

Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps

Abhishek Singharoy^{1††}, Ivan Teo^{1,2†}, Ryan McGreevy^{1†}, John E Stone¹, Jianhua Zhao³, Klaus Schulten^{1,2*}

¹Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, United States; ²Department of Physics, University of Illinois at Urbana-Champaign, Urbana, United States; ³Department of Biochemistry and Biophysics, University of California San Francisco School of Medicine, San Francisco, United States

Abstract Two structure determination methods, based on the molecular dynamics flexible fitting (MDFF) paradigm, are presented that resolve sub-5 Å cryo-electron microscopy (EM) maps with either single structures or ensembles of such structures. The methods, denoted cascade MDFF and resolution exchange MDFF, sequentially re-refine a search model against a series of maps of progressively higher resolutions, which ends with the original experimental resolution. Application of sequential re-refinement enables MDFF to achieve a radius of convergence of ~25 Å demonstrated with the accurate modeling of β-galactosidase and TRPV1 proteins at 3.2 Å and 3.4 Å resolution, respectively. The MDFF refinements uniquely offer map-model validation and B-factor determination criteria based on the inherent dynamics of the macromolecules studied, captured by means of local root mean square fluctuations. The MDFF tools described are available to researchers through an easy-to-use and cost-effective cloud computing resource on Amazon Web Services.

DOI: [10.7554/eLife.16105.001](https://doi.org/10.7554/eLife.16105.001)

*For correspondence: singharo@illinois.edu (AS); kschulte@ks.uiuc.edu (KS)

†These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.

Funding: See page 21

Received: 16 March 2016

Accepted: 06 July 2016

Published: 07 July 2016

Reviewing editor: Axel T Brunger, Howard Hughes Medical Institute, Stanford University, United States

© Copyright Singharoy et al. This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

Introduction

Cryo-electron microscopy (cryo-EM) has evolved into one of the most effective structure determination tools in modern day structural biology, achieving in recent years resolutions rivalling those of X-ray crystallography or NMR spectroscopy (Cheng, 2015). Furthermore, cryo-EM based structure determination overcomes two major bottlenecks faced in traditional X-ray crystallography, namely, the arduous task of preparing well-ordered crystals of macromolecules (Unger, 2002), and the more fundamental problem with capturing these molecules in unphysiological states as a result of crystal contacts (Neutze et al., 2015). Consequently, cryo-EM provides a natural way of resolving the structures of large macromolecular complexes.

Historically, computational methods were required to bridge the resolution gap between crystallography and cryo-EM to produce atomic-resolution models of biomolecular complexes. Various real-space refinement methods that combine crystallographic structures and cryo-EM densities for structure determination have been developed, including DireX (Schröder et al., 2007), Flex-EM (Topf et al., 2008), Rosetta (DiMaio et al., 2015), FRODA (Jolley et al., 2008), Phenix real space refinement (Afonine et al., 2013), and Molecular Dynamics Flexible Fitting (MDFF) (Trabuco et al., 2008, 2009; McGreevy et al., 2016).

MDFF, in particular, has proven to be an extremely successful refinement method as evidenced by its numerous applications (Goh et al., 2015; McGreevy et al., 2016) ranging from the intricate

eLife digest To understand the roles that proteins and other large molecules play inside cells, it is important to determine their structures. One of the techniques that researchers can use to do this is called cryo-electron microscopy (cryo-EM), which rapidly freezes molecules to fix them in position before imaging them in fine detail.

The cryo-EM images are like maps that show the approximate position of atoms. These images must then be processed in order to build a three-dimensional model of the protein that shows how its atoms are arranged relative to each other. One computational approach called Molecular Dynamics Flexible Fitting (MDFF) works by flexibly fitting possible atomic structures into cryo-EM maps. Although this approach works well with relatively undetailed (or 'low resolution') cryo-EM images, it struggles to handle the high-resolution cryo-EM maps now being generated.

Singharoy, Teo, McGreevy et al. have now developed two MDFF methods – called cascade MDFF and resolution exchange MDFF – that help to resolve atomic models of biological molecules from cryo-EM images. Each method can refine poorly guessed models into ones that are consistent with the high-resolution experimental images. The refinement is achieved by interpreting a range of images that starts with a 'fuzzy' image. The contrast of the image is then progressively improved until an image is produced that has a resolution that is good enough to almost distinguish individual atoms.

The method works because each cryo-EM image shows not just one, but a collection of atomic structures that the molecule can take on, with the fuzzier parts of the image representing the more flexible parts of the molecule. By taking into account this flexibility, the large-scale features of the protein structure can be determined first from the fuzzier images, and increasing the contrast of the images allows smaller-scale refinements to be made to the structure.

The MDFF tools have been designed to be easy to use and are available to researchers at low cost through cloud computing platforms. They can now be used to unravel the structure of many different proteins and protein complexes including those involved in photosynthesis, respiration and protein synthesis.

DOI: [10.7554/eLife.16105.002](https://doi.org/10.7554/eLife.16105.002)

ribosomal machinery (Villa et al., 2009; Trabuco et al., 2011; Frauenfeld et al., 2011; Wickles et al., 2014) to a host of non-enveloped viruses (Zhao et al., 2013). So far this success has been limited to structure determination from typically low-resolution cryo-EM maps in the 7–25 Å range which, indeed, represented the state-of-the-art at the time of MDFF's inception (Trabuco et al., 2008). However, seminal advances in detection hardware and programs over the past three years (Li et al., 2013; Milazzo et al., 2011) have enabled now the routine availability of high-resolution (<5 Å) EM maps for a range of biological systems including ion channels (Liao et al., 2013), enzymes (Bartesaghi et al., 2014, 2015), membrane fusion machinery (Zhao et al., 2015), and key functional components of the ribosome (Fischer et al., 2015; Brown et al., 2015).

High-resolution maps pose an imminent challenge to the traditional map-guided structure determination methods as the maps now characterize near-atomic scale features, the interpretation of which requires extremely precise structure building and validation protocols (DiMaio et al., 2015). For example, conformation of the protein sidechains, which are more flexible than the backbone, are now discernible within the maps and, thus, require precise modeling of the dihedral angles up to C_β atoms while also respecting the map boundaries (Barad et al., 2015).

In order to produce atomic models with correct backbone and sidechain geometries, as well as minimal potential energy, structure determination tools must be augmented with chemically accurate force fields and exhaustive search algorithms respecting density constraints. Inspired by crystallographic modeling techniques, where such structure-building requirements have already been addressed for the resolution of 3–5 Å diffraction data (DiMaio et al., 2013; Murshudov et al., 2011; McGreevy et al., 2014), tools such as Rosetta have introduced Monte Carlo simulation-based segment building and refinement protocols with heuristic force fields (DiMaio et al., 2015), to handle high-resolution EM maps. Other notable automated model-building tools that can be used for

the refinement of high-resolution EM maps include Buccaneer (Cowtan, 2006), ARP/wARP (Langer et al., 2008), and Moulder (Topf et al., 2006).

Driven by a vision to extend the capabilities of flexible fitting approaches (Topf et al., 2008; Trabuco et al., 2008; Tama et al., 2004; Suhre et al., 2006; Kovacs et al., 2008; Wu et al., 2013) for addressing high-resolution maps, two new MDFF methods are introduced here. These methods, denoted cascade MDFF (cMDFF) and resolution exchange MDFF (ReMDFF), augment the traditional MDFF method (Trabuco et al., 2008, 2009; McGreevy et al., 2016) (called direct MDFF henceforth) with enhanced conformational sampling techniques, namely simulated annealing (Brünger, 1988) and replica exchange molecular dynamics (Sugita and Okamoto, 1999). The central idea behind the techniques introduced is to fit a search model sequentially to a series of maps of progressively higher resolutions, ending with the original experimental resolution; all but the last in the series are computationally blurred lower-resolution derivatives of the original map, so that larger-scale features of the structure are determined first by fitting to the blurred densities, and smaller-scale refinements are performed subsequently during the fitting to higher-resolution densities. Altogether, this treatment enables a richer conformational sampling of the model within the map than direct MDFF, thereby allowing accurate modeling of the global and local structural features from the map; a similar treatment has previously been employed to increase the radius of convergence of MDFF protocols, but with crystallographic data (Singharoy et al., 2015).

The cMDFF and ReMDFF methods are demonstrated for structure analysis based on 3.2-Å and 3.4-Å resolution maps of β -galactosidase (Bartesaghi et al., 2014) and the TRPV1 channel (Liao et al., 2013), respectively. The two methods were found to resolve atomic structures with accuracy greater than that of direct MDFF and comparable to that of Rosetta, even with poor choices of search models. The accuracy is evaluated in terms of the quality of fit measured through global and local cross-correlations (GCC and LCC), integrated Fourier shell coefficients (iFSC), and EMRinger scores (Barad et al., 2015), as well as in terms of the quality of structural integrity measures like MolProbity (Chen et al., 2010).

In the second part of the present study we establish that structural flexibility, as measured by root mean square fluctuations or RMSF within the MDFF simulation, provides an ensemble-based indicator of local and overall resolution of a map offering, thus, a quality measure of an EM map based on the inherent dynamics of the imaged macromolecule. In line with this new finding, RMSF values are shown to provide a physical basis for the determination of optimal sharpening B-factors that maximize the signal-to-noise ratio within a map. These B-factors are determined at three different levels of model description: whole-system, per-domain, and per-residue.

Finally, use of the ReMDFF method on cloud computing platforms is discussed. Cloud computing is now a highly suitable approach for computational biology and can be employed for large-scale scientific computing, data analysis, and visualization tasks. For example, Amazon Web Services has been previously demonstrated to be a low-cost cloud computing platform for processing cryo-EM data (Cianfrocco et al., 2015). We demonstrate now the usage of Amazon Web Services, highlighting the platform's capability for rapidly fitting structures to EM density with ReMDFF. The web-interface makes it readily possible for experimental groups around the world to deploy MDFF in an easy and economical way, bypassing the need for their own staff, software, and hardware resources.

Results

In the following section, we first describe the methodological advances achieved within cMDFF and ReMDFF for the resolution of sub-5 Å maps. Search model preparation, refinement, and structure validation protocols based on these advances are subsequently demonstrated for five exemplary protein complexes that were chosen based on the availability of high-resolution (3–5 Å) EM maps and atomic structures. Finally, the performance of ReMDFF on Amazon's cloud computing platforms is described, demonstrating that our MDFF software offers an efficient web-based resource for structure determination from EM maps.

Simulation concept

In direct MDFF, an initial atomic structure is subjected to an MD simulation with an additional potential energy term V_{EM} that is proportional to the sign-inverse of the EM map. Through V_{EM} , steering

forces locally guide atoms towards high-density regions, thereby fitting the structure to the map (see Materials and methods).

The equilibrium structure obtained in the simulation represents a global minimum in V_{EM} . For maps in the low resolution range (6–15 Å), this global minimum is broad, accommodating an ensemble of conformations defined by the overall shape of the macromolecule (Trabuco *et al.*, 2008, 2011). In contrast, at the mid-resolution range of 4–6 Å, densities corresponding to the backbones become discernible, and at sub-4 Å resolutions, even sidechains can be resolved. At such high resolutions, V_{EM} now features multiple proximal local minima which correspond to recurring spatial patterns within a macromolecule, such as helices aligned in parallel or strands in a β -sheet. As shown in **Figure 1**, the energy barriers separating these local minima are typically twice as high as those in the case of low-resolution maps. The existence of such potential minima in high-resolution maps exposes MDFF to a long-known weakness of traditional MD-based algorithms, namely entrapment of the fitted structure within undesired local minima instead of reaching the global minimum of V_{EM} . Not unexpectedly, therefore, direct MDFF yields structurally poor or functionally irrelevant models with high-resolution EM maps (**Figure 2—figure supplement 1**) (DiMaio *et al.*, 2015).

To mitigate this weakness in direct MDFF, we introduce cMDFF and ReMDFF. In cMDFF, the structure is fitted, in a series of MDFF simulations, to maps of gradually increasing resolution. First, the experimental map is smoothed by applying a series of Gaussian blurs with increasing half-widths, σ , to obtain a set of theoretical maps with gradually decreasing resolution; $\sigma = 0$ Å corresponds to the experimental map, and $\sigma > 0$ Å corresponds to a smoothed one (see Materials and methods). Illustrated in **Figure 1**, the density-dependent potential derived from the smoothest map (i.e. the one with the largest σ value) features a clear global minimum representing the large-scale structural features of the protein. Second, a search model is fitted to this map, allowing resolution of these large-scale features. Third, the resulting structure is employed as the search model for fitting to the next higher-resolution map in the series. These fitting and search model refreshment steps are repeated through the series of maps in order of decreasing σ , until the structure is finally fitted to the experimental map (see **Video 1** for TRPV1 refinement. An additional virtual reality version can be found at <https://www.youtube.com/watch?v=UwwVC6C9tw0>).

