A Visual Environment for Data Driven Protein Modeling and Validation

Martin Falk, Victor Tobiasson, Alexander Bock, Charles Hansen, and Anders Ynnerman

Abstract—In structural biology, validation and verification of new atomic models are crucial and necessary steps which limit the production of reliable molecular models for publications and databases. An atomic model is the result of meticulous modeling and matching and is evaluated using a variety of metrics that provide clues to improve and refine the model so it fits our understanding of molecules and physical constraints. In cryo electron microscopy (cryo-EM) the validation is also part of an iterative modeling process in which there is a need to judge the quality of the model during the creation phase. A shortcoming is that the process and results of the validation are rarely communicated using visual metaphors.

This work presents a visual framework for molecular validation. The framework was developed in close collaboration with domain experts in a participatory design process. Its core is a novel visual representation based on 2D heatmaps that shows all available validation metrics in a linear fashion, presenting a global overview of the atomic model and provide domain experts with interactive analysis tools. Additional information stemming from the underlying data, such as a variety of local quality measures, is used to guide the user's attention toward regions of higher relevance. Linked with the heatmap is a three-dimensional molecular visualization providing the spatial context of the structures and chosen metrics. Additional views of statistical properties of the structure are included in the visual framework. We demonstrate the utility of the framework and its visual guidance with examples from cryo-EM.

Index Terms—Molecular visualization, cryo-EM, model validation, verification.

1 Introduction

The modern field of structural biology is principally concerned with the documentation and analysis of three-dimensional structures of protein molecules. Either through diffraction experiments employed in X-ray crystallography, measurements of nuclear couplings via nuclear magnetic resonance (NMR), or by direct imaging in cryogenic electron microscopy (cryo-EM) the relative positions of the individual amino acids of proteins can be determined with great accuracy. Over the past 10 years cryo-EM has experienced a 10-fold increase in resolving power allowing it to rival the usage of X-ray crystallography in the determination of protein structures, a process which has been coined the resolution revolution [1], [2].

In the cryo-EM experiment 1–3 µL of purified protein solution is rapidly frozen in order to produce a sample containing protein molecules suspended in a layer of vitreous ice. Analogous to light microscopy, electron phase contrast images of the vitreous suspension are then recorded in the electron microscope, producing projections of the electrostatic potential through the sample. Each image contains tens to hundreds of individual protein molecules, and via a combination of signal processing and Fourier synthesis, the individual projections can be combined into a three-dimensional reconstruction. This resulting volume, referred to as a density map, is a representation of the distribution of electrostatic potential throughout the protein molecule [3]. The map is extremely information-rich however direct interpretation and communication of cryo-EM maps is visually challenging. Therefore, a protein model, being a three-dimensional chemical structure, is created as an analogue to this volumetric representation and deposited alongside the experimentally derived maps into databases such as the RSCB Protein Data Bank (PDB) (rcsb.org) [4] and the Electron Microscopy Data Bank (EMDB) (ebi.ac.uk/emdb) [5].

Creating a protein model from an electrostatic potential map is a manual and laborious process and while aided by computational tools [6], the construction of a single protein model can span weeks to months. As the final model is an interpretation of the experimentally recorded map by the individual structural biologist, careful validation is required to ensure its chemical and physical correctness as well as its accurate reflection of the underlying map. As such, while the construction of models represents a time consuming process in structural analysis, the final evaluation and improvement of those models represent the most crucial, challenging, and error prone steps. Atomic models should never be taken as absolute since the protein structure can change depending on environmental factors. A failure to accurately build and validate a protein model hence risks introducing persistent errors or allows for misinterpretation by the wider community.

In order to ensure a valid interpretation, a host of software and validation metrics are available to the structural biologist; however, the use of visual representations and graphical aids to convey such metrics are limited. Most of

- *M. Falk, A. Bock, and A. Ynnerman are with Linköping University, SE-581 83 Linköping, Sweden and the Swedish e-Science Research Centre (SeRC). Email: {martin.falk | alexander.bock | anders.ynnerman}@liu.se*
- *V. Tobiasson is with the Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, 17165 Solna, Sweden. Email: victor.tobiasson@scilifelab.se*
- *C. Hansen is with Linköping University and the Kahlert School of Computing, University of Utah, Salt Lake City, Utah 84112, USA. E-mail: hansen@cs.utah.edu*

Manuscript received December 15, 2022; revised June 1, 2023.

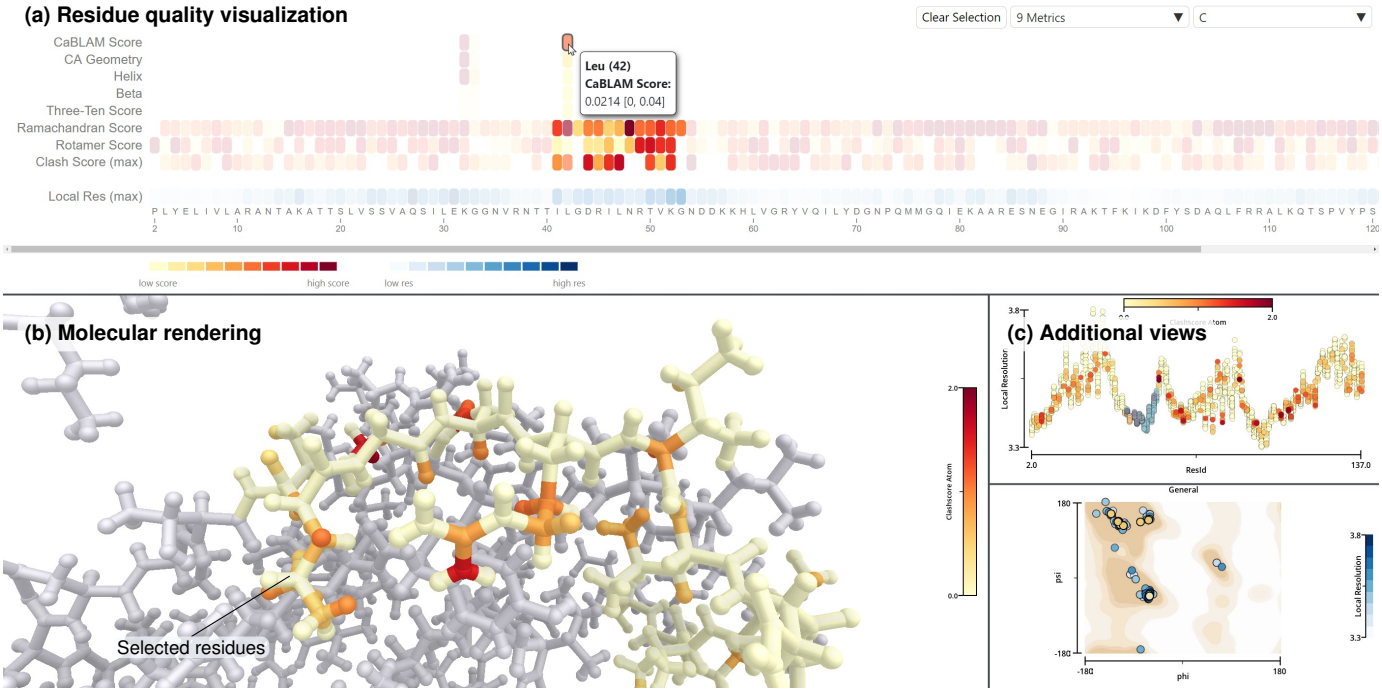


Fig. 1. Our visual guidance framework for map-to-model validation in cryo-EM showcasing the *initial* ribosomal protein bS6m. Multiple linked views enable interactive analysis and inspection of various quality metrics and the corresponding hot spots within an atomic model. Here, a subset of residues was selected in the residue quality panel. The framework consists of the following components: (a) The residue quality visualization, providing a linear overview of the atomic model. (b) A molecular rendering, showing the spatial context matching the residue selection while non-selected residues are depicted in gray for context. (c) Additional plots, facilitating in-depth analysis.

the validation metrics are communicated through tabulated values or, in some cases, compressed into scalar values as an overall score, intended to represent the whole model. As such, the individual metrics are often isolated from each other and with limited reference to the spatial domain.

