-0134. DOI: 10.1002/prot.26216. URL: <http://dx.doi.org/10.1002/prot.26216>.
- Crick, F. (Aug. 1970). “Central dogma of molecular biology”. en. In: *Nature* 227.5258, pp. 561–563. ISSN: 0028-0836,1476-4687. DOI: 10.1038/227561a0. URL: <http://dx.doi.org/10.1038/227561a0>.
- Croll, T. I. (June 2018). “ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps”. en. In: *Acta crystallographica. Section D, Structural biology* 74.Pt 6, pp. 519–530. ISSN: 2059-7983. DOI: 10.1107/S2059798318002425. URL: <http://dx.doi.org/10.1107/S2059798318002425>.
- Croll, T. I. and R. J. Read (Apr. 2021). “Adaptive Cartesian and torsional restraints for interactive model rebuilding”. en. In: *Acta crystallographica. Section D, Structural biology* 77.Pt 4, pp. 438–446. ISSN: 2059-7983. DOI: 10.1107/S2059798321001145. URL: <http://dx.doi.org/10.1107/S2059798321001145>.
- Dahmani, Z. L., A. L. Scott, C. Vénien-Bryan, D. Perahia, and M. G. S. Costa (July 2024). “MDFF_NM: Improved Molecular Dynamics Flexible Fitting into Cryo-EM Density Maps with a Multireplica Normal Mode-Based Search”. en.

- In: *Journal of chemical information and modeling* 64.13, pp. 5151–5160. ISSN: 1549-9596, 1549-960X. DOI: 10.1021/acs.jcim.3c02007. URL: <http://dx.doi.org/10.1021/acs.jcim.3c02007>.
- DiIorio, M. C. and A. W. Kulczyk (Jan. 2022). “A robust single-particle cryo-electron microscopy (cryo-EM) processing workflow with cryoSPARC, RELION, and Scipion”. en. In: *Journal of visualized experiments: JoVE* 179. ISSN: 1940-087X. DOI: 10.3791/63387. URL: <http://dx.doi.org/10.3791/63387>.
- DiMaio, F., Y. Song, X. Li, M. J. Brunner, C. Xu, V. Conticello, E. Egelman, T. Marlovits, Y. Cheng, and D. Baker (Apr. 2015). “Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement”. en. In: *Nature methods* 12.4, pp. 361–365. ISSN: 1548-7091,1548-7105. DOI: 10.1038/nmeth.3286. URL: <http://dx.doi.org/10.1038/nmeth.3286>.
- DiMaio, F., J. Zhang, W. Chiu, and D. Baker (June 2013). “Cryo-EM model validation using independent map reconstructions”. en. In: *Protein science: a publication of the Protein Society* 22.6, pp. 865–868. ISSN: 0961-8368,1469-896X. DOI: 10.1002/pro.2267. URL: <http://dx.doi.org/10.1002/pro.2267>.
- Ding, J., J. C. Deme, J. R. Stagno, P. Yu, S. M. Lea, and Y.-X. Wang (Oct. 2023). “Capturing heterogeneous conformers of cobalamin riboswitch by cryo-EM”. en. In: *Nucleic acids research* 51.18, pp. 9952–9960. ISSN: 0305-1048,1362-4962. DOI: 10.1093/nar/gkad651. URL: <http://dx.doi.org/10.1093/nar/gkad651>.
- Doyle, P. A. and P. S. Turner (May 1968). “Relativistic HartreeFock X-ray and electron scattering factors”. en. In: *Acta crystallographica. Section A, Crystal physics, diffraction, theoretical and general crystallography* 24.3, pp. 390–397. ISSN: 0567-7394,1600-8596. DOI: 10.1107/s0567739468000756. URL: <https://journals.iucr.org/paper?a05916>.
- Eastman, P., J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande (July 2017). “OpenMM 7: Rapid development of high performance algorithms for molecular dynamics”. en. In: *PLoS computational biology* 13.7, e1005659. ISSN: 1553-734X,1553-7358. DOI: 10.1371/journal.pcbi.1005659. URL: <http://dx.doi.org/10.1371/journal.pcbi.1005659>.
- Erdmann, P. S., Z. Hou, S. Klumpe, S. Khavnekar, F. Beck, F. Wilfling, J. M. Plitzko, and W. Baumeister (Sept. 2021). “In situ cryo-electron tomography reveals gradient organization of ribosome biogenesis in intact nucleoli”. en. In: *Nature communications* 12.1, p. 5364. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25413-w. URL: <https://www.nature.com/articles/s41467-021-25413-w>.
- Evans, R., M. O’Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. ídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstern, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, and D. Hassabis (Mar. 2022). “Protein complex prediction with AlphaFold-Multimer”. en. In: *bioRxiv*, p. 2021.10.04.463034. DOI: 10.1101/2021.10.04.463034. URL: <https://www.biorxiv.org/content/biorxiv/early/2022/03/10/2021.10.04.463034>.
- Fernandez-Gimenez, E., J. M. Carazo, and C. O. S. Sorzano (Dec. 2023). “Local defocus estimation in single particle analysis in cryo-electron microscopy”. en. In: *Journal of structural biology* 215.4, p. 108030. ISSN: 1047-8477,1095-8657. DOI: 10.1016/j.jsb.2023.108030. URL: <http://dx.doi.org/10.1016/j.jsb.2023.108030>.