The gradual increase in map resolution over the course of the simulations allows the structure to explore a greater conformational space than in direct MDFF. The structure thus avoids entrapment within local minima of the density-dependent potential and is accurately fitted to the near-atomic density features of the experimental map while also resolving the larger scale features.

In ReMDFF, the cascade scheme is infused with a greater degree of automation. Multiple MDFF simulation replica are run in parallel with each replica fitting a model to a map of a specific resolution. Based on a Metropolis formula analogous to that of conventional replica exchange molecular dynamics simulations (Sugita and Okamoto, 1999), but now derived in terms of density and not temperature (see Materials and methods), the models are exchanged at regular time intervals between maps from neighboring pairs of replica. Stepwise improvements in fit occur during exchanges between a poorly fitted model at a high resolution with a well-fitted model at a lower resolution. This well-fitted model is further refined against the high-resolution map until convergence is reached, and exchange between the chosen resolutions ceases. Further details are described in Materials and methods.

ReMDFF has advantages over cMDFF both in terms of efficiency and automation as it can take advantage of modern parallel computing hardware and the powerful and adaptive replica exchange interface of NAMD (Jiang *et al.*, 2014). Nonetheless, as presented in the following, both cMDFF and ReMDFF outperform direct MDFF in quality and speed across a range of high-resolution examples.

Search model preparation and refinement

In an initial proof-of-principle computation, cMDFF and ReMDFF were applied to fit a structure of carbon monoxide dehydrogenase to a 3-Å synthetic density map. The same techniques were subsequently applied to obtain refined structures of two more protein systems, namely TRPV1 (Liao *et al.*, 2013) and β -galactosidase (Bartesaghi *et al.*, 2014), for which experimental densities of 3.4 Å and 3.2 Å resolution respectively are available. In each case, a direct MDFF simulation was also performed for the purpose of comparison. The MDFF-derived structures were then subjected to a model validation analysis, to evaluate the quality of the models with established protocols in the

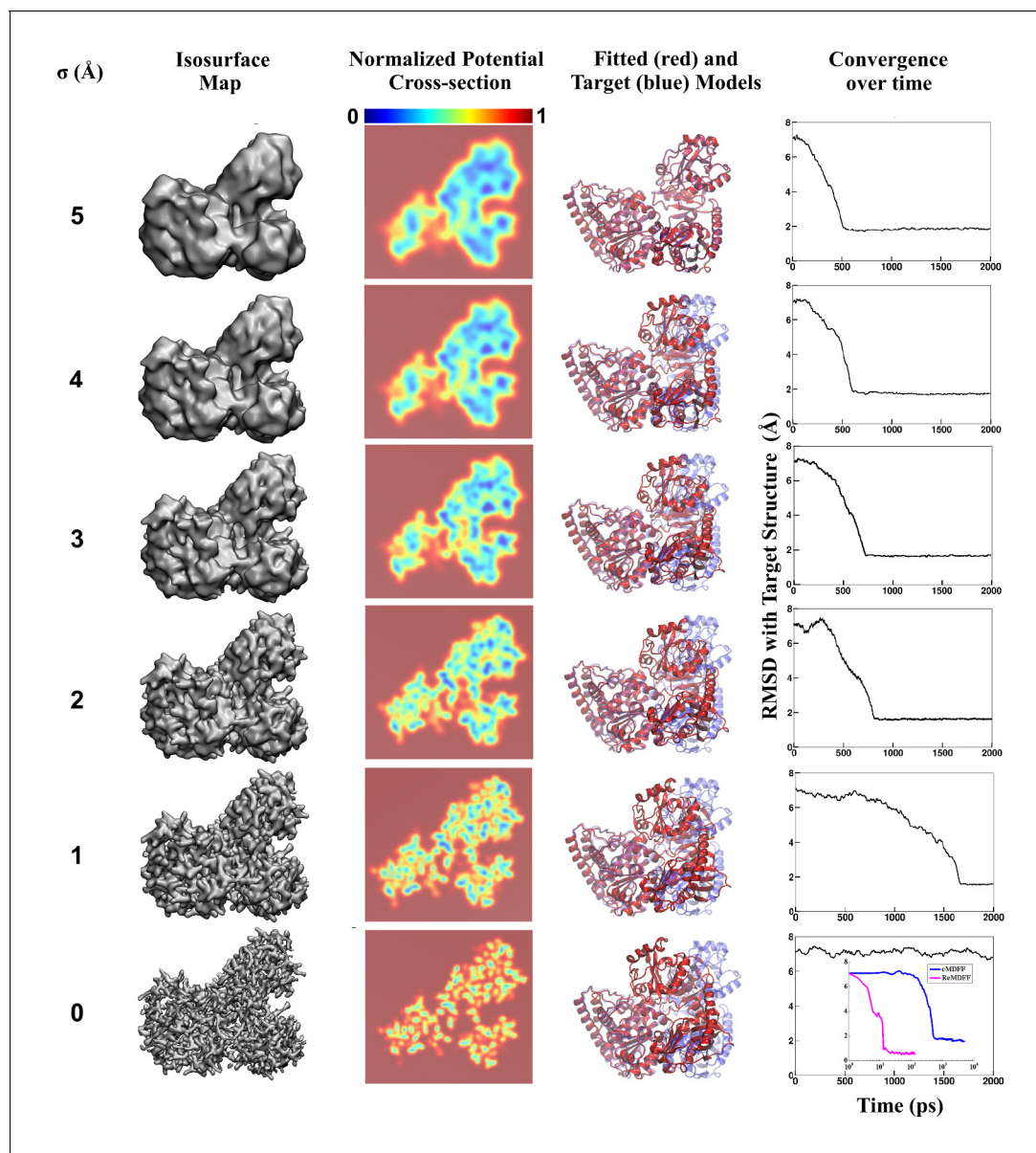
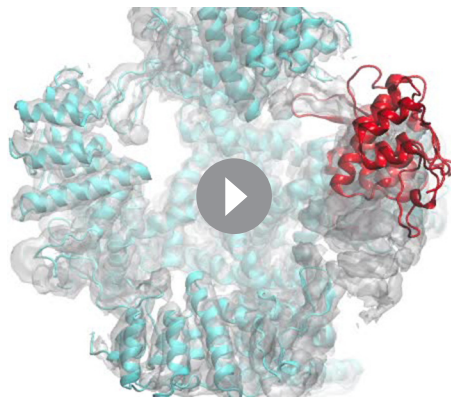


Figure 1. Visual summary of advanced MDFF methodology. A graphic table illustrating MDFF refinement of a model of carbon monoxide dehydrogenase using a high-resolution map. The map represents an open conformation while the initial search model was obtained through crystallography of a closed conformation. This search model was independently fitted, using direct MDFF, to individual members of a set of maps obtained by applying Gaussian blurs of various half-widths (σ , first column) to the experimental density. These maps are visualized as a 3D surface in the second column, while the resulting MDFF potentials V_{EM} are represented in cross-section in the third column. Notice the increase in number of contiguous density regions as σ increases. This increase in contiguity is manifested in the lowering of high V_{EM} barriers (red) for small σ values to low or flat energy profiles (blue) for larger σ values, as observed in the V_{EM} potential cross-sections. Reduced barrier heights allow the structure to explore the conformational space freely during fitting. The structure after 500 ps of fitting, shown in red, is superimposed on the known target structure, shown in blue, in the fourth column. The time evolution of RMSD with respect to the target during fitting is shown in the fifth column. The RMSD plots show that direct fitting to lower resolution maps requires fewer time steps to reach convergence. In fact, the structure never becomes less deviated than the initial 7-Å RMSD from the target in the direct MDFF of the highest-resolution map (i.e. in the absence of Gaussian blurring). The inset shows refinements of the same structure through cMDFF and ReMDFF employing the same set of maps. A clear improvement over direct MDFF is apparent, with convergence to within 1.7 Å and 1.0 Å of the target achieved within 1000 and 100 ps for cMDFF and ReMDFF respectively.

DOI: [10.7554/eLife.16105.003](https://doi.org/10.7554/eLife.16105.003)



Video 1. cMDFF Refinement of TRPV1.

DOI: [10.7554/eLife.16105.004](https://doi.org/10.7554/eLife.16105.004)

cryo-EM field. Additional examples chosen for this analysis include γ -secretase at 4.5 Å (Lu et al., 2014) and 3.4-Å (Bai et al., 2015), β -galactosidase at 2.2-Å (Bartesaghi et al., 2015) and the proteasome at 3.3 Å (Li et al., 2013) resolution. Comparisons between the direct and advanced MDFF protocols, and wherever possible, with other available fitting techniques, such as Rosetta (DiMaio et al., 2015), elucidate the general pros and cons of the flexible fitting strategy.

Proof of principle

The performance of cMDFF and ReMDFF was evaluated on a test system, carbon monoxide dehydrogenase, which exhibits a closed and an open conformation (Darnault et al., 2003). Both

these conformations have been crystallized, and are reported respectively in chains C and D of the PDB entry 1OAO. For our demonstration, the closed conformation (1OAO:chain C) was used as the search model, while the open one (1OAO:chain D) was the target.

First, a 3-Å resolution synthetic density map was constructed in Phenix (Adams et al., 2010), employing phases from the 1OAO structure and the associated diffraction data truncated at 3 Å. This map was then masked about chain D to yield a high-resolution envelope characterizing the open conformation. Assuming that the crystallographic model provides an accurate benchmark, the corresponding map for chain D determined here represents the best possible density data at 3 Å resolution that is experimentally attainable for the open conformation. Finally, through direct MDFF, cMDFF and ReMDFF, the search model constructed from chain C was fitted into this density to derive an atomic structure representing the open conformation. Both the cMDFF and ReMDFF refinements were performed for a set of six maps with σ values ranging from 5 to 0 Å at constant decrements of 1 Å.

Accuracy of the fitting protocols was evaluated by comparing the fitted chain C structures with the crystallographically reported target chain D model. Direct MDFF of the 3 Å synthetic map performed for 2 ns converged to a structure with an RMSD of 7 Å relative to the target model. In sharp contrast, the cMDFF- and ReMDFF-generated structures are within 1.7 Å and 1 Å RMSD of the target (see the inset of Figure 1). It is also noted that fitting to the lowest-resolution (i.e. one with $\sigma = 5$ Å) brings about an immediate decrease in RMSD from the target generating structures that are within 2 Å RMSD. Fitting of the structure to subsequent high-resolution maps brought the RMSD down to 1.0 Å.

The results demonstrate that the new protocols are capable of attaining well-fit structures where direct MDFF does not. In particular, one can think of the new protocols as extending the radius of

Table 1. β -galactosidase MDFF results. cMDFF and ReMDFF provide better fitted structures than direct MDFF according to various criteria. It is noteworthy that all structures refined by any form of MDFF display an improved MolProbity (Chen et al., 2010) score compared to the original de novo structure.

Structure	RMSD(Å)	EMRinger	iFSC1(Å)	iFSC2(Å)	MolProb.	GCC
de novo (Bartesaghi et al., 2014)	0.0	2.25	4.03	5.00	3.14	0.67
Refined de novo	0.6	4.23	4.19	5.20	1.23	0.68
Initial	7.7	0.24	0.14	0.15	1.49	0.48
Direct MDFF	3.7	2.31	2.11	2.74	1.38	0.56
cMDFF	0.7	3.16	4.22	5.22	1.37	0.67
ReMDFF	0.9	3.45	3.76	4.66	1.13	0.67

DOI: [10.7554/eLife.16105.014](https://doi.org/10.7554/eLife.16105.014)

convergence to at least 7 Å, rendering the fitting procedures less dependent on the quality of the starting structure.

Refinement of β -galactosidase

In a second test case, a search model was fitted into the 3.2 Å map (*Bartesaghi et al., 2014*) of β -galactosidase employing direct MDFF, cMDFF, and ReMDFF. Noting that the radius of convergence of the proposed MDFF protocols was at least 7 Å for the aforementioned test case, the initial search model was prepared such that it had an RMSD of 7 Å from the reported structure. This model was obtained by applying to the reported structure (obtained by de novo modeling within the EM map [*Bartesaghi et al., 2014*]) a high temperature MD protocol described in Appendix 1, Section 3 and, subsequently, choosing from the collection of trajectory structures one of RMSD 7 Å from the reported structure and with the lowest GCC with respect to the reported map (*Bartesaghi et al., 2014*) (*Figure 2—figure supplement 2a*).