In this work we are presenting a visual environment supporting this important component of the protein structure modeling workflow. The tool, depicted in Fig. 1, comprises multiple views to tackle the problems of exploring three-dimensional atomic models while integrating data from various validation metrics by means of visualization. Since the visualization of the available high-dimensional validation metrics and their embedding in 3D often leads to information overflow and results in higher cognitive loads on the user, we developed a novel linear representation inspired by heatmaps, the residue quality plot. This visual representation uses the linear backbone of biomolecules, like proteins, as the principal axis while the various validation metrics are shown for each residue on the second axis. To guide the user's attention we incorporate local quality measures to mask parts of the residue quality plot besides the apparent visual cues provided by the heatmap and color mapping. A 3D molecular visualization complements the quality visualization and provides the spatial context, which is important when considering geometric constraints like overlapping atoms and bonds. Additional views including scatter plots and Ramachandran plots enable the analysis and inspection of statistical properties for detecting outliers. In summary, we make the following contributions in this paper.

- A novel framework utilizing visualization for supporting model validation and verification of atomic models developed in close collaboration with domain experts.
- A visual metaphor for visualizing high-dimensional data of biomolecules based on heatmaps which serves as a linear overview of the complex 3D structure as well as interface for interactive analysis.
- Incorporating local quality measures like the local resolution derived from the cryo-EM density map to highlight areas of higher relevance.
- The utility of the framework is demonstrated with a variety of cryo-EM examples.

2 Background

Cryo-EM is a recent technology for determining the 3D structure of proteins. These macromolecular structures are a prerequisite for the modeling of the atomic structure of proteins and understanding their function. A simplified workflow for the process of determining a macromolecular structure is depicted in Fig. 2. The main components are the purification of a biological sample, followed by collection of transmission images in the electron microscope. These transmission images can then be used to reconstruct a 3D approximation of the electrostatic potential of the original protein, often referred to as a map. This volume is then interpreted and modeled as a set of atomic coordinates via a manual process of interpretation referred to as model building.

2.1 Map generation

The cryo-EM process starts with a liquid sample of purified protein. Roughly 3 µL containing 1 µg–5 µg of protein is

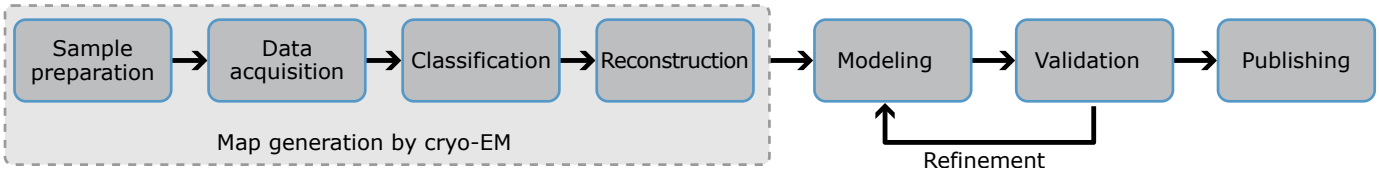


Fig. 2. Cryo-EM workflow with this work focusing on modeling and validation, the most time- and labor-intensive component of the protein reconstruction process.

applied to a sample holder consisting of a $400\,\mu m$ metal mesh covered in a layer of thin carbon or gold foil. The foil contains a series of holes, $1\,\mu m$–$2\,\mu m$ in diameter, through which the sample can be imaged. The sample is then rapidly frozen by immersion in liquid ethane. This rapid vitrification process prevents crystallization of the water solvent and suspends the individual protein molecules in close to random orientations, allowing for the recording of transmission images using the electron microscope. A full dataset contains thousands of images with each image potentially containing hundreds of unique protein molecules. Using image processing, a set of such protein projection images can be averaged and the original three-dimensional protein structure is reconstructed using Fourier synthesis. The resulting volume is referred to as a cryo-EM or density map and is an approximation of the electrostatic electron-scattering potential of the atoms constituting the protein. In addition to 3D reconstruction, clustering and classification algorithms allow for partitioning of the data into distinct subsets of structures which can represent and account for biological heterogeneity in the protein sample. This heterogeneity is often present either as compositional heterogeneity as proteins might bind additional components or subunits, or as conformational flexibility within the protein structure itself. Despite classification efforts, each reconstruction usually retains some level of heterogeneity which, after averaging, appears as blurring or structured noise in the final reconstruction. Therefore, the relative noise level, map fidelity and local resolution can vary significantly across the map. To provide a uniform noise level and aid interpretation, the maps are low pass filtered based on their "local resolution", as estimated by the local Fourier Shell Correlation (FSC) [7], [8]. This variation in local resolution poses significant challenges for model building, validation, and interpretation and makes validation metrics calculated on overall statistics less informative.

2.2 Modeling

The reconstructed map obtained from the cryo-EM experiment contains a wealth of information regarding the atomic composition of the imaged protein. The model building process typically entails a lengthy, semi-manual interpretation of the map density with the aim of producing an accurate and physically probable model of the target protein. The result is a set of atomic coordinates. As an example, in a recent study from our collaborators a model consisting of more than $282\,000$ atoms organized into 95 unique protein polypeptides was created, which required 6 months to complete [9]. Due to trade-offs between model to map correlation and prior knowledge of physical and geometric constraints on atomic structures, model building requires extensive knowledge and subsequent statistical validation in order to produce an accurate model useful for further inquiry.

2.3 Validation

The most crucial step in the cryo-EM workflow is the validation of the molecular model. During the validation process both the agreement between model and map is scored, as well as its conformity to prior statistical models of chemical bond lengths and angles. As a part of this process several well-established geometry metrics are available. Primarily chemical *bond lengths*, that is the distance between two covalently bound atoms, should be consistent with known statistics, and usually vary less than $0.02\,\text{Å}$ from ideal values. Similar tabulated values exist for *bond angles* between neighboring atoms as well as their torsion angles. In addition to distance and angular metrics, the *clash score* is a common metric which accounts for the improper overlap between atomic radii. This is evaluated on both a per atom basis as well as present as a scalar value for the whole model. As large atomic overlaps are highly improbable, the overall clash score should be minimized, however improperly refined models with high clash scores are still widely present in databases.

These strict geometric distance and angle restraints form the most rigorous basis for model validation. Apart from local distances and angles, the peptide backbone in proteins typically adheres to a set of torsion angles, the distribution of which form the basis for the two dimensional Ramachandran plot. Proper adherence to the Ramachandran distribution is a common and well-known metric for model validation and is readily utilized during model building. This in turn limits its usefulness for independent validation and due to common overfitting to Ramachandran statistics, higher-order metrics have been developed. One such metric is the CaBLAM (C-Alpha Based Low-resolution Annotation Method) score developed to provide an evaluation of the peptide backbone spanning several residues [10]. In addition to the peptide backbone geometry, the torsion angles of amino acid side-chains typically conform to a set of probable geometric configurations. These distinct side chain configurations are referred to as rotamers and by evaluating high resolution structures where the atomic positions can be determined directly, a library of likely configurations is created. Adherence to this distribution of rotamers is then useful for validation. Finally, in the specific context of cryo-EM, both the EM ringer and FSC-Q score evaluate a mixture of main chain, rotameric, and local correlation to the map and is cited as an overall scalar or per residue metric for

rotamer fitting accuracy [11], [12]. It is worthwhile to note that in rare cases statistical outliers are still valid as long as they can be adequately supported by the map, a judgement left to the model builder which can significantly increase the difficulty of the validation process.