- Filonov, G. S., J. D. Moon, N. Svensen, and S. R. Jaffrey (2014). “Broccoli: rapid selection RNA mimic green fluorescent protein fluorescence- based selection directed evolution”. In: *J Am Chem Soc* 136, pp. 16299–16308.
- Fischer, N., A. L. Konevega, W. Wintermeyer, M. V. Rodnina, and H. Stark (July 2010). “Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy”. en. In: *Nature* 466.7304, pp. 329–333. ISSN: 0028-0836,1476-4687. DOI: 10.1038/nature09206. URL: <http://dx.doi.org/10.1038/nature09206>.
- Fiser, A. and A. Sali (2003). “Modeller: generation and refinement of homology-based protein structure models”. en. In: *Methods in enzymology* 374, pp. 461–491. ISSN: 0076-6879,1557-7988. DOI: 10.1016/S0076-6879(03)74020-8. URL: [http://dx.doi.org/10.1016/S0076-6879\(03\)74020-8](http://dx.doi.org/10.1016/S0076-6879(03)74020-8).
- Glaeser, R. M., G. McMullan, A. R. Faruqi, and R. Henderson (Jan. 2011). “Images of paraffin monolayer crystals with perfect contrast: minimization of beam-induced specimen motion”. en. In: *Ultramicroscopy* 111.2, pp. 90–100. ISSN: 0304-3991,1879-2723. DOI: 10.1016/j.ultramic.2010.10.010. URL: <http://dx.doi.org/10.1016/j.ultramic.2010.10.010>.
- “Introduction and overview” (May 2021). In: *Single-particle Cryo-EM of Biological Macromolecules*. Ed. by R. M. Glaeser, E. Nogales, and W. Chiu. IOP Publishing. ISBN: 9780750330398. DOI: 10.1088/978-0-7503-3039-8ch1. URL: <http://dx.doi.org/10.1088/978-0-7503-3039-8ch1>.
- Go, N. (1983). “Theoretical studies of protein folding”. en. In: *Annual review of biophysics and bioengineering* 12.1, pp. 183–210. ISSN: 0084-6589,2327-9885. DOI: 10.1146/annurev.bb.12.060183.001151. URL: <http://dx.doi.org/10.1146/annurev.bb.12.060183.001151>.
- Grant, T. and N. Grigorieff (May 2015). “Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6”. en. In: *eLife* 4, e06980. ISSN: 2050-084X. DOI: 10.7554/eLife.06980. URL: <http://dx.doi.org/10.7554/eLife.06980>.
- Grigorieff, N. (June 2016). “Frealign: An exploratory tool for single-particle cryo-EM”. en. In: *Methods in enzymology* 579, pp. 191–226. ISSN: 0076-6879,1557-7988. DOI: 10.1016/bs.mie.2016.04.013. URL: <http://dx.doi.org/10.1016/bs.mie.2016.04.013>.
- Grubisic, I., M. N. Shokhirev, M. Orzechowski, O. Miyashita, and F. Tama (Jan. 2010). “Biased coarse-grained molecular dynamics simulation approach for flexible fitting of X-ray structure into cryo electron microscopy maps”. en. In: *Journal of structural biology* 169.1, pp. 95–105. ISSN: 1047-8477,1095-8657. DOI: 10.1016/j.jsb.2009.09.010. URL: <http://dx.doi.org/10.1016/j.jsb.2009.09.010>.
- Harauz, G. and M. Van Heel (1986). Exact filters general geometry three dimensional reconstruction.
- Heel, M. van and M. Schatz (Sept. 2005). “Fourier shell correlation threshold criteria”. en. In: *Journal of structural biology* 151.3, pp. 250–262. ISSN: 1047-8477,1095-8657. DOI: 10.1016/j.jsb.2005.05.009. URL: <http://dx.doi.org/10.1016/j.jsb.2005.05.009>.
- Henderson, R., S. Chen, J. Z. Chen, N. Grigorieff, L. A. Passmore, L. Ciccarelli, J. L. Rubinstein, R. A. Crowther, P. L. Stewart, and P. B. Rosenthal (Nov. 2011). “Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy”. en. In: *Journal of molecular biology* 413.5,

- pp. 1028–1046. ISSN: 0022-2836,1089-8638. DOI: 10.1016/j.jmb.2011.09.008. URL: <http://dx.doi.org/10.1016/j.jmb.2011.09.008>.
- Hinsen, K. (Nov. 1998). “Analysis of domain motions by approximate normal mode calculations”. en. In: *Proteins* 33.3, pp. 417–429. ISSN: 0887-3585,1097-0134. DOI: 10.1002/(sici)1097-0134(19981115)33:3<417::aid-prot10>3.0.co;2-8. URL: [http://dx.doi.org/10.1002/\(sici\)1097-0134\(19981115\)33:3%3C417::aid-prot10%3E3.0.co;2-8](http://dx.doi.org/10.1002/(sici)1097-0134(19981115)33:3%3C417::aid-prot10%3E3.0.co;2-8).
- Hinsen, K., E. Beaumont, B. Fournier, and J.-J. Lacapère (2010). “From electron microscopy maps to atomic structures using normal mode-based fitting”. en. In: *Methods in molecular biology* (Clifton, N.J.) 654, pp. 237–258. ISSN: 1064-3745,1940-6029. DOI: 10.1007/978-1-60761-762-4_13. URL: http://dx.doi.org/10.1007/978-1-60761-762-4_13.
- Hopf, T. A., L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks (June 2012). “Three-dimensional structures of membrane proteins from genomic sequencing”. en. In: *Cell* 149.7, pp. 1607–1621. ISSN: 0092-8674,1097-4172. DOI: 10.1016/j.cell.2012.04.012. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3641781/>.
- Hryc, C. F., D.-H. Chen, P. V. Afonine, J. Jakana, Z. Wang, C. Haase-Pettingell, W. Jiang, P. D. Adams, J. A. King, M. F. Schmid, and W. Chiu (Mar. 2017). “Accurate model annotation of a near-atomic resolution cryo-EM map”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 114.12, pp. 3103–3108. ISSN: 0027-8424,1091-6490. DOI: 10.1073/pnas.1621152114. URL: <http://dx.doi.org/10.1073/pnas.1621152114>.
- Huang, K., X. Chen, C. Li, Q. Song, H. Li, L. Zhu, Y. Yang, and A. Ren (Dec. 2021). “Structure-based investigation of fluorogenic Pepper aptamer”. en. In: *Nature chemical biology* 17.12, pp. 1289–1295. ISSN: 1552-4450,1552-4469. DOI: 10.1038/s41589-021-00884-6. URL: <https://www.nature.com/articles/s41589-021-00884-6>.
- Igaev, M., C. Kutzner, L. V. Bock, A. C. Vaiana, and H. Grubmüller (Mar. 2019). “Automated cryo-EM structure refinement using correlation-driven molecular dynamics”. en. In: *eLife* 8. ISSN: 2050-084X. DOI: 10.7554/eLife.43542. URL: <http://dx.doi.org/10.7554/eLife.43542>.
- Jamali, K., L. Käll, R. Zhang, A. Brown, D. Kimanius, and S. H. W. Scheres (Feb. 2024). “Automated model building and protein identification in cryo-EM maps”. en. In: *Nature*. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-024-07215-4. URL: <http://dx.doi.org/10.1038/s41586-024-07215-4>.
- Jones, D. T., D. W. A. Buchan, D. Cozzetto, and M. Pontil (Jan. 2012). “PSI-COV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments”. en. In: *Bioinformatics* (Oxford, England) 28.2, pp. 184–190. ISSN: 1367-4811,1367-4803. DOI: 10.1093/bioinformatics/btr638. URL: <https://pubmed.ncbi.nlm.nih.gov/22101153/>.
- Joseph, A. P., S. Malhotra, T. Burnley, C. Wood, D. K. Clare, M. Winn, and M. Topf (May 2016). “Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment”. en. In: *Methods* 100, pp. 42–49. ISSN: 1046-2023,1095-9130. DOI: 10.1016/j.ymeth.2016.03.007. URL: <http://dx.doi.org/10.1016/j.ymeth.2016.03.007>.
- Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. deK, A. Potapenko, A. Bridgland, C. Meyer,