Summarized in *Table 1*, the fitting results, in terms of quality of fit as well as model quality, are significantly better for cMDFF and ReMDFF than for direct MDFF: (i) RMSD of the fitted structure with respect to the reported de novo model is 0.7 Å and 0.9 Å for cMDFF and ReMDFF respectively, much lower than the 3.7 Å RMSD attained with direct MDFF (*Figure 2—figure supplement 3a*); (ii) EMRinger scores for cMDFF and ReMDFF are 3.16 and 3.45 respectively, higher than the 1.91 obtained for direct MDFF, implying accurate fitting of sidechains into the density; (iii) MolProbity scores are consistently small for all the flexible fitting techniques in part due to fewer, less severe steric clashes and fewer Ramachandran outliers (further detailed in *Table 2*); (iv) integrated FSC (iFSC2, corresponding to the range 3.4–10 Å on the FSC plot obtained as per Appendix 1 - Section 6), considered a more stringent measure of model quality than CC (*DiMaio et al., 2015*), attained higher values of 5.22 Å and 4.66 Å for cMDFF and ReMDFF, respectively, than 2.74 Å for direct MDFF. iFSC1, evaluated at the lower resolution range of 5–10 Å improves from 2.11 Å for direct MDFF to 4.22 Å and 3.76 Å for cMDFF and ReMDFF, respectively, showing a trend similar to that of iFSC2 corresponding to the high-resolution range; and (v) GCCs improved from an initial value of 0.48 to 0.56, 0.67 and 0.67 for direct, cMDFF, and ReMDFF protocols respectively. Similarly, typical residue LCC values improved from about 0 to greater than 0.80 (*Figure 2—figure supplement 4a* and *Figure 2—figure supplement 5*). Overall, cMDFF and ReMDFF refinements produce structures that interpret the 3.2-Å β -galactosidase map much more accurately than direct MDFF does.

Table 2. Structure quality indicators for β -galactosidase structures. β -galactosidase structures investigated in the present study were uploaded to the MolProbity server (<http://molprobity.biochem.duke.edu>) to extract the quantities presented below. The results show that the cMDFF- and ReMDFF-refined structures not only exhibit good measures of fit, but also improve the clash score and rotamer geometries, relative to the de novo and initial structures, while incurring only a small expense in Ramachandran statistics, bad angles, and C β deviations.

	de novo (<i>Bartesaghi et al., 2014</i>)	Refined de novo	Initial	Direct MDFF	cMDFF	ReMDFF
Clashscore	53.7	0.0	0.0	0.0	0.0	0.0
Poor rotamers (%)	11.6	3.8	4.2	3.0	4.4	1.37
Favored rotamers (%)	67.4	90.8	87.8	92.1	89.8	95.3
Ramachandran outliers (%)	0.2	0.7	2.7	3.0	1.6	2.7
Ramachandran favored (%)	97.4	95.8	91.1	91.1	94.4	90.9
MolProbity	3.14	1.23	1.49	1.38	1.37	1.13
C β deviations (%)	0.0	0.05	4.92	0.18	0.29	0.39
Bad bonds (%)	0.09	0.04	3.61	0.02	0.01	0.03
Bad angles (%)	0.03	0.60	3.98	0.63	0.49	0.37
RMS distance (Å)	0.007 (0.025%)	0.019 (0%)	0.035 (0.237%)	0.022 (0%)	0.019 (0%)	0.021 (0%)
RMS angle (degrees)	1.1 (0.009%)	2.2 (0.009%)	3.6 (1.177%)	2.4 (0.103%)	2.1 (0.018%)	2.3 (0.085%)
Cis prolines (%)	8.06	8.06	6.45	6.45	6.45	8.06
Cis non-prolines (%)	1.15	1.15	0.0	0.0	1.15	0.0

DOI: 10.7554/eLife.16105.015

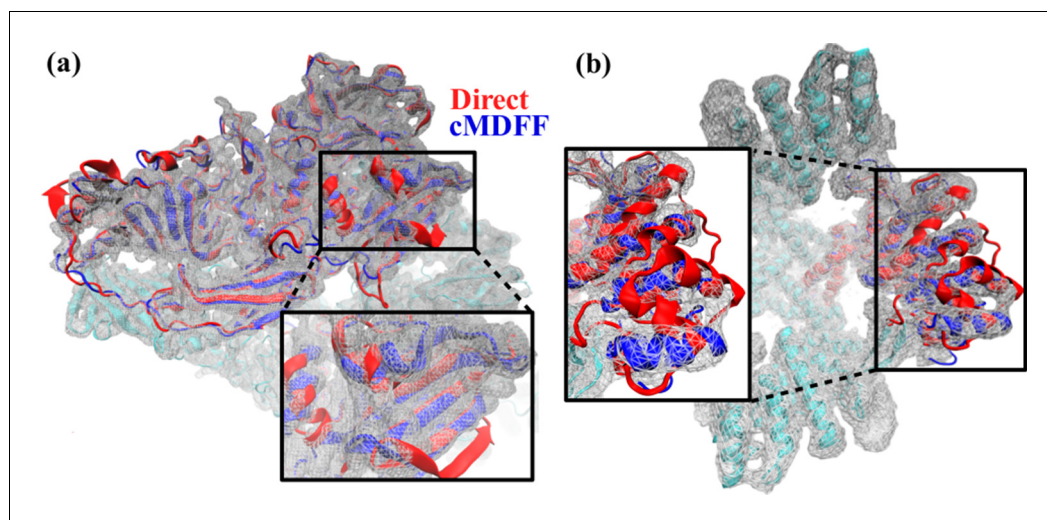


Figure 2. Comparison between cMDF and direct MDF fitted models. Models of (a) β -galactosidase and (b) TRPV1, obtained from cMDF (blue) and direct MDF (red) fitting simulations are superimposed. The cMDF-fitted models fit well into the high-resolution maps (grey) of each molecule, whereas the direct MDF models have become trapped in local minima that result in portions of the models protruding from the maps. ReMDF-fitted models are almost identical to those from cMDF and are therefore not shown.

DOI: [10.7554/eLife.16105.005](https://doi.org/10.7554/eLife.16105.005)

The following figure supplements are available for figure 2:

Figure supplement 1. Global cross-correlation as a measure of fit.

DOI: [10.7554/eLife.16105.006](https://doi.org/10.7554/eLife.16105.006)

Figure supplement 2. Comparison of initial models to target (published) models.

DOI: [10.7554/eLife.16105.007](https://doi.org/10.7554/eLife.16105.007)

Figure supplement 3. Convergence of cMDF, ReMDF, and direct MDF simulations.

DOI: [10.7554/eLife.16105.008](https://doi.org/10.7554/eLife.16105.008)

Figure supplement 4. Local cross-correlations during cMDF.

DOI: [10.7554/eLife.16105.009](https://doi.org/10.7554/eLife.16105.009)

Figure supplement 5. Local cross-correlations during direct MDF to refine de novo structures.

DOI: [10.7554/eLife.16105.010](https://doi.org/10.7554/eLife.16105.010)

Figure supplement 6. Equilibration of cMDF-refined model of β -galactosidase.

DOI: [10.7554/eLife.16105.011](https://doi.org/10.7554/eLife.16105.011)

Figure supplement 7. Residues of β -galactosidase fitted within density map.

DOI: [10.7554/eLife.16105.012](https://doi.org/10.7554/eLife.16105.012)

Figure supplement 8. FSC cross-validation plots.

DOI: [10.7554/eLife.16105.013](https://doi.org/10.7554/eLife.16105.013)

Figure 2a shows, visually, how the cMDF-derived structure differs from the direct MDF structure in terms of fit. In judging the RMSD values to the target model the reader is reminded that equilibrium MD simulations of a single structure at room temperature typically exhibit RMSD values relative to the initial structure or the average structure of about 3 Å; the same is true for β -galactosidase (see **Figure 2—figure supplement 6** in Appendix 1 - Section 7). Consequently, an RMSD of 0.7 Å of the cMDF/ReMDF-fitted model relative to the target implies a high-quality refinement. The high quality of this refinement is further supported by visualizations of accurate sidechain placements within the density, shown in **Figure 2—figure supplement 7**.

The cMDF- and ReMDF-refined structures were found to be comparable in every quality measure in comparison to the reported de novo structure (*Bartesaghi et al., 2014*); in fact, the overall Molprobrity and EMRinger scores are significantly better in cMDF and ReMDF. However, a closer look at the Molprobrity score (**Table 2**) reveals that even though cMDF vastly improves clash score and poor rotamers, it marginally increases the percentage of Ramachandran outliers and C β deviations relative to the de novo structure. Nonetheless, both cMDF and ReMDF improved structural

statistics with respect to the initial model (**Table 1**, third row) which was intentionally chosen to have a large deviation (RMSD of 7.7 Å) from the de novo structure.

Noting that the quality of MDFF output depends strongly on that of the search model, a second cMDFF simulation was also performed to refine the de novo structure within the reported map. The simulation, labeled 'refined de novo' in **Table 1**, yielded a structure that was superior in all the quality measures considered in comparison to the de novo structure as well as to the structures obtained from the various MDFF fittings of the other, 7.7 Å deviated initial model. A closer look at the Molprobity scores (**Table 2**) now reveals that not only are clash score and poor rotamers vastly improved, but the Ramachandran outliers and C β deviations are also comparable to the de novo structure.

A third cMDFF refinement was further performed with a search model of even lower structural quality (**Supplementary file 1E**, third row) compared to that of **Table 1** (third row). Structural statistics for the refinements from this lower-quality search model are provided in **Supplementary file 1F**. Comparison of the three cMDFF refinements starting with three search models of varying structural quality reveal that the poorer the secondary structure of the search model, measured in terms of higher percentages of rotamer and Ramachandran outliers, the worse the local structural statistics of the MDFF/cMDFF/ReMDFF-refined model. Surprisingly, large scale map-model validation measures, such as RMSD, GCC, or iFSC values, remain insensitive to such local discrepancies in the refined model. This insensitivity is apparent in the similarity of RMSD, GCC, and iFSC values of the three refined models of β -galactosidase (**Table 1** and **Supplementary file 1E**), which indeed feature very different number of Ramachandran outliers: 0.7% (**Table 2**-Refined de novo), 1.6% (**Table 2**-cMDFF) and 7.8% (**Supplementary file 1F**-cMDFF).

The high-precision of sub-5 Å maps demands correct assignment of secondary structure, which determines backbone geometry directly and sidechain conformations indirectly. In light of the initial model quality dependence of MDFF, it is advisable to begin refinement with an initial model of maximal secondary structure information, yet with minimal tertiary structure. Dependable initial secondary structure is also required since MD simulations have limited capability of recovering the structure of a protein fold if the search model begins with a random coil conformation (**Freddolino et al., 2010**). To this end, notwithstanding observed cMDFF improvements of the secondary structure (**Table 2** columns 4 vs. 6, and also demonstrated for γ -secretase [**Supplementary file 1D**]), moderate- to high-confidence homology models will serve as the most optimal starting point, as has been successfully shown for MDFF with low-resolution EM (**Chan et al., 2011; Noble et al., 2013; Wickles et al., 2014**) and crystallographic data (**McGreevy et al., 2014; Li et al., 2014**). Quantitatively speaking, employing the current example (β -galactosidase) and the following one (TRPV1), it is now shown that a search model which deviates from the target by an RMSD of 7–25 Å featuring Ramachandran outliers of ~3%, poor rotamers of 11–38%, overall Molprobity score of ~4 and EMRinger score of ~0.20 can be accurately refined against a sub-5 Å EM map with cMDFF or ReMDFF.

In terms of efficiency, the ReMDFF protocol exhibits the quickest convergence, arriving at steady state within 0.1 ns of simulation, whereas cMDFF requires around 0.8 ns. Both methods employed eleven maps with Gaussian blurs starting from a width of 5 Å and decreasing in steps of 0.5 Å towards the original reported map. To ensure that the cMDFF procedure did not over-fit the structures, cross-validation using EMRinger and FSC analysis was performed using half-maps from the EMD-5995 entry. iFSC and EMRinger values were found to be almost identical in both direct and cross comparisons. Details are provided in Appendix 1 - Section 11.

In addition to the MDFF simulations described so far for β -galactosidase, other simulations were performed to investigate in more detail the capabilities of MDFF. These simulations are described in Appendix 1 - Section 7. First, it was found that fitting of the C β atoms to the density is crucial for accurate placement of the sidechains. In agreement with prior EMRinger results (**Barad et al., 2015**), it is confirmed that MDFF placement of the backbone does not guarantee correct sidechain geometries, even with state-of-the-art CHARMM36 (**Klada et al., 2010**) force fields. Second, MD simulation of the cMDFF-fitted β -galactosidase model revealed that this model is indeed an excellent representation of the most probable structures of the thermodynamic ensemble that characterizes the 3.2 Å map.