One of the most popular tools for structure validation and verification is MolProbity [13], which is part of the PHENIX toolbox [14]. It includes many of the commonly used scoring and validation functions described above and also reports overall scores for some of the individual metrics. The different metrics are shown either in text form, tabular form, or plotted in line plots without any spatial context with the exception of Ramachandran plots for Ramachandran statistics. However, it is possible to highlight the matching residue or atom in the external modeling tool Coot [15] or PyMOL, a molecular visualization system [16]. In an iterative process, potentially problematic areas identified during validation are then remodeled and validated again. Once the reported issues are addressed appropriately, the molecular model of the structure can be published. In the remainder of this work, we describe how our tool can support the structural biologist in this modeling-validation loop through the use of visualization and visual guidance.

3 Related Work

Following the basic cryo-EM workflow as described in Section 2, we divide the related work into four parts: signal processing and map generation, protein modeling, validation, and visualization. A more detailed introduction to cryo-EM can be found in the literature [17], [18].

Data processing — RELION [19] and cryoSPARC [20] are commonly used to generate density maps from cryo-EM data which benefits from utilizing GPUs [21]. Scipion [22] serves as unified interface between different cryo-EM software and formats, but can also generate density maps. Our work relies on the resulting density maps independent of the tools used to create these.

Modeling — The reconstructed density maps serve as a reference for the manual modeling of atomic models that is typically performed in dedicated tools like Coot, the Crystallographic Object-Oriented Toolkit [15]. The entire process is facilitated and augmented by toolboxes and frameworks like PHENIX [14], a suite of tools developed for molecular structure determination and validation using data from either cryo-EM or crystallography, or CCP-EM [23]. An alternative to manual modeling is the prediction of protein structures with AlphaFold [24] and subsequent alignment with the cryo-EM density map in a second step followed by model validation using aforementioned tools [25], [26].

Validation — Even though cryo-EM has been utilized for a number of years, from a visualization perspective the validation and analysis of cryo-EM data is rather limited to scatter plots, line plots, and Ramachandran plots for dihedral angles [27]. These plots are typically shown next to linked views with simple molecular renderings in Coot [15] or KiNG (Kinemage, Next Generation) [28]. MolProbity [13], available as standalone as well as integral part of PHENIX, enables the validation of atomic models on the atomic scale with a number of different measures and metrics. The resulting metrics are typically shown in form of tables or the aforementioned plots. There is, however, a lack of visualizations that provide insight into the high-dimensional space spanned by the various validation metrics and thus allow for a visual quality control.

In their work from 2021, Lawson et al. [29] look at data from the Cryo-EM Model Challenge in 2019 with respect to assessing the model quality, reproducibility, and a comparison of scoring metrics. They recommend to consider multiple scoring parameters considering the underlying cryo-EM density map to fully evaluate the atomic model in question. We follow this recommendation by including all available metrics in our visualization.

Visualization — Atomic models can be rendered in different representations using popular molecular visualization tools like VMD [30], PyMOL [16], Jmol [31], VIAMD [32] and UCSF Chimera [33]. Modeling tools, like Coot and KiNG, provide basic visualizations of the atomic models to be build while depicting isosurfaces of the density map as triangle meshes at the same time. Volume rendering is often employed in USCF ChimeraX [34] to depict the underlying electron density maps. Our framework uses its own molecular visualization based on glyph ray casting [35] in order to provide an integrated view along with multiple linked views. For an overview on molecular visualization we refer the reader to the work by Kozlíková et al. [36].

In order to augment the molecular rendering in our work, we include a heatmap-based visualization in a separate view. Heatmaps have been applied successfully to different areas in biology in particular for high-dimensional data [37]. In the context of microarray data, Gehlenborg et al. [38] introduce a 2D colormap for the heatmap that uses the normalized relevance of an expression as second axis. Genome data is often concerned with very long sequences and heatmaps are utilized as a space-efficient approach to encode and visualize the data, for example in the Interactive Genomics Viewer [39]. Our work applies this idea to biomolecular data along with its quality metrics. As proteins can be unraveled into long linear chains, heatmaps can provide a matching overview of the quality along the chains.

4 The Framework

The goal of this work is to provide the tool for visual analysis and guidance in the context of cryo-EM, presenting an overview over validation metrics in the spatial context of the atomic model. Our framework consists of three main components (Fig. 1): 1. molecular visualization; 2. a linear overview of the model provided by the residue quality visualization with metrics; 3. supporting plots like Ramachandran plots, scatter plots, and parallel coordinate plots. All the concepts are implemented in the visualization framework Inviwo [40] using C++ and OpenGL as well as Python and D3.js. Brushing and linking in all views supports seamless interaction between the different visualizations. Additional contextual information, like residue numbers, atom type, and metric scores, is shown in tooltips.

Regular exchanges between the authors from visualization and structural biology were part of the iterative design process and helped to cater for the specific needs and requirements of structural biologists. Additional feedback was provided by coworkers of that author in a small number of

Fig. 3. Residue quality heatmap with the local resolution as weighting factor (ribosomal protein bS6m). No weighting (left), linear weighting (center), and quadratic weighting (right). A higher local resolution corresponds to higher opacity and thus to more precision, that is atom positions in those areas are more reliable.

meetings. These discussions resulted in three major design iterations and multiple minor revisions.

4.1 Data Sources and Processing

In order to avoid unnecessary pre-processing and format conversions, our framework supports data formats commonly used in the cryo-EM workflow. The data needed for the visual guidance can be categorized into three types.

Density map: This scalar volume obtained from volumetric reconstruction (see Section 2.1) holds electron density occupancy values. The data is stored in the MRC 2014 file format with floating point precision. The component is essential during modeling, for validation, and is part of the published structure. The local resolution information per voxel is derived from the density map and the resulting scalar volume is stored using the same data format.

Atomic model: An atomic model consists of atoms and matches the features in the density map. The atom positions, atom type, and additional data like parent residue are stored in a macromolecular Crystallographic Information File (mmCIF) or in the Protein Data Bank format (PDB). Typically, information on bonds is not included and has to be determined after loading the structure.

Validation metrics: MolProbity only shows the resulting metrics reported in tabular form or text as outlined in Section 2.3. Besides saving the summary as text file there are no other options to export the data. Therefore, we added some export functionality to each validation approach in PHENIX, which was incorporated into the most recent release of PHENIX, resulting in one data table per validation metric. Except for atom clash scores, the metrics are given on a per-residue basis. We then combine these tables with information from the atomic model which results in one table containing all available metrics per residue. Note that validation scores are by no means complete either due to thresholding or only outliers being reported. Thus, the residue table is sparsely filled.

We use Python modules, namely mrcfile and Biopython, to import the data from the different sources into our prototype. The validation metrics are loaded directly.

4.2 Molecular Rendering

A large portion of the available screen-space is dedicated to the 3D visualization of the atomic model. The main purpose of this view is to provide a spatial context when inspecting the validation metrics. In line with existing modeling tools, we decided to use a bond-centric model, the ball-and-stick representation. This representation is quite sparse and provides a sense of familiarity to the structural biologists at the same time. Space-filling models and the Van der Waals surfaces are not recommended, even though supported, since both are dense representations. The overlapping atoms cause lots of occlusion and bonds not being visible potentially hides important information.