- S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis (Aug. 2021). “Highly accurate protein structure prediction with AlphaFold”. en. In: *Nature* 596.7873, pp. 583–589. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: <http://dx.doi.org/10.1038/s41586-021-03819-2>.
- Kidmose, R., J. Juhl, P. Nissen, T. Boesen, J. Karlsen, and B. P. Pedersen (Dec. 2018). “Namdinator automatic molecular dynamics flexible fitting of structural models into cryo-EM and crystallography experimental maps”. In: *bioRxiv* 6, pp. 526–531. ISSN: 2052-2525. DOI: 10.1101/501197. URL: <https://scripts.iucr.org/cgi-bin/paper?eh5002>.
- Kim, D. N., N. W. Moriarty, S. Kirmizialtin, P. V. Afonine, B. Poon, O. V. Sobolev, P. D. Adams, and K. Sanbonmatsu (Oct. 2019). “Cryo_fit: Democratization of flexible fitting for cryo-EM”. en. In: *Journal of structural biology* 208.1, pp. 1–6. ISSN: 1047-8477,1095-8657. DOI: 10.1016/j.jsb.2019.05.012. URL: <http://dx.doi.org/10.1016/j.jsb.2019.05.012>.
- Kim, D. N. and K. Y. Sanbonmatsu (Dec. 2017). “Tools for the cryo-EM gold rush: going from the cryo-EM map to the atomistic model”. en. In: *Bioscience reports* 37.6. ISSN: 0144-8463,1573-4935. DOI: 10.1042/bsr20170072. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5715128/>.
- Kretsch, R. C., L. Xu, I. N. Zheludev, X. Zhou, R. Huang, G. Nye, S. Li, K. Zhang, W. Chiu, and R. Das (Mar. 2024). “Tertiary folds of the SL5 RNA from the 5’ proximal region of SARS-CoV-2 and related coronaviruses”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 121.10, e2320493121. ISSN: 0027-8424,1091-6490. DOI: 10.1073/pnas.2320493121. URL: <http://dx.doi.org/10.1073/pnas.2320493121>.
- Kryshtafovych, A., S. Malhotra, B. Monastyrskyy, T. Cragolini, A.-P. Joseph, W. Chiu, and M. Topf (Dec. 2019). “Cryo-electron microscopy targets in CASP13: Overview and evaluation of results”. en. In: *Proteins* 87.12, pp. 1128–1140. ISSN: 0887-3585,1097-0134. DOI: 10.1002/prot.25817. URL: <http://dx.doi.org/10.1002/prot.25817>.
- Kryshtafovych, A., T. Schwede, M. Topf, K. Fidelis, and J. Moult (Dec. 2021). “Critical assessment of methods of protein structure prediction (CASP)-Round XIV”. en. In: *Proteins* 89.12, pp. 1607–1617. ISSN: 0887-3585,1097-0134. DOI: 10.1002/prot.26237. URL: <http://dx.doi.org/10.1002/prot.26237>.
- Kucukelbir, A., F. J. Sigworth, and H. D. Tagare (Jan. 2014). “Quantifying the local resolution of cryo-EM density maps”. en. In: *Nature methods* 11.1, pp. 63–65. ISSN: 1548-7091,1548-7105. DOI: 10.1038/nmeth.2727. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3903095/>.
- Kühlbrandt, W. (Mar. 2014). “Biochemistry. The resolution revolution”. en. In: *Science* 343.6178, pp. 1443–1444. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1251652. URL: <http://dx.doi.org/10.1126/science.1251652>.
- Kulik, M., M. L. Chodkiewicz, and P. M. Dominiak (Aug. 2022). “Theoretical 3D electron diffraction electrostatic potential maps of proteins modeled with a multipolar pseudoatom data bank”. In: *Acta crystallographica. Section D, Structural biology* 78.8, pp. 1010–1020. ISSN: 2059-7983. DOI: 10.1107/s2059798322005836. URL: <http://dx.doi.org/10.1107/s2059798322005836>.