Refinement of TRPV1

In the third test case, cMDFF and ReMDFF protocols were employed to fit an initial model to the 3.4-Å map (*Liao et al., 2013*) of the temperature-sensing protein TRPV1. The search model was prepared from the reported de novo structure through an interactive MD protocol described in Appendix 1 - Section 3. The model deviated from the de novo structure by an overall RMSD of 10 Å and locally, by about 25 Å in the vicinity of the ankyrin repeats represented by residues 199 to 430. This degree of deviation is in the ballpark of the lowest resolutions of usable EM maps and, therefore, represents the upper limit of uncertainty between a search model and the fitted structure that MDFF can still reconcile. Having to address an RMSD of 25 Å between search and target models, the present example represents an extreme test case for judging the radius of convergence of the proposed MDFF methods.

Fitting results for TRPV1, described in Appendix 1 - Section 8, were significantly better for cMDFF and ReMDFF than for direct MDFF, but now with much poorer search models than those employed for the β -galactosidase refinements. For example, cMDFF and ReMDFF refinements produced structures that interpret the 3.4-Å TRPV1 map within an RMSD of 2.4 Å and 2.5 Å from the target de novo model, much more accurately than does direct MDFF which converges to structures at an RMSD of 7.9 Å.

The cMDFF- and ReMDFF-obtained structures were found to be better in every overall quality metric in comparison to the de novo structure. **Figure 2b** illustrates the contrast in fit between the cMDFF and direct MDFF-derived structures. To observe the effect of MDFF on a substantially well-fitted initial structure, a direct MDFF simulation was also performed to refine the de novo structure within the reported map. The simulation yielded a structure that was comparable in all the quality measures considered to the structures obtained from cMDFF and ReMDFF (Appendix 1 - Section 8). However, the TRPV1 fitting results show that cMDFF and ReMDFF can have a radius of convergence as high as 25 Å in RMSD, whereas direct MDFF requires at the outset a well-fitted structure to deliver a satisfactory model. Also, since the number of Ramachandran outliers were minimal in both the 25 Å-deviated and de novo initial models, the cMDFF, ReMDFF, and direct MDFF-refined models exhibited low percentages of Ramachandran outliers, as reflected in **Supplementary file 1B**.

As was observed already for β -galactosidase, the ReMDFF protocol exhibited the quickest convergence, arriving at steady state within 0.02 ns of simulation, whereas cMDFF required around 0.27 ns. Both methods employed six maps with Gaussian blurs starting from a width of $\sigma = 5$ Å and decreasing in steps of 1 Å to the reported $\sigma = 0$ Å map. Cross-validation with half-maps was also performed on the cMDFF structure, as per the β -galactosidase simulations, to ensure that it was not over-fitted. As in the case of β -galactosidase, iFSC and EMRinger scores for direct and cross comparisons were similar. FSC analysis results are described in Appendix 1 - Section 11.

A separate set of model validation analyses was performed on the well-resolved TM portion of TRPV1 to pursue a direct comparison of a MDFF refined model with one from Rosetta (*Barad et al., 2015*). As reported in Appendix 1 - Section 9, MDFF produced results comparable to those of Rosetta when all the heavy atoms are coupled to the density: though the EMRinger score is marginally lower relative to that from Rosetta, the GCC and iFSCs are higher for MDFF; also MDFF provides a marginally higher MolProbity score. Altogether, major discrepancies between Rosetta and direct MDFF that were reported for the high-resolution EM maps (*DiMaio et al., 2015*) are now absent when employing the cMDFF and ReMDFF protocols, even with poorer choices of search models than those used with Rosetta. Thus, cMDFF and ReMDFF enable flexible fitting techniques to pursue resolution of structures within state-of-the-art maps obtained via cryo-EM.

Model validation

An EM density map represents a thermodynamic ensemble of atomic conformations (*Schröder et al., 2007; Brunger et al., 2012; Schröder et al., 2010*). Conventionally, however, only a single model representing a best fit to the map is reported. One may ask how statistically representative a single model can be. To quantify the deviation of a fitted model from the rest of an ensemble of simulated molecules, root mean square fluctuation (RMSF) of the model relative to the ensemble-averaged structure was computed during an MDFF refinement simulation employing the protocol described in Methods. In the following, the RMSF of a fitted model is first shown to be indicative of the quality of fit of the model, as well as to represent the degree of natural

conformational variation exhibited within the thermodynamic ensemble underlying the map. Second, the RMSF values are found to correlate both locally and globally with the resolution of an EM map, providing an interpretation of map quality based on the inherent (i.e., natural) dynamics of the macromolecule under observation. Finally, RMSF values are also employed to identify optimal B-factor values for the sharpening of a map. Altogether, the results of the present study demonstrate that RMSF of a fitted model during an MDFF refinement provides valuable information on the model.

RMSF and quality of fit

The relationship between RMSF values and quality of model fit is demonstrated for the cMDFF refinement of β -galactosidase at 3.2 Å resolution. The initial conformation is a poor fit of the map, characterized by low values of GCC, LCC, and iFSC (the row containing 'initial' structure in [Table 1](#)). Such conformations belong to a diverse ensemble of poorly fit structures, explored by the search model in the early phase of the fitting, that gives rise to high initial RMSF values shown in [Figure 4—figure supplement 1](#). In the ending phase of the refinement, well-fitted structures are obtained with improved GCC, LCC, and iFSC (cMDFF row of [Table 1](#)). Owing to the high resolution (3.2 Å) of the EM density map, the population of these well-fitted structures is much smaller than that of partially-fitted structures. Thus, as can be observed in [Figure 4—figure supplement 1](#), the converged ensemble of the well-fitted structures exhibit much smaller fluctuations than the initial one. Low RMSF values for the fitted structure indicate therefore, that (i) the structure has been modeled unambiguously within the map, and (ii) the structure can be regarded as representative of the ensemble underlying the 3.2-Å β -galactosidase map; since per-residue fluctuations about the fitted structure ([Figure 4—figure supplement 1](#)) are less than 1 Å, only marginal backbone and sidechain variations within the ensemble arise ([Singharoy et al., 2013](#)).

RMSF and quality of map

Apart from representing the quality of fit, RMSF values monitored during an MDFF simulation correlate closely with the overall and local resolution of an EM map. Even though high-resolution cryo-EM data are becoming increasingly obtainable, resolution is not always uniform throughout a map. For example, [Figure 3](#) reflects the variation in local resolution of map regions corresponding to residues of β -galactosidase and TRPV1. Conformational flexibility can cause heterogeneity in the cryo-EM data ([Leschziner and Nogales, 2007](#)), producing local resolutions lower than that of the overall map.

Local resolution analysis ([Kucukelbir et al., 2014](#)) can be especially important for determining the parts of a high-resolution map that realistically contain side chain information and the parts that do not, preventing over-interpretation of the latter. MDFF protocols can be adjusted to account for such local variations and better inform the process of model validation. For example, if analysis shows that certain residues reside in low-resolution regions of the density, the per-atom weighting factor applied to the forces derived from the density can be lowered.

The RMSF-resolution correlation is found to hold in the cases of TRPV1 (PDB 3J5P), 2.2-Å β -galactosidase (PDB 5A1A), γ -secretase (PDB 5A63 and 4UPC), and the T20S Proteasome (PDB 3J9I). Generally, the lower the resolution of the map, the higher the corresponding overall RMSF during MDFF simulations. For example, the overall RMSF during MDFF of the 4.5-Å γ -secretase model and map is greater than that of the 3.4-Å model and map which, in turn, is greater than that of the 2.2-Å map and model of β -galactosidase (see RMSF labels on the upper column of [Figure 4](#)). The correlation between map resolution and model RMSF extends further to local features within the density. In [Figure 4](#) (upper row), RMSFs of atoms plotted against local resolutions of the corresponding map regions display linear correlation between the two quantities. Again, higher RMSF indicates lower local resolution.

The physical basis for considering RMSF values of a group of fitted atoms during an MDFF refinement as an indicator of the map resolution follows from the linear correlation of these values with RMSF of the same set of atoms during unbiased MD ([Figure 4](#), bottom row). Noting that the RMSF value during unbiased MD simulations reflects flexibility ([Karplus, 1990](#)), this linear correlation clearly establishes the dependence of RMSF during MDFF refinement on the inherent flexibility of the macromolecule, at least for our four demonstration systems (see Appendix 1 - Section 2 for MD simulation details). Since the flexibility of a molecule during the imaging process contributes to the

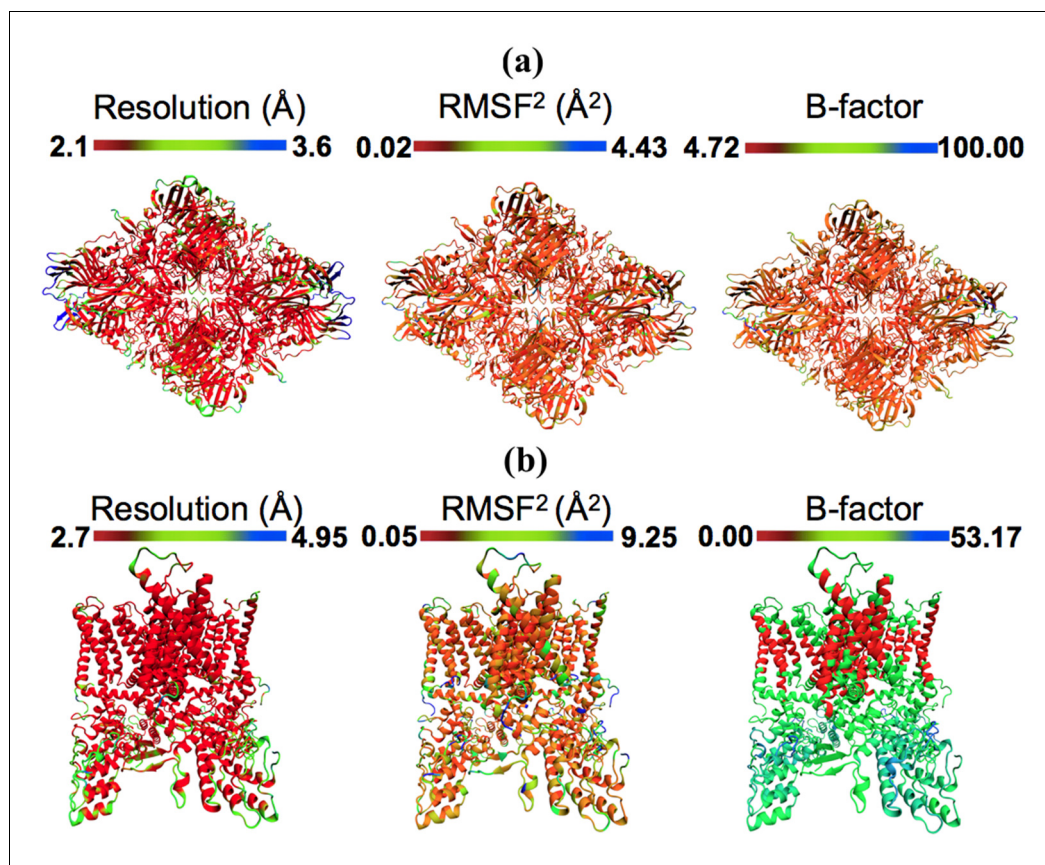


Figure 3. Models colored by local resolution, square of RMSF, and B-factor. The published models of (a) β -galactosidase (PDB 5A1A) and (b) TRPV1 (PDB 3J5P) are colored by the local EM map resolutions, the per-residue mean square fluctuations (RMSF²) during MDFF simulation, and published B-factors. Comparison of these figures shows qualitative agreement between local resolution, RMSF², and B-factor. In fact, the local resolutions and B-factors correlate linearly with RMSF² of a fitted model both in the presence as well as absence of the EM map (more details in [Figure 4](#)).

DOI: [10.7554/eLife.16105.016](https://doi.org/10.7554/eLife.16105.016)

limiting resolution of the resulting EM density map ([Kucukelbir et al., 2014](#)), it is not surprising that overall and local map resolutions correlate with overall and local RMSF values of the best fitted model ([Figure 4](#)).

In broad terms, through the present study we establish that RMSF, together with GCC, LCC, EMRinger ([Barad et al., 2015](#)), and iFSC, provide a comprehensive set of criteria for evaluating model and map quality on both global and local levels. The added value of RMSF is particularly evident on the local level, where the other measures may not perform as consistently. For example, a high LCC may be the result of a highly flexible structure fitting to a low-resolution region of the map, and not necessarily of a good representation of the local structure. As a result, although multiple low-resolution regions of the model in [Figure 4—figure supplement 2a](#) possess similar LCCs, disparate RMSFs of the same regions clearly indicate differences in local quality of the model. Likewise, EMRinger scoring, when applied to small groups of residues, does not correlate with local resolution ([Figure 4—figure supplement 2b](#)), and, therefore, is incapable of distinguishing regions of small number of atoms by local model quality. In contrast, RMSF clearly resolves local resolution, and, thus, resolves the map and model quality even with as few as 100 atoms.