Instead of coloring the atoms by element, any validation metric can be mapped to color. The reasoning is that the structural biologist is already familiar with the spatial positions and shape of the structures, having spent months working on it. Thus the additional key by atom color is not needed, though it can be toggled if necessary. Mapping validation metrics to the atomic model can provide insight into spatial interrelation of non-neighboring residues. For example, in case of atom clash scores pairs of overlapping atoms are easily spotted and identified as depicted in Fig. 4.

Besides rendering the entire atomic model, the tool supports rendering selections of sub-chains, that is a subset of connected residues, or individual residues. Non-selected parts are de-emphasized by either desaturating the colors of all other structures or hiding them entirely. This facilitates and guides the user to focus on a specific part of the molecule while reducing the visual clutter around it.

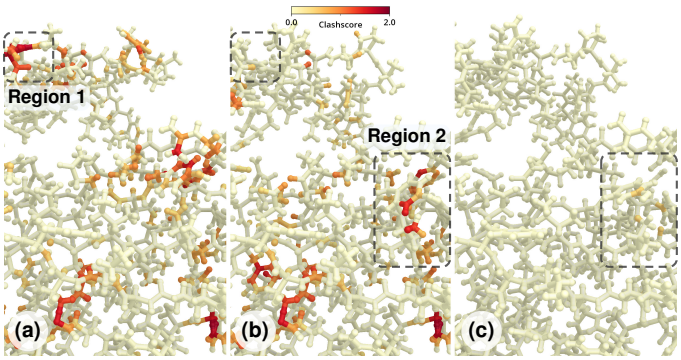


Fig. 4. Visualizing the clash scores for the *deteriorated* (a), *initial* (b), and *refined* model (c) of the ribosomal protein bS6m. In region 1, the atoms of residues 104 (Alanine) and 107 (Phenylalanine) clash due to deterioration while the lower part is mostly unaffected. The refinement of the model reduces most clashes except for residues 63 (Arginine), 64 (Tyrosine), 140 (Leucine), and 146 (Alanine) in region 2.

4.3 Residue Quality Visualization

The visualization of the overall residue quality is an important part of our framework which enables the inspection of various per-residue validation metrics. It provides both a linear overview over the entire modeled structure as well as the means for navigation and selection. Both individual residues and ranges can be added or removed from the current selection.

This visualization itself is based on a 2D heatmap. The horizontal axis represents all residues ordered by their occurrence in the backbone of the atomic model. The labels refer both to the residue index and its name's one-letter abbreviation to ease navigation. The available metrics, which can be filtered, are stacked vertically above each corresponding residue. Since the results of some of the validation metrics are sparse, entries for which there is no data are not shown. For atom-specific metrics, the available data is merged on a per-residue basis for all enclosed atoms. In case of the clash score, we decided to show the maximum overlap as this often indicates the most severe modeling issues. In addition, potential clashing atoms being part of the same residue are close by and easily identifiable in the 3D visualization.

For both visualization and interaction purposes, we decided to use rather large squares with constant size for each entry resulting in a coarsely pixelated representation of the validation metrics. Here, problematic areas appear clearly in this overview and outliers are also easily spotted.

As some of the molecules consist of a large number of residues, this heatmap would be very wide. To counteract this, it is possible to show only specific chains of the molecule in line with the molecular rendering. This reduces the need for horizontal scrolling and at the same time suits the needs of the structural biologists who quite often focus on a particular area during the refinement process.

The information from the validation metrics can optionally be augmented with the local resolution, which expresses uncertainty (see Section 2.1). The local resolution of a single residue is obtained by averaging the local resolutions for all of its atoms. It can directly be used as a cut-off threshold or for importance weighting by means of transparency. If enabled, the cut-off threshold determines the visibility of the data values. When using the importance weighting, the local resolution factors are normalized and then mapped to opacity either linearly or quadratically for further emphasis. This mapping allows the user to focus on regions of higher reliability, that is higher local resolution (Fig. 3).

The heatmap is implemented in D3.js and shown next to the molecular rendering and additional plots.

4.4 Additional Views

Scatter plots are provided to complement the aforementioned residue quality plot and molecular rendering to support further investigation of statistical properties and validation metrics including local resolution information. Since the available metrics and data contains both per-residue data and per-atom data, this can result in a 1-to-n mapping or vice versa. For example, plotting per-atom clash scores against residue indices results in multiple clash scores per residue, one for each related atom. Similarly, selecting an individual residue will highlight all belonging atoms as well

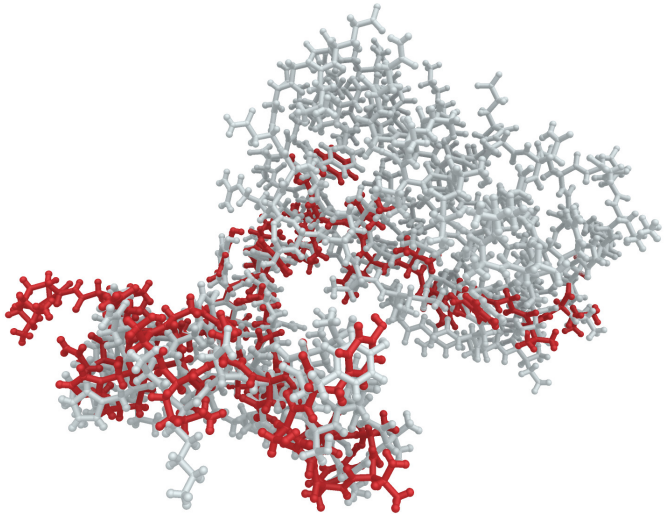


Fig. 5. Atomic model of a mitoribosomal protein bS6m with the initially modeled structure (gray) and a modified, purposefully worsened model with partially shifted atoms (red).

(blue highlights in the center plot of Fig. 7). By incorporating local resolution information these plots can give more insight into data reliability, for instance correlations between a low local resolution in the density map and higher clash scores due to imprecise positioning of atoms.

Ramachandran plots [27], a specialized type of scatter plots commonly used in biochemistry for detecting outliers, are included as well. This type of plot can be shown individually for the four peptide types of residues: general, proline, pre-proline, and glycine (see Figures 7 and 8). It depicts the dihedral angles (Φ, Ψ) of each residue against a distribution of the top-500 proteins. The filled area in the background corresponds to enclose $98\,\%$, $99\,\%$, $99.8\,\%$, and $99.999\,\%$ of the top-500 proteins' angles, respectively. Thus, residues with dihedral angles outside those regions are easily spotted and can be subjected to further inspection and verification.

4.5 Visual Attributes / Color Maps

The color maps used in the framework are taken from D3.js, which are based on ColorBrewer (colorbrewer2.org). We use a continuous yellow-orange-red map for the validation metrics where red correlates with high validation metric scores thereby emphasizing problematic areas of the structure. The data values are normalized with respect to the largest value in order to provide a higher dynamic range. Angles are shown either with a diverging red-white-blue color map or the regular yellow-red color map depending on their range, that is either $[-180°; 180°]$ or $[0°, 360°)$.

The color map of the local resolution is a discretized white-blue gradient. Using a different hue clearly marks the difference between the local resolution and validation metrics when visualizing both at the same time.

The atomic model is rendered using the RasMol CPKnew color scheme [41] or any color-mapped validation metric. Deselected residues and atoms are desaturated or hidden.