- Leontis, N. B. and E. Westhof (Apr. 2001). “Geometric nomenclature and classification of RNA base pairs”. en. In: RNA 7.4, pp. 499–512. ISSN: 1355-8382. DOI: 10.1017/s1355838201002515. URL: <http://dx.doi.org/10.1017/s1355838201002515>.
- Leontis, N. B., A. Lescoute, and E. Westhof (June 2006). “The building blocks and motifs of RNA architecture”. en. In: Current opinion in structural biology 16.3, pp. 279–287. ISSN: 0959-440X,1879-033X. DOI: 10.1016/j.sbi.2006.05.009. URL: <http://dx.doi.org/10.1016/j.sbi.2006.05.009>.
- Li, X., P. Mooney, S. Zheng, C. R. Booth, M. B. Braunfeld, S. Gubbens, D. A. Agard, and Y. Cheng (June 2013). “Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM”. en. In: Nature methods 10.6, pp. 584–590. ISSN: 1548-7091,1548-7105. DOI: 10.1038/nmeth.2472. URL: <http://dx.doi.org/10.1038/nmeth.2472>.
- Liebschner, D., P. V. Afonine, N. W. Moriarty, B. K. Poon, V. B. Chen, and P. D. Adams (Jan. 2021). “CERES: a cryo-EM re-refinement system for continuous improvement of deposited models”. en. In: Acta crystallographica. Section D, Structural biology 77.Pt 1, pp. 48–61. ISSN: 2059-7983. DOI: 10.1107/S2059798320015879. URL: <http://dx.doi.org/10.1107/S2059798320015879>.
- Lindert, S. and J. A. McCammon (Mar. 2015). “Improved cryoEM-guided iterative molecular dynamics–Rosetta protein structure refinement protocol for high precision protein structure prediction”. en. In: Journal of chemical theory and computation 11.3, pp. 1337–1346. ISSN: 1549-9618,1549-9626. DOI: 10.1021/ct500995d. URL: <http://dx.doi.org/10.1021/ct500995d>.
- Liu, D., F. A. Th  lot, J. A. Piccirilli, M. Liao, and P. Yin (May 2022). “Sub-3-   cryo-EM structure of RNA enabled by engineered homomeric self-assembly”. en. In: Nature methods 19.5, pp. 576–585. ISSN: 1548-7105,1548-7091. DOI: 10.1038/s41592-022-01455-w. URL: <https://pubmed.ncbi.nlm.nih.gov/35501384/>.
- Liu, T. and A. M. Pyle (Apr. 2024). “Highly reactive group I introns ubiquitous in pathogenic fungi”. en. In: Journal of molecular biology 436.8, p. 168513. ISSN: 0022-2836,1089-8638. DOI: 10.1016/j.jmb.2024.168513. URL: <http://dx.doi.org/10.1016/j.jmb.2024.168513>.
- Marques, M. A., M. D. Purdy, and M. Yeager (Oct. 2019). “CryoEM maps are full of potential”. en. In: Current opinion in structural biology 58, pp. 214–223. ISSN: 0959-440X,1879-033X. DOI: 10.1016/j.sbi.2019.04.006. URL: <http://dx.doi.org/10.1016/j.sbi.2019.04.006>.
- Martin, T. G., T. A. M. Bharat, A. C. Joerger, X.-C. Bai, F. Praetorius, A. R. Fersht, H. Dietz, and S. H. W. Scheres (Nov. 2016). “Design of a molecular support for cryo-EM structure determination”. en. In: Proceedings of the National Academy of Sciences of the United States of America 113.47, E7456–E7463. ISSN: 0027-8424,1091-6490. DOI: 10.1073/pnas.1612720113. URL: <http://dx.doi.org/10.1073/pnas.1612720113>.
- Mitton-Fry, R. M., S. J. DeGregorio, J. Wang, T. A. Steitz, and J. A. Steitz (Nov. 2010). “Poly(A) tail recognition by a viral RNA element through assembly of a triple helix”. en. In: Science (New York, N.Y.) 330.6008, pp. 1244–1247. ISSN: 0036-8075,1095-9203. DOI: 10.1126/science.1195858. URL: <http://dx.doi.org/10.1126/science.1195858>.
- Morcos, F., T. Hwa, J. N. Onuchic, and M. Weigt (2014). “Direct coupling analysis for protein contact prediction”. en. In: Methods in molecular biology (Clifton,