RMSF and B-factor determination

High contrast within an EM map allows clear identification of secondary structural elements. However, experimental imaging discrepancies arising from specimen movement and charging, radiation damage, and partial microscope coherence, or computational discrepancies due to inaccurate

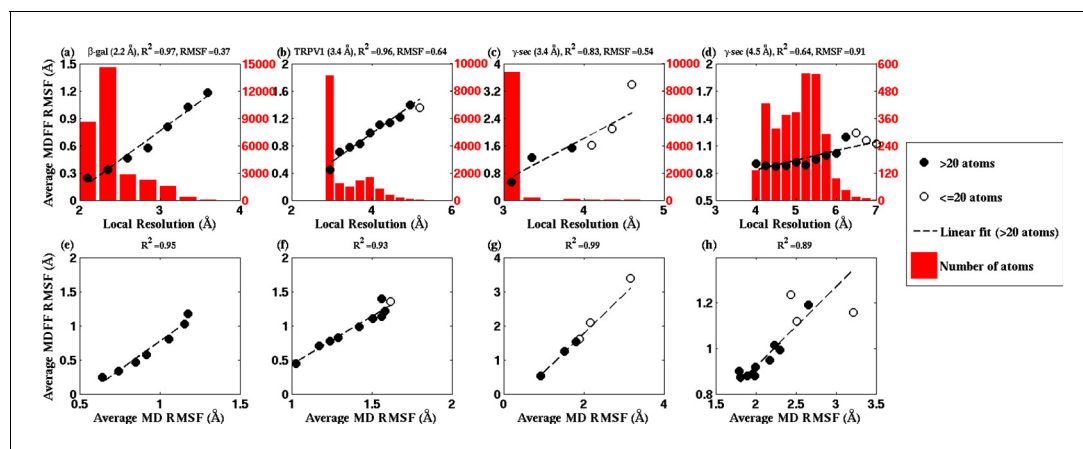


Figure 4. RMSF vs. local resolution plots for various simulations. For each test case shown, atoms in the MDFF-refined structure are classified by local resolution of the map regions they are fitted into. The average RMSF value of atoms (during MDFF simulation) in each resolution bin is calculated and plotted against the local resolution in the cases of (a) β -galactosidase (β -gal) at 2.2 Å, (b) TRPV1 at 3.4 Å, γ -secretase (γ -sec) at (c) 3.4 Å and (d) 4.5 Å resolution, and proteasome (see **Figure 4—figure supplement 3**). The numbers of atoms in the resolution bins are displayed as a histogram (in red) spanning a system-specific range of resolutions. The lowest resolution bins contained low (<20) populations and visual inspection consistently revealed the atoms to be on the edges of the density or were otherwise located inside map noise, and were therefore ignored during further analysis. A clear linear correlation between RMSF and local resolution can be found in each case, such that applying a linear fit produces the high R^2 value shown in each graph heading. Also displayed in each heading is an overall RMSF, averaged over all atoms in the system. The overall RMSF reflects the conformational variety of structures that fit within the map, and is found to correspond to the map resolution such that higher resolutions produce lower RMSFs. The second row of plots show that the RMSF during MDFF simulation also linearly correlates with RMSF during unbiased MD simulations of (e) β -gal, (f) TRPV1 and (g,h) γ -sec, establishing that fluctuations during MDFF reflect the inherent flexibility of a system.

DOI: [10.7554/eLife.16105.017](https://doi.org/10.7554/eLife.16105.017)

The following figure supplements are available for figure 4:

Figure supplement 1. Per-residue RMSFs over β -galactosidase cMDFF fitting.

DOI: [10.7554/eLife.16105.018](https://doi.org/10.7554/eLife.16105.018)

Figure supplement 2. EMRinger score and LCC do not predict local resolution in TRPV1.

DOI: [10.7554/eLife.16105.019](https://doi.org/10.7554/eLife.16105.019)

Figure supplement 3. Average RMSF vs. local resolution during MDFF simulation of proteasome.

DOI: [10.7554/eLife.16105.020](https://doi.org/10.7554/eLife.16105.020)

Figure supplement 4. RMSF values of individual residues during direct MDFF of published β -galactosidase models.

DOI: [10.7554/eLife.16105.021](https://doi.org/10.7554/eLife.16105.021)

determination of the single particle order parameters (Wade, 1992; Fernández et al., 2008), introduce fuzziness to the EM map, thus hindering secondary structure identification. B-factor sharpening restores lost contrast by resolving the fuzzy features and, therefore, is a crucial step of map generation that affects map interpretation. Here, we describe the use of RMSF as a physical basis for the determination of optimal B-factors that preserve contrast during map sharpening.

Figure 5, Figure 5—figure supplement 3, Figure 4—figure supplement 4 demonstrate the relationship between RMSF and B-factor for the 2.2-Å map of β -galactosidase, 3.4-Å map of TRPV1 and 4.5-Å map of γ -secretase; these three systems were chosen due to their reasonable size and availability of unsharpened data. An initial decrease in the RMSF values is observed with increase in B-factor sharpening of the map. However, the RMSF eventually reaches a minimum before increasing as the B-factor is further increased. Surprisingly, for all the three structures the B-factor corresponding to the minimum RMSF coincides with the one determined by Guinier analysis (Fernández et al., 2008; Rosenthal and Henderson, 2003) to provide maximum contrast. In fact, for TRPV1, B-factors determined from the RMSF minima are found to be higher for the soluble regions than for the transmembrane helices (Figure 5—figure supplement 1, b vs. c), again, in agreement with those B-factors derived from the Guinier analysis. Therefore, the RMSF analysis of maps with varying B-factor sharpening provides an alternate procedure for the determination of the optimal B-factor, which has been traditionally determined by Guinier analysis.

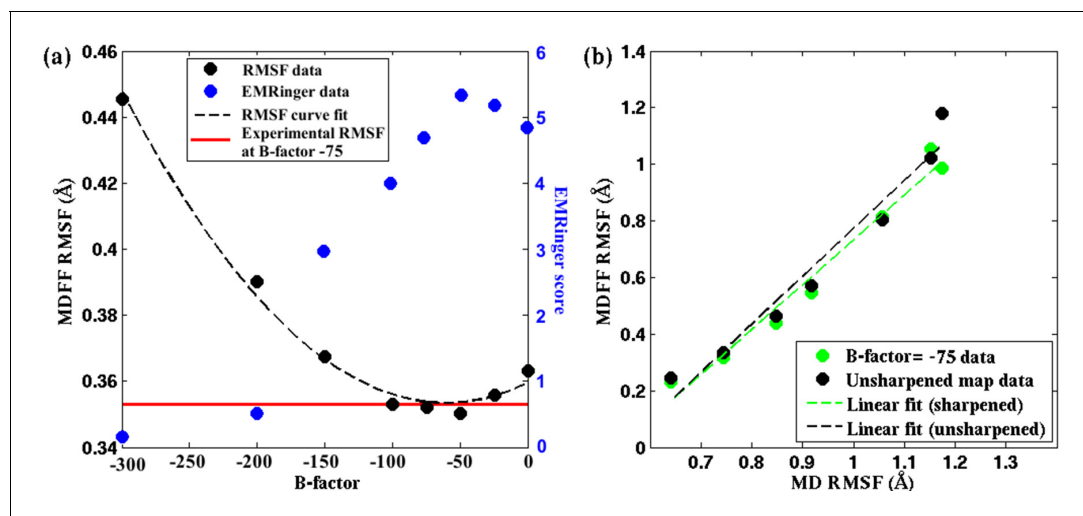


Figure 5. Effect of map sharpening on residue flexibility of β -galactosidase. (a) Overall RMSF of a fitted 2.2Å β -galactosidase structure (PDB 5A1A) during direct MDFF fitting as a function of the B-factor of the fitting map exhibits a parabolic trend. Guinier analysis identifies a B-factor of -75 as optimal, for which the corresponding RMSF (shown in red) coincides with the minimum of the trend line. EMRinger scores (shown in blue) of the same structures show a negative parabolic trend, with the peak coinciding with the minimum of the RMSF plot. (b) The linear relationships between local RMSF during MDFF and during unbiased MD for the unsharpened map and optimally sharpened map are compared. While the linear relationship is preserved in both cases, RMSFs in the sharpened case are slightly lower than in the unsharpened case.

DOI: [10.7554/eLife.16105.022](https://doi.org/10.7554/eLife.16105.022)

The following figure supplements are available for figure 5:

Figure supplement 1. Effect of map sharpening on residue flexibility in TRPV1.

DOI: [10.7554/eLife.16105.023](https://doi.org/10.7554/eLife.16105.023)

Figure supplement 2. Effect of map sharpening on residue flexibility in γ -secretase.

DOI: [10.7554/eLife.16105.024](https://doi.org/10.7554/eLife.16105.024)

Figure supplement 3. EMRinger scores as a function of B-factor.

DOI: [10.7554/eLife.16105.025](https://doi.org/10.7554/eLife.16105.025)

Figure supplement 4. Atom-by-atom B-factor for a β -galactosidase monomer.

DOI: [10.7554/eLife.16105.026](https://doi.org/10.7554/eLife.16105.026)

The rationale for optimal B-factor selection based on an analysis of RMSF values is the following. As pointed out above and in **Figure 4**, RMSF of a fitted model is a function of the quality of the corresponding map. Following this argument, the maps in **Figure 5**, **Figure 5—figure supplement 1** and **Figure 5—figure supplement 2** that produce the model with the lowest RMSF during fitting represent the density envelopes where atoms can be positioned with the least uncertainty at the given experimental resolutions. The corresponding sharpening B-factor thus provides the optimal contrast for atom placement and secondary structure determination.

Our argument for the RMSF-based selection of B-factors is supported by the well-established quadratic relationship of B-factors with the RMSF of atoms of a structure in an experimental setting, i.e. $8\pi^2/3(\text{RMSF})^2$ (*Rosenthal and Henderson, 2003*; *Liu and Xiong, 2014*). This relationship indeed implies that the smaller the average atomic fluctuations, the lower are the B-factors, and by definition (*Liu and Xiong, 2014*) the higher are the measured structure factor amplitudes, and hence the contrast, of the resulting map in both crystallographic and EM experiments (*Liu and Xiong, 2014*; *Fernández et al., 2008*). By analogy, RMSF within MDFF is a measure of these average atomic fluctuations under the experimental setting (**Figure 4**, bottom row) and, therefore, within a set of maps of varying B-factors, the highest contrast is indicated by the lowest RMSF obtained during fitting of a model. Since Guinier analysis of EM data selects B-factors based on the same atomic displacement-structure factor relationship that justifies the RMSF-based selection of B-factors (*Rosenthal and Henderson, 2003*), the B-factor selected from our analysis of the lowest RMSF matches excellently with those from the Guinier analysis of EM maps.

The higher contrast of the B-factor sharpened map, over the unsharpened one, is evident from the higher quality of the models derived from the sharpened maps. Presented in **Figure 5a**,

Figure 5—figure supplement 3b, and **Figure 5—figure supplement 1a**, for the cases of β -galactosidase, γ -secretase, and TRPV1, respectively, are EMRinger scores of models fitted to unsharpened and sharpened maps. These scores indicate clearly that the models derived from maps at B-factor sharpening of -50 (**Figure 5a**), -100 (**Figure 5—figure supplement 3b**), and -100 (**Figure 5—figure supplement 1a**) are more accurate than those derived from the unsharpened ones. Indeed, as further illustrated in **Figure 5a**, **Figure 5—figure supplement 3b**, and **Figure 5—figure supplement 1**, the maximum EMRinger score is attained for the same B-factor that produces the minimum RMSF. Sharpening by B-factors of magnitudes any higher than 50 for β -galactosidase, and 100 for γ -secretase or TRPV1 damages key density features thus reducing the quality of the associated fitted models. In addition, correlation of the RMSF of the fitted structure with that from an unbiased MD of the system is still preserved (**Figure 5b** and **Figure 5—figure supplement 2b**) for the B-factor sharpened map, confirming that the dynamical fluctuations of the structure within the sharpened map reflects the inherent dynamics of the system. Altogether, RMSF establishes a unique map-model validation criterion that represents an ensemble view of the fitted structures, while also preventing over-sharpening of the EM maps.