5 Use Cases

The datasets used for showcasing our framework were chosen by the domain expert, one of the authors. Both

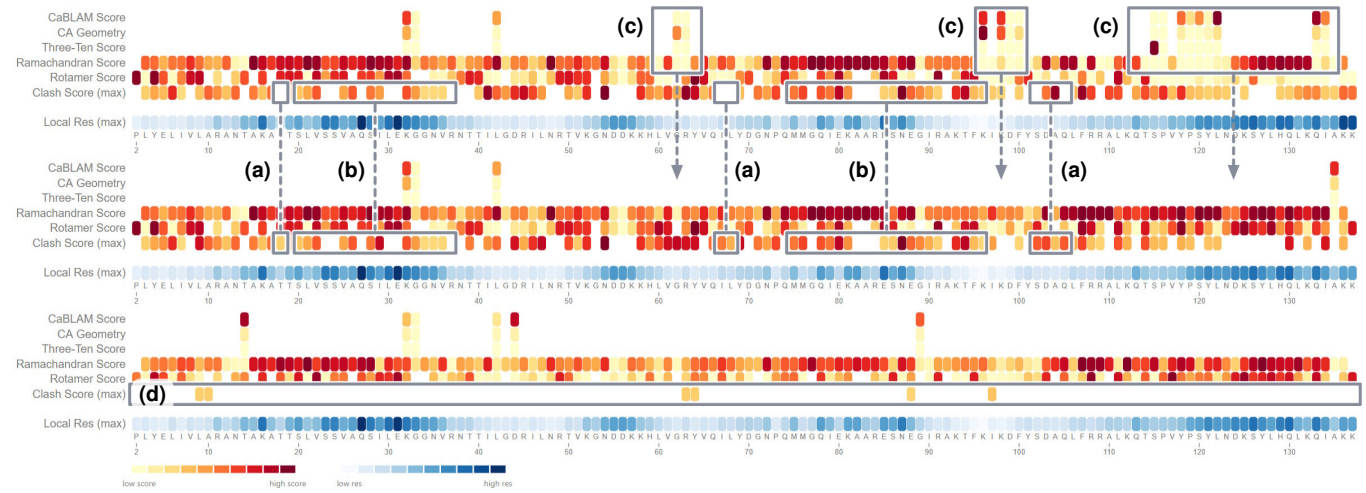


Fig. 6. Comparison between different models of a ribosomal protein bS6m. From top to bottom: artificially *deteriorated* structure with obvious problems, the *initial* atomic model after the first modeling iteration, and the *refined* structure. (a) the deteriorated structure locally improved some clash scores. (b) identical clash scores as only parts were modified. (c) the modified parts show up in various metrics. (d) six clashes are left after refining.

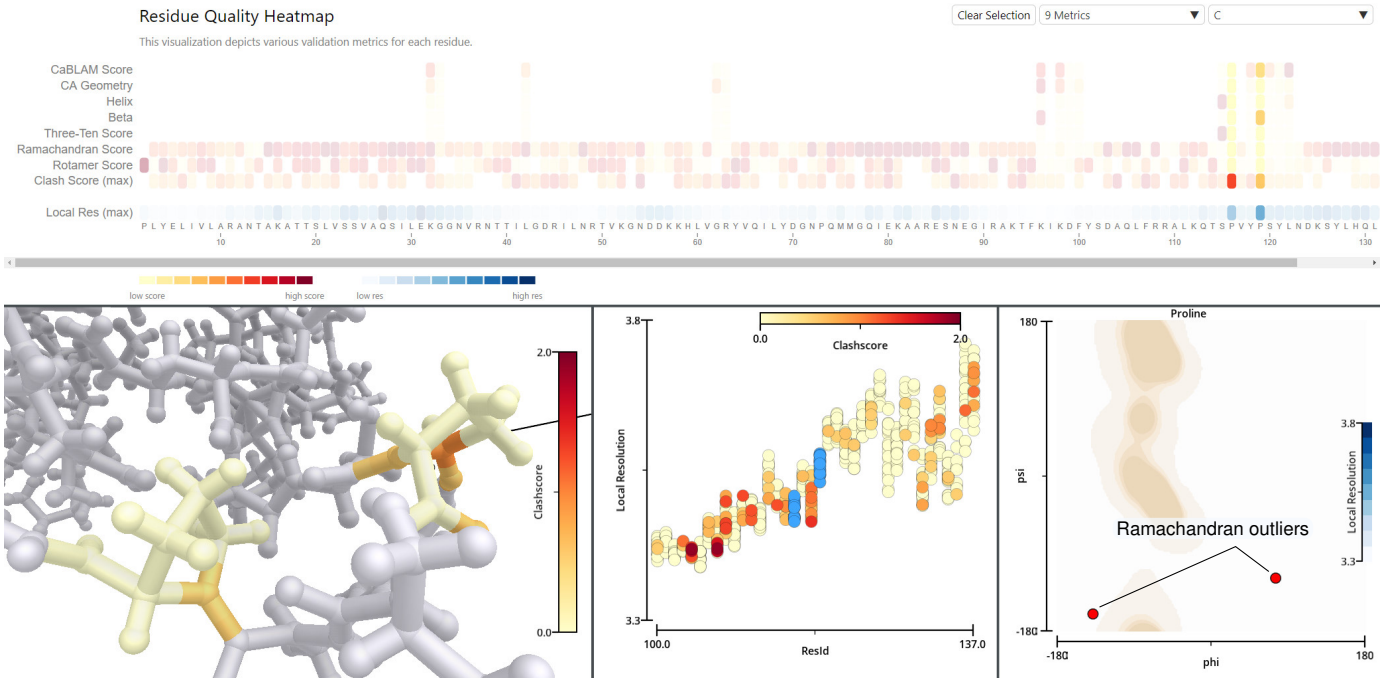


Fig. 7. Close-up inspection of two Ramachandran outliers in the *deteriorated* ribosomal protein bS6m. In the Ramachandran plot, the two selected residues are outside the $99.8\,\%$ interval of the top-500 proteins. In this case, the outlier was caused by one of the modifications to the model.

datasets are part of his research [42], [43]. The smaller one, a ribosomal protein, is intended to demonstrate the basic functionality and concepts while keeping the overall complexity low and avoid clutter and occlusion in the 3D visualization. Furthermore, this facilitates comparisons between slight variations of the structure performed by the domain expert as described below. The chloroplast ribosome contains two orders of magnitude as many atoms and residues than the ribosomal protein and illustrates the scalability of the framework while also highlighting issues with published structures.

5.1 Ribosomal protein (bS6m)

This protein is part of the mitochondrial ribosome of yeast *Saccharomyces cerevisiae* [42] and consists of 2206 atoms of which 1126 are hydrogen atoms that were added after modeling. The corresponding density map has dimensions of $60 \times 60 \times 60$ voxels. For evaluation and illustrative purposes, we use different versions of the atomic model of this molecule. The density map remained unchanged for all three of them. This dataset is featured in most of the illustrations if not otherwise mentioned.

Initial model This model is the first completed model worthy of validation after a couple iterations of modeling and refinement.