- N.J.) 1137, pp. 55–70. ISSN: 1064-3745,1940-6029. DOI: 10.1007/978-1-4939-0366-5_5. URL: http://dx.doi.org/10.1007/978-1-4939-0366-5_5.
- Mulvaney**, T., R. C. Kretsch, L. Elliott, J. G. Beton, A. Kryshtafovych, D. J. Rigden, R. Das, and M. Topf (Dec. 2023). “CASP15 cryo-EM protein and RNA targets: Refinement and analysis using experimental maps”. en. In: *Proteins* 91.12, pp. 1935–1951. ISSN: 0887-3585, 1097-0134. DOI: 10.1002/prot.26644. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26644>.
- Murray, C. J. L. et al. (Feb. 2022). “Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis”. en. In: *The Lancet* 399.10325, pp. 629–655. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(21)02724-0. URL: <http://www.thelancet.com/article/S0140673621027240/abstract>.
- Murray, L. J. W., W. B. Arendall 3rd, D. C. Richardson, and J. S. Richardson (Nov. 2003). “RNA backbone is rotameric”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.24, pp. 13904–13909. ISSN: 0027-8424,1091-6490. DOI: 10.1073/pnas.1835769100. URL: <https://pubmed.ncbi.nlm.nih.gov/14612579/>.
- Nicholls, R. A., F. Long, and G. N. Murshudov (Apr. 2012). “Low-resolution refinement tools in REFMAC5”. en. In: *Acta crystallographica. Section D, Biological crystallography* 68.Pt 4, pp. 404–417. ISSN: 0907-4449,1399-0047. DOI: 10.1107/S090744491105606X. URL: <http://dx.doi.org/10.1107/S090744491105606X>.
- Nilaratanakul, V., D. A. Hauer, and D. E. Griffin (May 2017). “Development and characterization of Sindbis virus with encoded fluorescent RNA aptamer Spinach2 for imaging of replication and immune-mediated changes in intracellular viral RNA”. en. In: *The Journal of general virology* 98.5, pp. 992–1003. ISSN: 0022-1317,1465-2099. DOI: 10.1099/jgv.0.000755. URL: <http://dx.doi.org/10.1099/jgv.0.000755>.
- (Mar. 2020). “Development of encoded Broccoli RNA aptamers for live cell imaging of alphavirus genomic and subgenomic RNAs”. en. In: *Scientific reports* 10.1, p. 5233. ISSN: 2045-2322. DOI: 10.1038/s41598-020-61573-3. URL: <https://www.nature.com/articles/s41598-020-61573-3>.
- Oldfield, T. J. (Jan. 2001). “A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent”. en. In: *Acta crystallographica. Section D, Biological crystallography* 57.Pt 1, pp. 82–94. ISSN: 0907-4449,1399-0047. DOI: 10.1107/s0907444900014098. URL: <https://pubmed.ncbi.nlm.nih.gov/11134930/>.
- Olek, M., K. Cowtan, D. Webb, Y. Chaban, and P. Zhang (Apr. 2022). “IceBreaker: Software for high-resolution single-particle cryo-EM with non-uniform ice”. en. In: *Structure* 30.4, 522–531.e4. ISSN: 0969-2126, 1878-4186. DOI: 10.1016/j.str.2022.01.005. URL: <http://dx.doi.org/10.1016/j.str.2022.01.005>.
- Orzechowski, M. and F. Tama (Dec. 2008). “Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations”. en. In: *Biophysical journal* 95.12, pp. 5692–5705. ISSN: 0006-3495,1542-0086. DOI: 10.1529/biophysj.108.139451. URL: https://scholar.google.com/citations?view_op=view_citation&hl=en&citation_for_view=TFZcKRIAAAJ:ufrVoPGSRksC.
- Ozden, B., A. Kryshtafovych, and E. Karaca (Dec. 2023). “The impact of AI-based modeling on the accuracy of protein assembly prediction: Insights from

- CASP15". en. In: *Proteins* 91.12, pp. 1636–1657. ISSN: 0887-3585,1097-0134. DOI: 10.1002/prot.26598. URL: <http://dx.doi.org/10.1002/prot.26598>.
- Paige, J. S., K. Y. Wu, and S. R. Jaffrey (July 2011). “RNA mimics of green fluorescent protein”. en. In: *Science (New York, N.Y.)* 333.6042, pp. 642–646. ISSN: 0036-8075,1095-9203. DOI: 10.1126/science.1207339. URL: <http://dx.doi.org/10.1126/science.1207339>.
- Pandurangan, A. P. and M. Topf (2012a). RIBFIND: a web server for identifying rigid bodies in protein structures and to aid flexible fitting into cryo EM maps. DOI: 10.1093/bioinformatics/bts446. URL: <http://dx.doi.org/10.1093/bioinformatics/bts446>.
- Pandurangan, A. P. and M. Topf (Feb. 2012b). “Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps”. en. In: *Journal of structural biology* 177.2, pp. 520–531. ISSN: 1047-8477, 1095-8657. DOI: 10.1016/j.jsb.2011.10.011. URL: <http://dx.doi.org/10.1016/j.jsb.2011.10.011>.
- Pastor, R. W. (1994). “Techniques and Applications of Langevin Dynamics Simulations”. In: *The Molecular Dynamics of Liquid Crystals*. Dordrecht: Springer Netherlands, pp. 85–138. ISBN: 9789401045094,9789401111683. DOI: 10.1007/978-94-011-1168-3_5. URL: http://dx.doi.org/10.1007/978-94-011-1168-3_5.
- Peng, L. M. (July 1998). “Electron Scattering Factors of Ions and their Parameterization”. In: *Acta crystallographica. Section A, Foundations of crystallography* 54.4, pp. 481–485. ISSN: 0108-7673,1600-5724. DOI: 10.1107/s0108767398001901. URL: <http://dx.doi.org/10.1107/s0108767398001901>.
- (Dec. 1999). “Electron atomic scattering factors and scattering potentials of crystals”. en. In: *Micron (Oxford, England: 1993)* 30.6, pp. 625–648. ISSN: 0968-4328,1878-4291. DOI: 10.1016/s0968-4328(99)00033-5. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0968432899000335>.
- Peng, L. M., G. Ren, S. L. Dudarev, and M. J. Whelan (Mar. 1996). “Robust parameterization of elastic and absorptive electron atomic scattering factors”. In: *Acta crystallographica. Section A, Crystal physics, diffraction, theoretical and general crystallography* 52.2, pp. 257–276. ISSN: 0567-7394. DOI: 10.1107/s0108767395014371. URL: <https://journals.iucr.org/a/issues/1996/02/00/zh0006/zh0006.pdf>.
- Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin (Oct. 2004). “UCSF Chimera—a visualization system for exploratory research and analysis”. en. In: *Journal of computational chemistry* 25.13, pp. 1605–1612. ISSN: 0192-8651. DOI: 10.1002/jcc.20084. URL: <http://dx.doi.org/10.1002/jcc.20084>.
- Phillips, J. C., D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot, and E. Tajkhorshid (July 2020). “Scalable molecular dynamics on CPU and GPU architectures with NAMD”. en. In: *The Journal of chemical physics* 153.4, p. 044130. ISSN: 0021-9606,1089-7690. DOI: 10.1063/5.0014475. URL: <http://dx.doi.org/10.1063/5.0014475>.
- Pintilie, G. and W. Chiu (Sept. 2021). “Validation, analysis and annotation of cryo-EM structures”. en. In: *Acta crystallographica. Section D, Structural biology*