RMSF and per-residue B-factors

The quadratic relationship between RMSF and B-factors can be further employed to determine per-residue B-factors. Presented in **Figure 3**, the per-residue RMSF^2 of β -galactosidase and TRPV1, derived from MDFF, show excellent agreement with the distribution of local resolution and experimentally reported B-factors: regions with higher RMSF^2 correspond to lower local resolution and higher per-residue B-factors, and vice versa. In fact, the quantitative agreement between the B-factors derived from MDFF through the computation of $8\pi^2/3(\text{RMSF})^2$ ($0.02 < (\text{RMSF})^2 < 4.43$ [**Figure 3**]) and those reported in the experiments is remarkable (**Figure 5—figure supplement 4**); a cross-correlation of 55% is found between the data sets, which improves to 60% in the structured regions. Larger discrepancies are observed between the RMSF-based and reported B-factors of TRPV1, particularly in the soluble region (**Figure 3b**). This discrepancy is expected, as for poorly resolved regions MD provides up to six-fold higher B-factors (**Kuzmanic et al., 2014**); the higher B-factors indeed have been demonstrated to be a more accurate representation of conformational diversity (**Kuzmanic et al., 2014**). For the TRPV1 example, the highest computed B-factor is $8\pi^2/3 \times 9.25 = 240.5$ (**Figure 3b**) compared to the reported value 53.17. However, the majority of the computed B-factors fluctuate about the value of $8\pi^2/3 \times 1.73 = 44.98$, which is in fair agreement with the reported values of 20–30.

Overall, B-factors of cryo-EM maps are typically calculated by Guinier analysis for maps with resolutions better than ~ 10 Å (**Fernández et al., 2008**; **Rosenthal and Henderson, 2003**). Using a mask around specific regions of interest, estimation of local B-factors for different parts of a map is also possible. However, this method is limited to large domains of macromolecular complexes due to problems associated with tight masking of cryo-EM maps. For maps with resolutions better than ~ 3 Å, local B-factors could be estimated and refined in X-ray crystallography programs. However, most high-resolution cryo-EM maps have resolutions between 3 to 5 Å (**Liao et al., 2013**; **Bartesaghi et al., 2014**; **Zhao et al., 2015**; **Fischer et al., 2015**). In this resolution range, B-factors could be estimated from MDFF and used as prior information to improve model building and refinement. For resolutions better than 3 Å, MDFF-derived values may serve as initial estimates of B-factors. Furthermore, many cryo-EM maps show local variations in resolution, complicating the model building process. A possible solution may be to combine local B-factor refinement for high-resolution regions of a map and B-factors derived from MDFF for lower resolution regions. A resolution-dependent weighting scheme could be incorporated to combine the different values for optimal performance, which may help improve the accuracy of atomic models derived from high-resolution cryo-EM maps.

Accessibility through cloud computing

ReMDFF involves many independent, though sporadically communicating, MDFF simulations that can be run well on a parallel computer as exploited by the NAMD software for the case of replica exchange simulations (**Jiang et al., 2014**). As a result, ReMDFF provides an efficient and automated method which can converge on a final fitted structure more quickly than direct or cMDFF (**Figure 2—**

Table 3. Performance and cost results for ReMDFF of carbon monoxide dehydrogenase on Amazon Web Services (AWS) Elastic Compute Cloud (EC2) platform. Costs are incurred on a per-hour basis, with a 1 hr minimum.

Instance type	CPU	Performance (ns/day)	Time (hours)	Simulation cost (\$)
c3.8xlarge	30	5.88	0.41	1.68
c3.4xlarge	12	3.33	0.72	0.84
c3.2xlarge	6	1.35	1.78	0.84

DOI: [10.7554/eLife.16105.027](https://doi.org/10.7554/eLife.16105.027)

figure supplement 3). However, a potential bottleneck may exist with respect to the computing hardware that a researcher has access to. Cloud computing offers a potential solution, allowing a researcher to focus on the scientific challenges of their project without having to worry about local availability and administration of suitable computer hardware.

To prove the feasibility of performing ReMDFF simulations on a cloud platform, we performed ReMDFF for a test system, carbon monoxide dehydrogenase (PDB 1OAO:chain C), on the Amazon Web Services (AWS) Elastic Compute Cloud (EC2) platform. The test system converges to the known target structure in approximately 0.1 ns simulation time. The time to convergence requires very little wall clock time, and, therefore, incurs a small monetary cost to a user (**Table 3**). However, it should be noted that human error and varying experience level can easily add to the incurred cost of cloud usage. Some systems may require multiple simulations to achieve a high quality structure and, therefore, additional time beyond the example discussed here. Furthermore, preparing a structure for simulation may require additional time and resources over the purely simulation-oriented results presented here. The files and information necessary to run ReMDFF on the test system using EC2 cloud computing resources are available at (<http://www.ks.uiuc.edu/Research/cloud/>). The Implementation section of the Methods contains further details for setting up and running ReMDFF simulations.

Discussion

Flexible fitting methods have facilitated structure determination from low-resolution EM maps for more than a decade (*Tama et al., 2004; Suhre et al., 2006; Velazquez-Muriel et al., 2006; Orzechowski and Tama, 2008; Topf et al., 2008; Kovacs et al., 2008; Lopéz-Blanco and Chacón, 2013; Wu et al., 2013*) and continue to be the methods of choice for resolving molecular systems with atomic resolution. MDFF, in particular, has been a front-runner among methods that have facilitated the discovery of some of the most complicated structures in modern day structural biology (*Hsin et al., 2009; Sener et al., 2009; Gumbart et al., 2011; Frauenfeld et al., 2011; Zhao et al., 2013; Wickles et al., 2014*).

cMDFF and ReMDFF, the new variants of MDFF introduced above, offer now accurate fitting of atomic-level structures within sub-5 Å EM maps, a feat thus far inaccessible to direct MDFF. These new methods extend the radius of convergence of MDFF to at least 25 Å, fitting models to maps of resolutions as high as 3.2 Å. This radius of convergence is at least twice that reported for Rosetta refinements of the 20S proteasome (*DiMaio et al., 2015*). Such a broad radius of convergence will allow the refinement of extremely poorly guessed initial models with MDFF, as demonstrated in the cases of β -galactosidase and TRPV1 reported here.

ReMDFF simulations involving the so-called replica-exchange molecular dynamics converge quickly using a small number of replicas and are thus amenable to cloud computing applications. Running ReMDFF on the cloud lowers greatly the barrier to usage of the method, providing a cost-effective and practical solution to fitting structures to high-resolution cryo-EM densities for researchers who neither own nor can administer their own advanced computer hardware.

The accuracy of structures refined by cMDFF and ReMDFF has been confirmed by standard error analysis protocols, both in terms of quality of fit as well as in terms of the quality of the model. The results clearly show that sidechain refinements through MDFF produce accurate placement of C_{α} and C_{β} atoms and modeling of the associated dihedrals.

Beyond the standard error analysis protocols, which apply to single static structures, the quality of both the fit and the model can be evaluated by ensemble-based measures. An example

demonstrated in this study pertains to the use of RMSF during MDFF refinement for simultaneously evaluating the quality of model, map, and fit. Furthermore, utilizing the fact that the inherent flexibility of a macromolecule is a key determinant of the achievable resolution of the corresponding map, RMSF values have been employed to identify the optimal amount of sharpening for a given map offering the highest contrast, as established typically through Guinier analysis. The RMSF computations in MDFF provide a viable means of determining per-residue B-factors.

Altogether, interpretation of a map as being representative of an ensemble rather than a single model brings to light new ways of model validation. EM maps, including those at high-resolution, typically do not have a uniform local resolution (*Kucukelbir et al., 2014*) and contain low-resolution regions, such as flexible exterior or transmembrane segments (*Leschziner and Nogales, 2007*). The sub-5 Å EM maps have also been able to resolve proteins in multiple conformations (*Matthies et al., 2016*). Both these type of maps will continue to benefit from accurate ensemble-based flexible fitting techniques for the foreseeable future. MDFF provides a natural method to model flexible regions; de novo models constructed for one conformational state can be flexibly fitted into the density of the other state(s), thus avoiding the arduous task of model construction for all the conformational states for capturing a conformational transition process with cryo-EM.

Materials and methods

In the following section, we first outline the methodology underlying direct MDFF, cMDFF, and ReMDFF, along with specific protocols applied for the refinement of β -galactosidase and TRPV1 models. Second, we describe the computations of local and global root mean square fluctuation (RMSF) values, that are utilized for the evaluation of map-model quality. Finally, details on the computational implementation of all the three MDFF protocols are provided.

Direct MDFF

MDFF requires, as input data, an initial structure and a cryo-EM density map. A potential map is generated from the density and subsequently used to bias a MD simulation of the initial structure. The structure is subject to the EM-derived potential while simultaneously undergoing structural dynamics as described by the MD force field.

Let the Coulomb potential associated with the EM map be $\Phi(\mathbf{r})$. Then the MDFF potential map is given by

$$V_{EM}(\mathbf{r}) = \begin{cases} \zeta \left(\frac{\Phi(\mathbf{r}) - \Phi_{thr}}{\Phi_{max} - \Phi_{thr}} \right) & \text{if } \Phi(\mathbf{r}) \geq \Phi_{thr}, \\ \zeta & \text{if } \Phi(\mathbf{r}) < \Phi_{thr}. \end{cases} \quad (1)$$

where ζ is a scaling factor that controls the strength of the coupling of atoms to the MDFF potential, Φ_{thr} is a threshold for disregarding noise, and $\Phi_{max} = \max(\Phi(\mathbf{r}))$. The potential energy contribution from the MDFF forces is then

$$U_{EM} = \sum_i w_i V_{EM}(\mathbf{r}_i), \quad (2)$$

where i labels the atoms in the structure and w_i is an atom-dependent weight, usually the atomic mass.

During the simulation, the total potential acting on the system is given by

$$U_{total} = U_{MD} + U_{EM} + U_{SS} \quad (3)$$

where U_{MD} is the MD potential energy as provided by MD force fields (e.g. CHARMM) and U_{SS} is a secondary structure restraint potential that prevents warping of the secondary structure by the potentially strong forces due to U_{EM} . A detailed description of the potentials arising in **Equation 3** is given in Trabuco et al (*Trabuco et al., 2008, 2009*).

After the MDFF and restraint potentials are created through the MDFF plugin of VMD (*Humphrey et al., 1996*), the initial structure is rigid-body docked (e.g. with Situs [*Wriggers, 2010*]) into the density map. Prior to simulation, MDFF-specific parameters can be modified and include ζ and the subset of atoms to be coupled to the MDFF potential. The latter typically consists of all non-

hydrogen atoms or backbone atoms and ζ is usually set to 0.3. MDFF can be performed in various simulated conditions, including different temperatures and vacuum, membrane, and explicit or implicit (Tanner *et al.*, 2011) solvent environments. The choice of parameters and conditions depends on the requirements of each specific case. For example, a highly polar molecule would be more accurately simulated in explicit solvent rather than in vacuum, but the computational cost would be much higher in this case. The MDFF simulation is run until the system has reached stationarity, as determined by RMSD; typical run times are nanoseconds.

Cascade MDFF

In cascade MDFF (cMDFF), the initial structure is sequentially fitted to a series of potential maps of successively higher resolution, with the final potential map being the original one derived from the EM map. Starting with $i = 1$, the i th map in the series is obtained by applying a Gaussian blur of width $\sigma_i \geq 0 \sim \text{\AA}$ to the original potential map, such that σ_i decreases as the structure is fitted in the sequence $i = 1, 2, \dots, L$, where L is the total number of maps in the series, so that $\sigma_L \geq 0 \sim \text{\AA}$. One can intuitively understand cMDFF as fitting the simulated structure to an initially large and ergodic conformational space that is shrinking over the course of the simulation towards the highly corrugated space described by the original MDFF potential map.

Figure 1 provides a visual representation of cMDFF. For a mathematical illustration, suppose that the original potential map can be written as a sum of Gaussians

$$V_{EM}(\mathbf{r}) = \sum_n c_n G(\mathbf{r}; \mathbf{r}'_n, s_n), \quad (4)$$

where c_n is a weight, $G(\mathbf{r}; \mathbf{r}'_n, s_n)$ is a Gaussian function centered at \mathbf{r}'_n with half-width s_n and evaluated at \mathbf{r} . The result of a Gaussian blur of half-width σ_i on V_{EM} is (see Appendix 1 - Section 1 for details)

$$V_{\sigma_i}(\mathbf{r}) = \sum_n c_n G\left(\mathbf{r}; \mathbf{r}'_n, \sqrt{s_n^2 + \sigma_i^2}\right). \quad (5)$$

Hence, the half-width σ_i allows one to tune the characteristic width of the potential map through the half-widths of its component Gaussians $\sqrt{s_n^2 + \sigma_i^2}$. The initial fitting starts with a large σ_1 , corresponding to a diffuse potential which allows much structural mobility, and proceeds along decreasing values of σ_i , corresponding to narrower potentials with steeper gradients, so that the structure is gradually settled into the original potential map, characterized by $\sigma_L \geq 0 \sim \text{\AA}$.