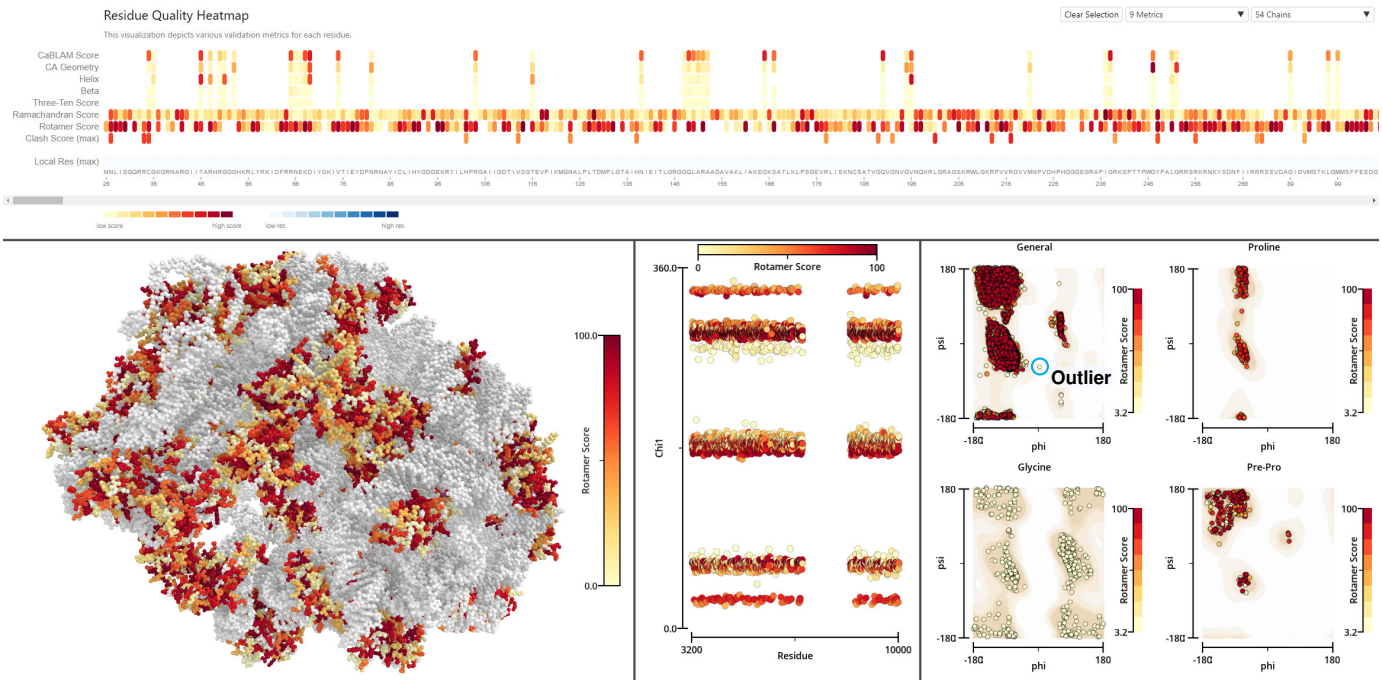


Fig. 8. Analyzing the rotamers of a spinach chloroplast ribosome (PDB-ID: 6ERI, EM Map: EMD-3941). Unknown residues without rotamer information are grayed out in the molecular visualization (lower left). The Ramachandran plot showing the general peptides reveals one residue in the unfavored region (cyan).

Refined model The refined model resulted from iteratively refining the initial model through automated fitting to the density map and manually fixing issues which were indicated throughout the validation. This reflects the regular modeling-validation loop of the cryo-EM workflow as described in Section 2.

Deteriorated model Here, roughly a third of the atoms of the initial model were purposefully modified by moving them slightly, see Fig. 5. The movement had to be large enough to have an effect but at the same time small enough to not sever any covalent bonds. The net effect of this operation was sought to be overall negative with respect to fitting of the density map and, thus, the different validation metrics. Particularly the clash score of overlapping atoms and the binding angles are affected.

Fig. 4 depicts the clash scores corresponding to the three model instances. The differences are quite apparent. The initial model suffers from a number of atom clashes, most of which are also present in the deteriorated case besides more affected areas like in the top and center part of Fig. 4. In case of the refined model, almost all problematic clashes have disappeared. This is confirmed by the residue quality plots show in Fig. 6 where only six residues remain with clashes in the refined model. Since only a part of the model has been modified, the clash scores of some residues are identical between the deteriorated and initial case. See for example residues 19–36 and 71–95. There are even some cases where the modification removes clashes like for residues 18, 67, 68, 102, and 106. The variations in the local resolution (blue color map) are caused by the fact that atoms have been moved.

When looking at the Ramachandran plot for proline peptides in the deteriorated model there are two obvious outliers, see Fig. 7. Selecting the two outliers reveals two close-by residues whose validation metrics are in the acceptable range including the Ramachandran scores. Since these outliers occur neither in the initial nor the refined model, they are caused by the deterioration and not the modeling and, thus, can be ignored after a brief inspection of the molecular rendering.

5.2 The chloroplast ribosome

This spinach chloroplast ribosome (PDB-ID: 6ERI, EM Map: EMD-3941) is involved in oxygenic photosynthesis [43]. The structure was obtained at 3 Å resolution and comprises 145 000 atoms distributed over 11 094 residues. For some of the chains, including about 5000 residues, we were not able to produce any validation metrics as those are marked as "unknown". These chains appear as grayed-out regions in the residue plot.

The large number of residues require browsing and scrolling in the residue plot or reducing the number of selected chains to obtain an overview. At a quick glance, the heatmap reveals a small number of clashing atoms and only a few potential problems with geometric constraints.

The rotamer scores, however, are quite high in places as shown in Fig. 8. A scatterplot of the chi 1 angles of the rotamers features five clusters adopting predetermined torsion angles. Interestingly, the second cluster from the top shows a distinct pattern where residues with low rotamer scores accumulate at a slightly different angle. The Ramachandran scores indicate that most residues lie within the favored and allowed regions with one clear outlier as confirmed by the Ramachandran plots (cyan circle).

As this data concerns a published structure, the density map ($420 \times 420 \times 420$ voxels) has been cleaned up and

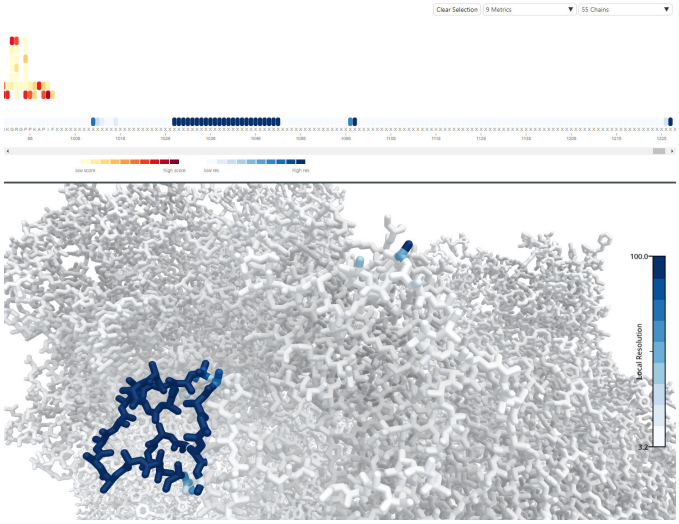


Fig. 9. Spinach chloroplast ribosome (PDB-ID: 6ERI, EM Map: EMD-3941). Mapping the local resolution onto the atomic model reveals potential problems with the underlying map and/or tools used to compute the local resolution from the density map.

smoothed. This in turn affects the local resolution, derived from the density map, which is constant almost everywhere and, thus, is not considered a sensible weighting factor in the residue quality plot. Using the plot and the molecular visualization with the local resolution mapped to color, a small number of outlying regions can be identified as shown in Fig. 9. These regions could either indicate issues with the tool used for computing the local resolution or the published density map. In either case the original data would be required for further analysis.

5.3 User Feedback

The impact of the proposed framework and visualizations is unfortunately hard to quantify directly in the context of the cryo-EM workflow as it augments the mostly manual process of model building of individual structures. In the following, we therefore summarize the feedback and thoughts of our domain expert instead.