- 77.Pt 9, pp. 1142–1152. ISSN: 2059-7983. DOI: 10.1107/S2059798321006069. URL: <http://dx.doi.org/10.1107/S2059798321006069>.
- Punjani, A., J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker (Mar. 2017). “cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination”. en. In: *Nature methods* 14.3, pp. 290–296. ISSN: 1548-7091,1548-7105. DOI: 10.1038/nmeth.4169. URL: <http://dx.doi.org/10.1038/nmeth.4169>.
- Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan (July 1963). “Stereochemistry of polypeptide chain configurations”. en. In: *Journal of molecular biology* 7, pp. 95–99. ISSN: 0022-2836,1089-8638. DOI: 10.1016/s0022-2836(63)80023-6. URL: [http://dx.doi.org/10.1016/s0022-2836\(63\)80023-6](http://dx.doi.org/10.1016/s0022-2836(63)80023-6).
- Ranson, N. A., G. W. Farr, A. M. Roseman, B. Gowen, W. A. Fenton, A. L. Horwich, and H. R. Saibil (Dec. 2001). “ATP-bound states of GroEL captured by cryo-electron microscopy”. en. In: *Cell* 107.7, pp. 869–879. ISSN: 0092-8674,1097-4172. DOI: 10.1016/s0092-8674(01)00617-1. URL: [http://dx.doi.org/10.1016/s0092-8674\(01\)00617-1](http://dx.doi.org/10.1016/s0092-8674(01)00617-1).
- Richardson, J. S., B. Schneider, L. W. Murray, G. J. Kapral, R. M. Immormino, J. J. Headd, D. C. Richardson, D. Ham, E. Hershkovits, L. D. Williams, K. S. Keating, A. M. Pyle, D. Micalef, J. Westbrook, H. M. Berman, and RNA Ontology Consortium (Mar. 2008). “RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution)”. en. In: *RNA (New York, N.Y.)* 14.3, pp. 465–481. ISSN: 1355-8382,1469-9001. DOI: 10.1261/rna.657708. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2248255/>.
- Rohou, A. and N. Grigorieff (Nov. 2015). “CTFFIND4: Fast and accurate defocus estimation from electron micrographs”. en. In: *Journal of structural biology* 192.2, pp. 216–221. ISSN: 1047-8477,1095-8657. DOI: 10.1016/j.jsb.2015.08.008. URL: <http://dx.doi.org/10.1016/j.jsb.2015.08.008>.
- Rosenthal, P. B. and R. Henderson (Oct. 2003). “Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy”. en. In: *Journal of molecular biology* 333.4, pp. 721–745. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2003.07.013. URL: <http://dx.doi.org/10.1016/j.jmb.2003.07.013>.
- Rosenthal, P. B. and J. L. Rubinstein (Oct. 2015). “Validating maps from single particle electron cryomicroscopy”. en. In: *Current opinion in structural biology* 34, pp. 135–144. ISSN: 0959-440X,1879-033X. DOI: 10.1016/j.sbi.2015.07.002. URL: <http://dx.doi.org/10.1016/j.sbi.2015.07.002>.
- Sali, A., L. Potterton, F. Yuan, H. van Vlijmen, and M. Karplus (Nov. 1995). “Evaluation of comparative protein modeling by MODELLER”. en. In: *Proteins* 23.3, pp. 318–326. ISSN: 0887-3585,1097-0134. DOI: 10.1002/prot.340230306. URL: <http://dx.doi.org/10.1002/prot.340230306>.
- Sampedro Vallina, N., E. K. S. McRae, B. K. Hansen, A. Boussebayle, and E. S. Andersen (May 2023). “RNA origami scaffolds facilitate cryo-EM characterization of a Broccoli-Pepper aptamer FRET pair”. en. In: *Nucleic acids research* 51.9, pp. 4613–4624. ISSN: 0305-1048,1362-4962. DOI: 10.1093/nar/gkad224. URL: <https://academic.oup.com/nar/article/51/9/4613/7097662>.
- Scheres, S. H. W. (Dec. 2012). “RELION: implementation of a Bayesian approach to cryo-EM structure determination”. en. In: *Journal of structural biology* 180.3,