In practice, the series of cMDFF maps is generated from the original potential map using VMD's `volutil` Gaussian blur tool. Optimal values for the first half-width σ_1 and the change in σ_i from one map to the next are case-dependent. Values used in the present study were obtained through trial and error. In general, structures far from the ideal conformation benefit from a large σ_1 ($>5 \text{\AA}$) so as to explore a large conformation space. On the other hand, if the original map has a high resolution, small changes in σ_i ($<1 \text{\AA}$) would allow a gradual convergence required to avoid being trapped in local potential minima. In our simulations, the change in σ_i is initially 1\AA . A concrete example is $\sigma_1 = 5 \text{\AA}$, $\sigma_2 = 4 \text{\AA}$, $\sigma_3 = 3 \text{\AA}$, $\sigma_4 = 2 \text{\AA}$, $\sigma_5 = 1 \text{\AA}$, $\sigma_6 = 0 \text{\AA}$. A second trial using changes of 0.5\AA was performed ($\sigma_1 = 5 \text{\AA}$, $\sigma_2 = 4.5 \text{\AA}$, $\sigma_3 = 4 \text{\AA}$, ..., $\sigma_{10} = 0.5 \text{\AA}$, $\sigma_{11} = 0 \text{\AA}$), and if the resulting structure of the second trial presented a better fit, then the first trial was discarded.

Resolution exchange MDFF

Replica Exchange MD (ReMD) is an advanced sampling method that explores conformational phase space in search of conformational intermediates, which are separated by energy barriers too high to be overcome readily by fixed temperature simulations. Instead of working with a single, fixed MD simulation, ReMD carries out many simulations in parallel, but at different temperatures $T_1 < T_2 < T_3 < \dots$ where the lowest temperature T_1 is the temperature of actual interest, typically, room temperature. The simulations of several copies of the system, the so-called replicas, run mainly independently, such that ReMD can be easily parallelized on a computer, but at regular time points the instantaneous conformations of replicas of neighboring temperatures are compared in terms of energy and transitions between replicas are permitted according to the so-called Metropolis

criterion (*Sugita and Okamoto, 1999*). This way the highest temperature replicas overcome the energy barriers between conformational intermediates and through the Metropolis criterion moves the T_1 replica benefits from it such that transitions between intermediates occur frequently. The application of the Metropolis criterion in the protocol guarantees that the conformations of the T_1 replica are Boltzmann-distributed.

ReMDFF extends the concept of ReMD to MDFF by simply differentiating replicas not by temperatures $T_1 < T_2 < T_3 < \dots$, but by the half-width parameters $\sigma_1 > \sigma_2 > \sigma_3 > \dots$. Numerical experiments showed that ReMDFF works extremely well as documented in the present study. As NAMD can parallelize ReMD well (*Jiang et al., 2014*), it can do the same for ReMDFF, such that the enhanced sampling achieved translates into extremely fast MDFF convergence. At certain time instances replicas i and j , of coordinates \mathbf{x}_i and \mathbf{x}_j and fitting maps of blur widths σ_i and σ_j , are compared energetically and exchanged with Metropolis acceptance probability

$$p(\mathbf{x}_i, \sigma_i, \mathbf{x}_j, \sigma_j) = \min\left(1, \exp\left(\frac{-E(\mathbf{x}_i, \sigma_j) - E(\mathbf{x}_j, \sigma_i) + E(\mathbf{x}_i, \sigma_i) + E(\mathbf{x}_j, \sigma_j)}{k_B T}\right)\right), \quad (6)$$

where k_B is the Boltzmann constant, $E(\mathbf{x}, \sigma)$ is the instantaneous total energy of the configuration \mathbf{x} within a fitting potential map of blur width σ .

MDFF protocols for β -galactosidase and TRPV1

Computational protocols for performing direct MDFF, cMDFF, and ReMDFF refinements of the two test systems, β -galactosidase and TRPV1, are now outlined.

Direct MDFF

In order to provide a basis for comparison with cMDFF and ReMDFF, direct MDFF simulations were performed for both β -galactosidase and TRPV1. Original published maps of resolutions 3.2 Å and 3.4 Å for β -galactosidase (EMD-5995 [*Bartesaghi et al., 2014*]) and TRPV1 (EMD-5778 [*Liao et al., 2013*]), respectively, were fitted with search models characterized by RMSD values of 7 Å and 25 Å, respectively, relative to the known de novo target models (PDB entries 3J7H and 3J5P, respectively); these search models were prepared as described in Appendix 1 - Section 3. Scale factors ζ (*Equation 1*) of values 1.0 and 0.3 were employed for β -galactosidase and TRPV1, respectively, to couple all the heavy atoms of the models to the respective maps. All other simulation parameters are noted in Appendix 1 - Section 2. The resulting structure from each MDFF simulation was then subjected to a final re-refinement applying a scaling factor of 1.0. Furthermore, the temperature was ramped down from 300 K to 0 K over 30 ps and held at 0 K for an additional 1 ns. This re-refinement step additionally improved the fitting of sidechains (*McGreevy et al., 2014*). Thus, all the MDFF results reported in the present study pertain to the structures that are obtained from direct MDFF, cMDFF, or ReMDFF simulations followed by the final re-refinement step.

cMDFF

The general simulation protocol consists of a series of consecutive MDFF simulations. The search models for β -galactosidase and TRPV1 were first minimized over between 500 to 1000 time steps. Next, MDFF simulation runs were sequentially performed, starting with the map of the lowest resolution, progressing through maps of successively higher resolution, and ending with the original map. Each run was chosen to be long enough for the structure to equilibrate within the MDFF potential. Taking advantage of the stochastic nature of MDFF simulations, multiple independent cMDFF simulations (10 in the present study) were performed for each system to be fitted, generating an ensemble of fitted structures. From the ensemble, the best structure was determined by the various quality indicators described in Results. This structure was then subjected to the final re-refinement step to allow for accurate resolution of sidechains.

For β -galactosidase, cMDFF was initiated with a map blurred with half-width $\sigma_1 = 5$ Å. Use of 0.5 Å resolution steps produced better-fitted structures than 1 Å steps. Hence, the final simulation utilized $L = 11$ maps in total, including the original. A search model was obtained by subjecting the published structure to a 5-ns equilibration MD run with temperature 1000 K, yielding a structure with backbone RMSD of 7.6 Å relative to the original structure (see Appendix 1 - Section 3). 70-ps MDFF simulations were performed at each of the 11 resolutions to achieve convergence during the

cMDFF protocol. Simulation parameters for all these MDFF runs are identical to those used for the direct MDFF simulations above.

For TRPV1, Gaussian blurred maps were generated starting with a half-width of $\sigma_1 = 5 \text{ \AA}$, and decreasing by 1 \AA for each subsequent map, thus yielding a series of $L = 6$ maps, including the original. The initial structure for this case was obtained using an interactive MD protocol, described in Appendix 1 - Section 2, to displace the ankyrin repeat region of one subunit. The resulting model deviated from the de novo model by an RMSD of about 10 \AA in the displaced region and about 25 \AA in the overall structure. Unlike β -galactosidase, now, 100-ps MDFF simulations were performed at each of the 6 resolutions to achieve convergence.

ReMDFF

ReMDFF was performed on both β -galactosidase and TRPV1 using the same search models, simulation parameters, and high-resolution maps as in the cMDFF simulations. 11 and 6 replicas were employed for β -galactosidase and TRPV1 respectively with an exchange trial interval of 1 ps. In each case, the total energy of each replica was monitored and the simulation was run until the energies reached a stationary level. The ReMDFF simulation was found to converge in 0.1 ns for the β -galactosidase refinement, and in 0.02 ns for that of TRPV1. Finally, similar to direct MDFF and cMDFF, the re-refinement step was performed to improve sidechain geometry.

Fluctuation analysis

The local resolutions of a density map can be computed with ResMap (*Kucukelbir et al., 2014*) and used within VMD to select the atoms of a structure that are contained in a range of resolutions found by the ResMap analysis. First, the local resolution map output by ResMap is loaded into VMD and then the *interpvol* keyword can be used to automatically select the atoms found inside the volume values specified, using interpolation. The average RMSF of each selection can then be calculated for a structure during the steps of a MDFF simulation after the initial fitting has occurred and the structure has stabilized. In principle any criteria for atom selection can be used for RMSF analysis, though we use local resolution of the EM density here to illustrate the correlation between the two measurements. Additionally, we compute a global average RMSF of the entire structure.

The ensemble-based nature of the RMSF analysis means that the quality metric is not dependent on a single structure, but instead a large family of structures can be employed as a better representative of the data. Ensemble-based analyses are a natural and powerful benefit of the MD-based nature of MDFF. RMSF analysis does not, however, require MDFF to be used as the method of refinement. In principle, any refinement method can be used to obtain the fitted model. A subsequent short MDFF simulation of the fitted model can then be performed to obtain the data necessary for the RMSF analysis.

Implementation

Incorporating advanced simulation techniques, such as multi-copy algorithms (*Jiang et al., 2014*), into the MDFF protocol creates a more efficient and accurate computational strategy in cMDFF and ReMDFF. However, these advanced simulation techniques come with an added complexity in the setup and execution of the methods. The current implementation of these methods in NAMD (*Phillips et al., 2005*) and VMD (*Humphrey et al., 1996*) aim to automate the steps previously discussed. The MDFF Graphical User Interface (GUI) (*McGreevy et al., 2016*) can be used to set up cMDFF and ReMDFF simulations and provides default parameters, including the number and extent of smoothed maps used for the fitting, with which to run. The parameters for the smoothed maps and number of steps used per map are set heuristically based on previous experience and represent an adequate initial starting point. The GUI automatically generates each of the smoothed maps and converts them to potentials for use in the ReMDFF simulation. All parameters can be tuned by a user to adapt the protocols to their specific system and preference. The GUI produces a series of NAMD configuration files and scripts used for running the simulation, as well sorting and visualizing the results in VMD. Instruction on the use of MDFF, including the GUI, is given in the tutorial found at http://www.ks.uiuc.edu/Training/Tutorials/science/mdff/tutorial_mdff-html/.

Future development will allow for the automatic generation of the smoothed maps in NAMD at runtime. NAMD will also analyse the dynamics of the system to determine when the simulation has

converged and move on to fitting to the next density map in the sequence in case of cMDFF calculations. Furthermore, advanced visualization and analysis techniques in VMD (e.g. new graphical representations) will be critical for properly understanding the RMSF analyses and to provide greater insight when examining the quality of a model.

The use of advanced simulation techniques also comes with an added cost of computational requirements. Adapting the algorithms to best utilize available computational hardware is key when developing efficient methods. Fortunately, the cMDFF and ReMDFF methods and associated analysis algorithms are well suited to highly efficient software implementations on contemporary multi-core CPUs and graphics processing unit (GPU) accelerators. We observe that by storing the complete cascade resolution series in efficient multi-resolution data structures such as mip maps (Williams, 1983), the MDFF cascade algorithm can access a continuously variable resolution representation of the original cryo-EM density map, while making efficient use of CPU and GPU processors and memory systems (Stone et al., 2007).

The parallel nature of ReMDFF presents an opportunity for efficient, automated sampling of maps of varying resolution. However, to achieve the best efficiency, the simulations should be performed on multi-core CPUs with relatively high core counts (i.e. at least 1 core per replica). Access to such multi-core computers could prohibit use of ReMDFF, however access to machines with the necessary hardware is easily achieved through cloud computing. The cloud computing model provides researchers with access to powerful computational equipment that would otherwise be too costly to procure, maintain, and administer on their own. A particular obstacle is that structural modeling often involves the use of different software suites, such as VMD (Humphrey et al., 1996) for simulation preparation and Situs (Wriggers, 2010) for initial rigid-body docking or VMD, NAMD (Phillips et al., 2005), and Rosetta (Leaver-Fay et al., 2011) for iterative refinement of models (Lindert and McCammon, 2015). Cloud platforms can easily bundle different software packages used in a modeling workflow to guarantee their availability and interoperability on a standardized system. Through the cloud version of our MDFF program suite a user does not need to be aware of any of the above mentioned technical issues.

To prove the feasibility of performing ReMDFF simulations on a cloud platform, we ran ReMDFF for the test system, carbon monoxide dehydrogenase (PDB 1OAO:chain C), on the Amazon Web Services (AWS) Elastic Compute Cloud (EC2) platform. For the purposes of testing ReMDFF on EC2, we ran benchmarks on a variety of compute-optimized EC2 instance types. The details of the instance types can be found in Table 3. We used the same 6 smoothed density maps as the previously discussed cMDFF and ReMDFF simulations in the Proof of Principle and, therefore, also 6 replicas. The files and information necessary to run ReMDFF on the test system using EC2 cloud computing resources are available at (<http://www.ks.uiuc.edu/Research/cloud/>).

Acknowledgements

This work is supported by grants NIH 9P41GM104601, NIH 5R01GM098243-02, and NIH U54GM087519 from the National Institutes of Health. The authors gratefully acknowledge computer time provided by the NSF-funded Extreme Science and Engineering Discovery Environment (XSEDE) MCA93S028, Sriram Subramanian for extremely insightful discussions on the high-resolution cryo-EM methods in general, and the β -galactosidase study in particular, and Benjamin Barad from the Fraser group at UCSF for guidance on the use of the EMRinger program. The authors also acknowledge the Beckman Postdoctoral Fellowship program for supporting A Singharoy.