Model building is an iterated and curated process asking and testing small hypotheses, especially for poor resolution data. How does this region fit into the density? What happens if you move these residues? Could there be shifts in the amino acid sequence etc. This coupled together with the fact that many times the actual insight and new knowledge that is generated comes primarily at this stage. It is the observations at the stage of model building which ultimately make it into the papers. This is also true for validation where anomalies are more often than not interesting and noteworthy, especially with the existence of more automated tools like AlphaFold [24] nowadays. Here, the main importance is then to convey as much data as possible in an interactive manner. Especially data which is spatially linked is extremely useful as it minimizes the overhead which is required to interpret and validate the individual error metrics. Having several scalar values presented together in a spatial context such as in the proposed framework is unfortunately not present in the currently available tools and both novel and useful.

Another aspect of this work is the overall view it provides for validation. All current tools provide a very local view of validation, typically centered on a single amino acid or clash. This is often misleading as errors often propagate in non-intuitive ways and clusters of errors can usually be resolved together, if one is aware that they are there in the first place.

For an evaluation people would have to actually work with the software. There are many more ephemeral issues which only become apparent when you work for prolonged periods with the software. What is pretty evident is the immediate improvement in visual presentation and data availability in the current implementation. It looks a lot better and easier to work with compared to what we have.

6 Conclusion

This work describes a framework for visual validation and verification of atomic models. The main component, the residue quality plot, is based on 2D heatmaps and visualizes the structure along with different validation metrics in a linear fashion. Thereby, it provides a compact overview over the atomic model and serves as navigational interface by means of brushing and linking. The visualization is augmented with additional plots and a molecular visualization for spatial context.

We used an iterative process involving domain experts to develop the framework with the goal of augmenting the currently established cryo-EM modeling pipeline. Design decisions were taken in close collaboration with domain experts in the field of structural biology. Visualization concepts uncommon in this field, like the parallel coordinates, were considered interesting but not required for validation and, thus, not utilized. The framework is able to highlight problematic areas with overlapping atoms, deviating binding angles, and more by incorporating data from established validation metrics. The main reason that the modeler wants to know the precise location of a problematic area so the molecular model can be fixed to address the issue rather than investigating correlations. In contrast to other existing tools, the available metrics are shown at the same time instead of individually. Additionally, local resolution information derived from the cryo-EM map is utilized to emphasize and filter areas where the underlying data provides more confidence.

The capabilities of visual environment were demonstrated with a few examples. Even though the presented datasets stem exclusively from cryo-EM, the framework is readily applicable to other modalities like X-ray crystallography, homology modeling, and structure prediction using AI and machine learning. Due to its extensible support for general per-residue validation metrics, additional metrics are easily included such as metrics particular to crystallography and pLDDT (predicted local distance difference test) in structure predictions. The same also applies to modeling structures using multi-modal workflows, the common factor being a validation based on per-residue metrics. Considering recent advances in the field, validation and verification will play an even bigger part when releasing new atomic models. This is especially true for the prediction of protein structures using machine learning, for example

AlphaFold [24], where many new structures are discovered in a short time.

For future work, we like to connect the framework directly to the PHENIX tool suite [14] since obtaining the various metrics from model validation is a cumbersome process for the time being. This would enable visual inspection and verification in sync with model validation, while a direct data transfer will simplify the multi-step preprocessing and also reach a wider audience.

ACKNOWLEDGMENTS

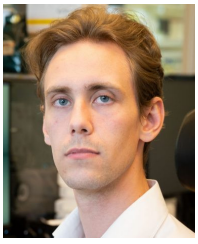
This work was supported through grants from the Excellence Center at Linköping and Lund in Information Technology (ELLIIT), Swedish e-Science Research Centre (SeRC), Swedish Research Council (VR) grant 2015-05462, and NIH grants R01EB023947, R01EB031872, the Knut and Alice Wallenberg Foundation through Grant KAW 2019.0024. The presented concepts have been developed and evaluated in the Inviwo framework (www.inviwo.org).

REFERENCES

[1] W. Kühlbrandt, "The resolution revolution," *Science*, vol. 343, no. 6178, pp. 1443–1444, 2014.
[2] X.-C. Bai, G. McMullan, and S. H. Scheres, "How cryo-EM is revolutionizing structural biology," *Trends in Biochemical Sciences*, vol. 40, no. 1, pp. 49–57, 2015.
[3] M. Carroni and H. R. Saibil, "Cryo electron microscopy to determine the structure of macromolecular complexes." *Methods*, vol. 95, pp. 78–85, Feb 2016.
[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 01 2000.
[5] C. L. Lawson, A. Patwardhan, M. L. Baker, C. Hryc, E. S. Garcia, B. P. Hudson, I. Lagerstedt, S. J. Ludtke, G. Pintilie, R. Sala, J. D. Westbrook, H. M. Berman, G. J. Kleywegt, and W. Chiu, "EMDataBank unified data resource for 3DEM," *Nucleic Acids Research*, vol. 44, no. D1, pp. D396–D403, 11 2015.
[6] K. Yamashita, C. M. Palmer, T. Burnley, and G. N. Murshudov, "Cryo-EM single-particle structure refinement and map calculation using *Servalcat*," *Acta Crystallographica Section D*, vol. 77, no. 10, pp. 1282–1291, Oct 2021.
[7] A. Kucukelbir, F. J. Sigworth, and H. D. Tagare, "Quantifying the local resolution of cryo-EM density maps," *Nature Methods*, vol. 11, no. 1, pp. 63–65, 2014.
[8] M. van Heel and M. Schatz, "Fourier shell correlation threshold criteria," *Journal of Structural Biology*, vol. 151, no. 3, pp. 250–262, 2005.
[9] V. Tobiasson and A. Amunts, "Ciliate mitoribosome illuminates evolutionary steps of mitochondrial translation," *Elife*, vol. 9, p. e59264, 2020.
[10] M. G. Prisant, C. J. Williams, V. B. Chen, J. S. Richardson, and D. C. Richardson, "New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink "waters," and NGL Viewer to recapture online 3D graphics," *Protein Science*, vol. 29, no. 1, pp. 315–329, 2020.
[11] B. A. Barad, N. Echols, R. Y.-R. Wang, Y. Cheng, F. DiMaio, P. D. Adams, and J. S. Fraser, "EMRinger: side chain–directed model and map validation for 3D cryo-electron microscopy," *Nature Methods*, vol. 12, no. 10, pp. 943–946, 2015.
[12] E. Ramírez-Aportela, D. Maluenda, Y. C. Fonseca, P. Conesa, R. Marabini, J. B. Heymann, J. M. Carazo, and C. O. S. Sorzano, "FSC-Q: a CryoEM map-to-atomic model quality validation based on the local Fourier shell correlation." *Nature Communications*, vol. 12, no. 1, p. 42, Jan 2021.
[13] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 1, pp. 12–21, 2010.
[14] D. Liebschner, P. V. Afonine, M. L. Baker *et al.*, "Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix," *Acta Crystallographica Section D*, vol. 75, no. 10, pp. 861–877, Oct 2019.
[15] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, "Features and development of Coot," *Acta Crystallographica Section D*, vol. 66, no. 4, pp. 486–501, Apr 2010.
[16] W. DeLano, "PyMOL: An open-source molecular graphics tool," *CCP4 Newsletter On Protein Crystallography*, vol. 40, 2002, pymol.sourceforge.net, (Online, Dec 2021).
[17] Y. Cheng, N. Grigorieff, P. A. Penczek, and T. Walz, "A primer to single-particle cryo-electron microscopy," *Cell*, vol. 161, no. 3, pp. 438–449, 2015.
[18] R. F. Thompson, M. Walker, C. A. Siebert, S. P. Muench, and N. A. Ranson, "An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology," *Methods*, vol. 100, pp. 3–15, 2016.
[19] S. H. W. Scheres, "RELION: implementation of a Bayesian approach to cryo-EM structure determination," *Journal of Structural Biology*, vol. 180, pp. 519–30, Dec 2012.
[20] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker, "cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination," *Nature Methods*, vol. 14, pp. 290–296, Mar 2017.
[21] D. Kimanius, B. O. Forsberg, S. H. Scheres, and E. Lindahl, "Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2," *eLife*, vol. 5, p. e18722, nov 2016.
[22] J. de la Rosa-Trevín, A. Quintana, L. del Cano, A. Zaldívar *et al.*, "Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy," *Journal of Structural Biology*, vol. 195, no. 1, pp. 93–99, 2016.
[23] T. Burnley, C. M. Palmer, and M. Winn, "Recent developments in the CCP-EM software suite," *Acta Crystallographica Section D*, vol. 73, no. 6, pp. 469–477, Jun 2017.
[24] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.
[25] M. A. Cianfrocco, M. Wong-Barnum, C. Youn, R. Wagner, and A. Leschziner, "COSMIC2: A science gateway for cryo-electron microscopy structure determination," in *Practice and Experience in Advanced Research Computing (PEARC)*, no. 22, 2017, pp. 1–5.
[26] M. Gupta, C. M. Azumaya, M. Moritz, S. Pourmal *et al.*, "CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes," *bioRxiv*, 2021. [Online]. Available: https://www.biorxiv.org/content/10.1101/2021.05.10.443524v1
[27] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations," *Journal of Molecular Biology*, vol. 7, no. 1, pp. 95–99, 1963.
[28] V. B. Chen, I. W. Davis, and D. C. Richardson, "KiNG (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program," *Protein science*, vol. 18, no. 11, pp. 2403–2409, 2009.
[29] C. L. Lawson, A. Kryshtafovych, P. D. Adams, P. V. Afonine, M. L. Baker, B. A. Barad, P. Bond, T. Burnley, R. Cao, J. Cheng *et al.*, "Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge," *Nature Methods*, vol. 18, no. 2, pp. 156–164, 2021.
[30] W. Humphrey, A. Dalke, K. Schulten *et al.*, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.
[31] A. Herraez, "Biomolecules in the computer: Jmol to the rescue," *Biochemistry and Molecular Biology Education*, vol. 34, no. 4, pp. 255–261, 2006.
[32] R. Skånberg, M. Linares, C. König, P. Norman, D. Jönsson, I. Hotz, and A. Ynnerman, "VIA-MD: Visual interactive analysis of molecular dynamics," in *EG Workshop on Molecular Graphics and Visual Analysis of Molecular Data*, 2018, pp. 19–27.
[33] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera - a visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.
[34] E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, and T. E. Ferrin, "UCSF ChimeraX: Structure visualization for researchers, educators, and developers," *Protein science*, vol. 30, pp. 70–82, Jan 2021.