- pp. 519–530. ISSN: 1047-8477,1095-8657. DOI: 10.1016/j.jsb.2012.09.006. URL: <http://dx.doi.org/10.1016/j.jsb.2012.09.006>.
- Schreiner, E., L. G. Trabuco, P. L. Freddolino, and K. Schulten (May 2011). “Stereochemical errors and their implications for molecular dynamics simulations”. en. In: *BMC bioinformatics* 12.1, p. 190. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-190. URL: <http://dx.doi.org/10.1186/1471-2105-12-190>.
- Schwede, T., J. Kopp, N. Guex, and M. C. Peitsch (July 2003). “SWISS-MODEL: An automated protein homology-modeling server”. en. In: *Nucleic acids research* 31.13, pp. 3381–3385. ISSN: 0305-1048,1362-4962. DOI: 10.1093/nar/gkg520. URL: <http://dx.doi.org/10.1093/nar/gkg520>.
- Shimomura, O., F. H. Johnson, and Y. Saiga (June 1962). “Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, *Aequorea*”. en. In: *Journal of cellular and comparative physiology* 59.3, pp. 223–239. ISSN: 0095-9898,1553-0809. DOI: 10.1002/jcp.1030590302. URL: <http://dx.doi.org/10.1002/jcp.1030590302>.
- Silvester, E., B. Vollmer, V. Praák, D. Vasishtan, E. A. Machala, C. Whittle, S. Black, J. Bath, A. J. Turberfield, K. Grünewald, and L. A. Baker (Feb. 2021). “DNA origami signposts for identifying proteins on cell membranes by electron cryotomography”. en. In: *Cell* 184.4, 1110–1121.e16. ISSN: 0092-8674,1097-4172. DOI: 10.1016/j.cell.2021.01.033. URL: <http://dx.doi.org/10.1016/j.cell.2021.01.033>.
- Simons, K. T., R. Bonneau, I. Ruczinski, and D. Baker (1999). “Ab initio protein structure prediction of CASP III targets using ROSETTA”. In: *Proteins* 37.S3, pp. 171–176. ISSN: 0887-3585,1097-0134. DOI: 10.1002/(sici)1097-0134(1999)37:3+<171::aid-prot21>3.3.co;2-q. URL: [http://dx.doi.org/10.1002/\(sici\)1097-0134\(1999\)37:3+%3C171::aid-prot21%3E3.3.co;2-q](http://dx.doi.org/10.1002/(sici)1097-0134(1999)37:3+%3C171::aid-prot21%3E3.3.co;2-q).
- Singharoy, A., I. Teo, R. McGreevy, J. E. Stone, J. Zhao, and K. Schulten (July 2016). “Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps”. en. In: *eLife* 5. ISSN: 2050-084X. DOI: 10.7554/eLife.16105. URL: <http://dx.doi.org/10.7554/eLife.16105>.
- Smart, O. S., T. O. Womack, C. Flensburg, P. Keller, W. Paciorek, A. Sharff, C. Vonrhein, and G. Bricogne (Apr. 2012). “Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER”. en. In: *Acta crystallographica. Section D, Biological crystallography* 68.Pt 4, pp. 368–380. ISSN: 0907-4449,1399-0047. DOI: 10.1107/S0907444911056058. URL: <https://journals.iucr.org/d/issues/2012/04/00/ba5178/index.html>.
- Souza, P. C. T., R. Alessandri, J. Barnoud, S. Thallmair, I. Faustino, F. Grünewald, I. Patmanidis, H. Abdizadeh, B. M. H. Bruininks, T. A. Wassenaar, P. C. Kroon, J. Melcr, V. Nieto, V. Corradi, H. M. Khan, J. Domaski, M. Javanainen, H. Martinez-Seara, N. Reuter, R. B. Best, I. Vattulainen, L. Monticelli, X. Periole, D. P. Tieleman, A. H. de Vries, and S. J. Marrink (Apr. 2021). “Martini 3: a general purpose force field for coarse-grained molecular dynamics”. en. In: *Nature methods* 18.4, pp. 382–388. ISSN: 1548-7091,1548-7105. DOI: 10.1038/s41592-021-01098-3. URL: <http://dx.doi.org/10.1038/s41592-021-01098-3>.
- Stahl, K., A. Graziadei, T. Dau, O. Brock, and J. Rappsilber (Dec. 2023). “Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning”. en. In: *Nature biotechnology* 41.12, pp. 1810–1819. ISSN: 1087-

- 0156,1546-1696. DOI: 10.1038/s41587-023-01704-z. URL: <https://www.nature.com/articles/s41587-023-01704-z>.
- Stasiewicz, J., S. Mukherjee, C. Nithin, and J. M. Bujnicki (Mar. 2019). “QRNAS: software tool for refinement of nucleic acid structures”. en. In: *BMC structural biology* 19.1, p. 5. ISSN: 1472-6807. DOI: 10.1186/s12900-019-0103-1. URL: <http://dx.doi.org/10.1186/s12900-019-0103-1>.
- Swope, W. C. and H. C. Andersen (Dec. 1984). “A molecular dynamics method for calculating the solubility of gases in liquids and the hydrophobic hydration of inert-gas atoms in aqueous solution”. en. In: *The journal of physical chemistry* 88.26, pp. 6548–6556. ISSN: 0022-3654,1541-5740. DOI: 10.1021/j150670a016. URL: <http://dx.doi.org/10.1021/j150670a016>.
- Tama, F., O. Miyashita, and C. L. Brooks 3rd (Apr. 2004). “Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis”. en. In: *Journal of molecular biology* 337.4, pp. 985–999. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2004.01.048. URL: <http://dx.doi.org/10.1016/j.jmb.2004.01.048>.
- Tan, Y. Z., P. R. Baldwin, J. H. Davis, J. R. Williamson, C. S. Potter, B. Carragher, and D. Lyumkis (Aug. 2017). “Addressing preferred specimen orientation in single-particle cryo-EM through tilting”. en. In: *Nature methods* 14.8, pp. 793–796. ISSN: 1548-7091,1548-7105. DOI: 10.1038/nmeth.4347. URL: <http://dx.doi.org/10.1038/nmeth.4347>.
- Topf, M., K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali (Feb. 2008). “Protein structure fitting and refinement guided by cryo-EM density”. en. In: *Structure* 16.2, pp. 295–307. ISSN: 0969-2126. DOI: 10.1016/j.str.2007.11.016. URL: <http://dx.doi.org/10.1016/j.str.2007.11.016>.
- Trabuco, L. G., E. Villa, K. Mitra, J. Frank, and K. Schulten (May 2008). “Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics”. en. In: *Structure* 16.5, pp. 673–683. ISSN: 0969-2126. DOI: 10.1016/j.str.2008.03.005. URL: <http://dx.doi.org/10.1016/j.str.2008.03.005>.
- Uversky, V. N. (Mar. 2016). “Dancing protein clouds: The strange biology and chaotic physics of intrinsically disordered proteins”. en. In: *The journal of biological chemistry* 291.13, pp. 6681–6688. ISSN: 1083-351X,0021-9258. DOI: 10.1074/jbc.R115.685859. URL: <https://pubmed.ncbi.nlm.nih.gov/26851286/>.
- Vögele, J., D. Hyman, J. Martins, J. Ferner, H. R. A. Jonker, A. E. Hargrove, J. E. Weigand, A. Wacker, H. Schwalbe, J. Wöhnert, and E. Duchardt-Ferner (Nov. 2023). “High-resolution structure of stem-loop 4 from the 5'-UTR of SARS-CoV-2 solved by solution state NMR”. en. In: *Nucleic acids research* 51.20, pp. 11318–11331. ISSN: 0305-1048,1362-4962. DOI: 10.1093/nar/gkad762. URL: <https://academic.oup.com/nar/article/51/20/11318/7288826>.
- Vuillemot, R., O. Miyashita, F. Tama, I. Rouiller, and S. Jonic (Apr. 2022). “NMMD: Efficient Cryo-EM Flexible Fitting Based on Simultaneous Normal Mode and Molecular Dynamics atomic displacements”. en. In: *Journal of molecular biology* 434.7, p. 167483. ISSN: 0022-2836, 1089-8638. DOI: 10.1016/j.jmb.2022.167483. URL: <http://dx.doi.org/10.1016/j.jmb.2022.167483>.
- Wang, J., Z. Liu, J. Frank, and P. B. Moore (July 2018). “Identification of ions in experimental electrostatic potential maps”. In: *IUCrJ* 5.4, pp. 375–381. ISSN: 2052-2525. DOI: 10.1107/s2052252518006292. URL: <https://pubmed.ncbi.nlm.nih.gov/30002838/>.

- Wang, J., S. K. Natchiar, P. B. Moore, and B. P. Klaholz (Apr. 2021). “Identification of Mg^{2+} ions next to nucleotides in cryo-EM maps using electrostatic potential maps”. en. In: *Acta crystallographica. Section D, Structural biology* 77.4, pp. 534–539. ISSN: 2059-7983. DOI: 10.1107/s2059798321001893. URL: <https://journals.iucr.org/d/issues/2021/04/00/vo5001/index.html>.
- Wang, Y., M. Shekhar, D. Thifault, C. J. Williams, R. McGreevy, J. Richardson, A. Singharoy, and E. Tajkhorshid (Nov. 2018). “Constructing atomic structural models into cryo-EM densities using molecular dynamics - Pros and cons”. en. In: *Journal of structural biology* 204.2, pp. 319–328. ISSN: 1047-8477,1095-8657. DOI: 10.1016/j.jsb.2018.08.003. URL: <http://dx.doi.org/10.1016/j.jsb.2018.08.003>.
- Wang, Z., A. Patwardhan, and G. J. Kleywegt (May 2022). “Validation analysis of EMDB entries”. en. In: *Acta crystallographica. Section D, Structural biology* 78.Pt 5, pp. 542–552. ISSN: 2059-7983. DOI: 10.1107/S205979832200328X. URL: <http://dx.doi.org/10.1107/S205979832200328X>.
- Wriggers, W., R. A. Milligan, and J. A. McCammon (1999). “Situs: A package for docking crystal structures into low-resolution maps from electron microscopy”. en. In: *Journal of structural biology* 125.2-3, pp. 185–195. ISSN: 1047-8477. DOI: 10.1006/jsbi.1998.4080. URL: <http://dx.doi.org/10.1006/jsbi.1998.4080>.
- Yonekura, K., R. Matsuoka, Y. Yamashita, T. Yamane, M. Ikeguchi, A. Kidera, and S. Maki-Yonekura (May 2018). “Ionic scattering factors of atoms that compose biological molecules”. en. In: *IUCrJ* 5.3, pp. 348–353. ISSN: 2052-2525. DOI: 10.1107/s2052252518005237. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5929380/>.
- Zhang, K. (Jan. 2016). “Gctf: Real-time CTF determination and correction”. en. In: *Journal of structural biology* 193.1, pp. 1–12. ISSN: 1047-8477,1095-8657. DOI: 10.1016/j.jsb.2015.11.003. URL: <http://dx.doi.org/10.1016/j.jsb.2015.11.003>.
- Zheng, S. Q., E. Palovcak, J.-P. Armache, K. A. Verba, Y. Cheng, and D. A. Agard (Apr. 2017). “MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy”. en. In: *Nature methods* 14.4, pp. 331–332. ISSN: 1548-7091,1548-7105. DOI: 10.1038/nmeth.4193. URL: <http://dx.doi.org/10.1038/nmeth.4193>.

5

Appendix

No chemicals used.

Acknowledgements

When I finished my undergraduate in 2012, I had no intention to pursue anything academic ever again. Yet here I am, handing in a PhD thesis. I would like to thank my supervisors Maya and Kay for giving me plenty of opportunities to explore, make mistakes, and pursue ideas that interest me. I also have to thank everyone in the lab for creating a collaborative and fun place to work. The papers which make up this thesis involved many people from the lab past and present such as Joe, Tristan and Sony. Without their ideas and help none of this work would have been possible. Like all cryo-EM related things there is a lot of background image processing theory that is often glossed over. Mauro, is a fountain of image processing knowledge, who is always willing to share his expertise. One individual, Matthias Pfeifer, was a major obstruction to the handing in of this thesis with his incessant “Submitted yet?” remarks. Laetitia, on the other hand who pulls all the strings and makes sure everything runs smoothly, cannot be thanked enough. Finally, I’d like to thank my wife Helen for all her support and encouragement. Moving overseas so I could do this PhD was not an easy decision to make, yet she did it in heartbeat.

Affidavit

I hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated. If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.

17.02.2025

Date



Signature
(author)