[35] G. Reina and T. Ertl, "Hardware-accelerated glyphs for mono- and dipoles in molecular dynamics visualization," in *EG/IEEE VGTC Symposium on Visualization*, 2005, pp. 177–182.
[36] B. Kozlíková, M. Krone, M. Falk, N. Lindow, M. Baaden, D. Baum, I. Viola, J. Parulek, and H.-C. Hege, "Visualization of biomolecular structures: State of the art revisited," *Computer Graphics Forum*, vol. 36, no. 8, pp. 178–204, 2017.
[37] N. F. Fernandez, G. W. Gundersen, A. Rahman, M. L. Grimes, K. Rikova, P. Hornbeck, and A. Ma'ayan, "Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data," *Scientific Data*, vol. 4, no. 1, p. 170151, Oct 2017.
[38] N. Gehlenborg, J. Dietzsch, and K. Nieselt, "A framework for visualization of microarray data and integrated meta information," *Information Visualization*, vol. 4, no. 3, pp. 164–175, 2005.
[39] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 178–192, 04 2012.
[40] D. Jönsson, P. Steneteg, E. Sundén, R. Englund, S. Kottravel, M. Falk, A. Ynnerman, I. Hotz, and T. Ropinski, "Inviwo – a visualization system with usage abstraction levels," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 11, pp. 3241–3254, 2019.
[41] H. J. Bernstein, "Recent changes to RasMol, recombining the variants," *Trends in Biochemical Sciences*, vol. 25, no. 9, pp. 453–455, 2000.
[42] N. Desai, A. Brown, A. Amunts, and V. Ramakrishnan, "The structure of the yeast mitochondrial ribosome," *Science*, vol. 355, no. 6324, pp. 528–531, 2017.
[43] A. P. Boerema, S. Aibara, B. Paul, V. Tobiasson, D. Kimanius, B. O. Forsberg, K. Wallden, E. Lindahl, and A. Amunts, "Structure of the chloroplast ribosome with chl-RRF and hibernation-promoting factor," *Nature plants*, vol. 4, no. 4, pp. 212–217, 2018.

Martin Falk Martin Falk is an Associate Professor in the Scientific Visualization Group at Linköping University. He received his Ph.D. degree (Dr.rer.nat.) from the University of Stuttgart in 2013. His research interests include large-scale volume rendering, visualizations in the context of pathology, systems biology and cryo-EM, large spatio-temporal data, topological analysis, glyph-based rendering, and GPU-based simulations.



Victor Tobiasson Victor Tobiasson received his Ph.D. in structural biology in 2022 from Stockholm University for his work on the evolution of mitochondrial ribosomes. His research interests include the effects of neutral evolution and biased variation on the long term generation of biological complexity. He specializes in the collection, processing, modeling, validation and interpretation of cryo-EM data. Currently he is working at the NCBI/NLM/NIH studying eukaryogenesis and the origin of the mitochondrial ribosome in the group of Eugene Koonin.

Alexander Bock Alexander Bock is an Assistant Professor at Linköping University and the Development Lead for the astrovisualization software OpenSpace. He uses visualization techniques to create applications that can make difficult data, for example CT/MRI scans or astronomical datasets,easier to understand for a specific target audience, such as domain experts or the general public.

Charles Hansen Charles D. Hansen received the Ph.D. degree in computer science from the University of Utah, in 1987. He is the Peter Wallenberg Jr. Guest Professor of Visual Science Communication at Linköping University, Norrköping, Sweden and a Distinguished Professor, Emeritus at the University of Utah. From 1989 to 1997, he was a technical staff member in the Advanced Computing Laboratory (ACL), Los Alamos National Laboratory, where he formed and directed the visualization efforts in the ACL. He was a Bourse de Chateaubriand postdoc fellow at INRIA, Rocquencourt, France, in 1987 and 1988. His research interests include large-scale scientific visualization and computer graphics. He is a fellow of the IEEE.

Anders Ynnerman Anders Ynnerman received a Ph.D. 1992 in physics from Gothenburg University. During the early 90's he was doing research at Oxford University, UK, and Vanderbilt University, USA. In 1996 he started the Swedish National Graduate School in Scientific Computing, which he directed until 1999. From 1997 to 2002 he directed the Swedish National Supercomputer Centre and from 2002 to 2006 he directed the Swedish National Infrastructure for Computing (SNIC). Since 1999 he is holding a chair in scientific visualization at Linköping University and he is the director of the Norrköping Visualization Center – C, which currently constitutes one of the main focal points for research and education in computer graphics and visualization in the Nordic region. The center also hosts a public arena with large scale visualization facilities. He is also one of the co-founders of the Center for Medical Image Science and Visualization (CMIV).