Additional information

Funding

Funder	Grant reference number	Author
National Institutes of Health	9P41GM104601	Abhishek Singharoy Ivan Teo Ryan McGreevy John E Stone Klaus Schulten

National Science Foundation	MCA93S028	Abhishek Singharoy Ivan Teo Ryan McGreevy John E Stone Klaus Schulten
Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign		Abhishek Singharoy
National Institutes of Health	5R01GM098243-02	Ryan McGreevy John E Stone Klaus Schulten
National Institutes of Health	U54GM087519	Klaus Schulten

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

AS, IT, RM, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; JES, Conception and design, Drafting or revising the article; JZ, Analysis and interpretation of data, Drafting or revising the article; KS, Computational simulation technologies, Conception and design, Analysis and interpretation of data, Drafting or revising the article

Author ORCIDs

Klaus Schulten,  <http://orcid.org/0000-0001-7192-9632>

Additional files

Supplementary files

- Supplementary file 1. Supplementary tables. (A) TRPV1 MDFF Results. (B) Structure quality indicators for TRPV1 structures. (C) MDFF for the TRPV1 TM region. (D) Structure quality indicators for γ -secretase. (E) Measures of fit for MDFF refinements of β -galactosidase prepared initially at 1000 K. (F) Structural quality indicators for MDFF-refined β -galactosidase prepared initially at 1000 K.

DOI: [10.7554/eLife.16105.028](https://doi.org/10.7554/eLife.16105.028)

- Supplementary file 2. Initial and refined structures.

DOI: [10.7554/eLife.16105.029](https://doi.org/10.7554/eLife.16105.029)

Major datasets

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database, license, and accessibility information
Darnault C, Volbeda A, Kim EJ, Legrand P, Vernede X, Lindahl PA, Fontecilla-Camps JC	2003	Ni-Zn-[Fe4-S4] and Ni-Ni-[Fe4-S4] Clusters in Closed and Open Alpha Subunits of Acetyl-Coa Synthase/Carbon Monoxide Dehydrogenase	http://www.rcsb.org/pdb/explore/explore.do?structureId=1OAO	Publicly available at the Protein Data Bank (accession no: 1OAO)
Liao M, Cao E, Julius D, Cheng Y	2013	Structure of TRPV1 ion channel determined by single particle electron cryo-microscopy	http://www.rcsb.org/pdb/explore/explore.do?structureId=3J5P	Publicly available at the Protein Data Bank (accession no: 3J5P)
Liao M, Cao E, Julius D, Cheng Y	2015	Structure of the capsaicin receptor, TRPV1, determined by single particle electron cryo-microscopy	http://emsearch.rutgers.edu/atlas/5778_summary.html	Publicly available at the EMDataBank (accession no: EMD-5778)
Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne J, Subramaniam S	2015	2.2 Å resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor	http://www.rcsb.org/pdb/explore/explore.do?structureId=5A1A	Publicly available at the Protein Data Bank (accession no: 5A1A)

Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne JL, Subramaniam S	2015	2.2 Å resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor	http://emsearch.rutgers.edu/atlas/2984_summary.html	Publicly available at the EMDDataBank (accession no: EMD-2984)
Bai X, Yan C, Yang G, Lu P, Ma D, Sun L, Zhou R, Scheres SHW, Shi Y	2015	Cryo-EM structure of the human gamma-secretase complex at 3.4 angstrom resolution.	http://www.rcsb.org/pdb/explore/explore.do?structureId=5A63	Publicly available at the Protein Data Bank (accession no: 5A63)
Bai XC, Yan CY, Yang GH, Lu PL, Ma D, Sun LF, Zhou R, Scheres SHW, Shi YG	2015	Cryo-EM structure of the human gamma-secretase complex at 3.4 angstrom resolution	http://emsearch.rutgers.edu/atlas/3061_summary.html	Publicly available at the EMDDataBank (accession no: EMD-3061)
Lu P, Bai XC, Ma D, Xie T, Yan C, Sun L, Yang G, Zhao Y, Zhou R, Scheres SHW, Shi Y	2014	Structure of a extracellular domain	http://www.rcsb.org/pdb/explore/explore.do?structureId=4UPC	Publicly available at the Protein Data Bank (accession no: 4UPC)
Lu PL, Bai XC, Ma D, Xie T, Yan CY, Sun LF, Yang GH, Zhao YY, Zhou R, Scheres SHW, Shi YG	2015	Three-dimensional structure of human gamma-secretase at 4.5 angstrom resolution	http://emsearch.rutgers.edu/atlas/2677_summary.html	Publicly available at the EMDDataBank (accession no: EMD-2677)
Li X, Mooney P, Zheng S, Booth C, Braunfeld MB, Gubbens S, Agard DA, Cheng Y	2015	Thermoplasma acidophilum 20S proteasome	http://www.rcsb.org/pdb/explore/explore.do?structureId=3J9I	Publicly available at the Protein Data Bank (accession no: 3J9I)
Li X, Mooney P, Zheng S, Booth C, Braunfeld MB, Gubbens S, Agard DA, Cheng Y	2015	3D reconstruction of archaeal 20S proteasome	http://emsearch.rutgers.edu/atlas/5623_summary.html	Publicly available at the EMDDataBank (accession no: EMD-5623)
Bartesaghi A, Matthies D, Banerjee S, Merk A, Subramaniam S	2014	Structure of beta-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy	http://www.rcsb.org/pdb/explore/explore.do?structureId=3J7H	Publicly available at the Protein Data Bank (accession no: 3J7H)
Bartesaghi A, Matthies D, Banerjee S, Merk A, Subramaniam S	2015	Structure of beta-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy	http://emsearch.rutgers.edu/atlas/5995_summary.html	Publicly available at the EMDDataBank (accession no: EMD-5995)

References

- Adams PD**, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH. 2010. PHENIX : a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D Biological Crystallography* **66**:213–221. doi: [10.1107/S0907444409052925](https://doi.org/10.1107/S0907444409052925)
- Afonine PV**, Headd JJ, Terwilliger TC, Adams PD. 2013. New tool: phenix.real_space_refine. *Computational Crystallography Newsletter* **4**:43–44.
- Bai XC**, Yan C, Yang G, Lu P, Ma D, Sun L, Zhou R, Scheres SH, Shi Y. 2015. An atomic structure of human γ -secretase. *Nature* **525**:212–217. doi: [10.1038/nature14892](https://doi.org/10.1038/nature14892), PMID: [26280335](https://pubmed.ncbi.nlm.nih.gov/26280335/)
- Barad BA**, Echols N, Wang RY, Cheng Y, DiMaio F, Adams PD, Fraser JS. 2015. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nature Methods* **12**:943–946. doi: [10.1038/nmeth.3541](https://doi.org/10.1038/nmeth.3541), PMID: [26280328](https://pubmed.ncbi.nlm.nih.gov/26280328/)
- Bartesaghi A**, Matthies D, Banerjee S, Merk A, Subramaniam S. 2014. Structure of β -galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy. *PNAS* **111**:11709–11714. doi: [10.1073/pnas.1402809111](https://doi.org/10.1073/pnas.1402809111), PMID: [25071206](https://pubmed.ncbi.nlm.nih.gov/25071206/)
- Bartesaghi A**, Merk A, Banerjee S, Matthies D, Wu X, Milne JL, Subramaniam S. 2015. 2.2 Å resolution cryo-EM structure of β -galactosidase in complex with a cell-permeant inhibitor. *Science* **348**:1147–1151. doi: [10.1126/science.aab1576](https://doi.org/10.1126/science.aab1576), PMID: [25953817](https://pubmed.ncbi.nlm.nih.gov/25953817/)
- Best RB**, Zhu X, Shim J, Lopes PE, Mittal J, Feig M, Mackerell AD. 2012. 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. *Journal of Chemical Theory and Computation* **8**:3257–3273. doi: [10.1021/ct300400x](https://doi.org/10.1021/ct300400x), PMID: [23341755](https://pubmed.ncbi.nlm.nih.gov/23341755/)

- Brown A**, Shao S, Murray J, Hegde RS, Ramakrishnan V. 2015. Structural basis for stop codon recognition in eukaryotes. *Nature* **524**:493–496. doi: [10.1038/nature14896](https://doi.org/10.1038/nature14896)
- Brunger AT**, Adams PD, Fromme P, Fromme R, Levitt M, Schröder GF. 2012. Improving the accuracy of macromolecular structure refinement at 7 Å resolution. *Structure* **20**:957–966. doi: [10.1016/j.str.2012.04.020](https://doi.org/10.1016/j.str.2012.04.020), PMID: [22681901](https://pubmed.ncbi.nlm.nih.gov/22681901/)
- Brünger AT**. 1988. *Crystallographic refinement by simulated annealing*. In: Isaacs N. W, Taylor M. R (Eds). *Crystallographic Computing 4: Techniques and New Technologies*. Oxford: Clarendon Press. doi: [10.1016/0022-2836\(88\)90211-2](https://doi.org/10.1016/0022-2836(88)90211-2)
- Chan KY**, Gumbart J, McGreevy R, Watermeyer JM, Sewell BT, Schulten K. 2011. Symmetry-restrained flexible fitting for symmetric EM maps. *Structure* **19**:1211–1218. doi: [10.1016/j.str.2011.07.017](https://doi.org/10.1016/j.str.2011.07.017), PMID: [21893283](https://pubmed.ncbi.nlm.nih.gov/21893283/)
- Chen VB**, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D Biological Crystallography* **66**:12–21. doi: [10.1107/S0907444909042073](https://doi.org/10.1107/S0907444909042073), PMID: [20057044](https://pubmed.ncbi.nlm.nih.gov/20057044/)
- Cheng Y**. 2015. Single-Particle Cryo-EM at crystallographic resolution. *Cell* **161**:450–457. doi: [10.1016/j.cell.2015.03.049](https://doi.org/10.1016/j.cell.2015.03.049), PMID: [25910205](https://pubmed.ncbi.nlm.nih.gov/25910205/)
- Cianfrocco MA**, Leschziner AE, Scheres SH. 2015. Low cost, high performance processing of single particle cryo-electron microscopy data in the cloud. *eLife* **4**:e06664. doi: [10.7554/eLife.06664](https://doi.org/10.7554/eLife.06664), PMID: [25955969](https://pubmed.ncbi.nlm.nih.gov/25955969/)
- Cowtan K**. 2006. The *Buccaneer* software for automated model building. 1. tracing protein chains. *Acta Crystallographica Section D Biological Crystallography* **62**:1002–1011. doi: [10.1107/S0907444906022116](https://doi.org/10.1107/S0907444906022116), PMID: [16929101](https://pubmed.ncbi.nlm.nih.gov/16929101/)
- Darnault C**, Volbeda A, Kim EJ, Legrand P, Vernède X, Lindahl PA, Fontecilla-Camps JC. 2003. Ni-Zn-[Fe4-S4] and Ni-Ni-[Fe4-S4] clusters in closed and open subunits of acetyl-CoA synthase/carbon monoxide dehydrogenase. *Nature Structural Biology* **10**:271–279. doi: [10.1038/nsb912](https://doi.org/10.1038/nsb912), PMID: [12627225](https://pubmed.ncbi.nlm.nih.gov/12627225/)
- DiMaio F**, Echols N, Headd JJ, Terwilliger TC, Adams PD, Baker D. 2013. Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nature Methods* **10**:1102–1104. doi: [10.1038/nmeth.2648](https://doi.org/10.1038/nmeth.2648), PMID: [24076763](https://pubmed.ncbi.nlm.nih.gov/24076763/)
- DiMaio F**, Song Y, Li X, Brunner MJ, Xu C, Conticello V, Egelman E, Marlovits TC, Cheng Y, Baker D. 2015. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nature Methods* **12**:361–365. doi: [10.1038/nmeth.3286](https://doi.org/10.1038/nmeth.3286), PMID: [25707030](https://pubmed.ncbi.nlm.nih.gov/25707030/)
- Fernández JJ**, Luque D, Castón JR, Carrascosa JL. 2008. Sharpening high resolution information in single particle electron cryomicroscopy. *Journal of Structural Biology* **164**:170–175. doi: [10.1016/j.jsb.2008.05.010](https://doi.org/10.1016/j.jsb.2008.05.010)
- Fischer N**, Neumann P, Konevega AL, Bock LV, Ficner R, Rodnina MV, Stark H. 2015. Structure of the E. coli ribosome-EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. *Nature* **520**:567–570. doi: [10.1038/nature14275](https://doi.org/10.1038/nature14275), PMID: [25707802](https://pubmed.ncbi.nlm.nih.gov/25707802/)
- Frauenfeld J**, Gumbart J, Sluis EO, Funes S, Gartmann M, Beatrix B, Mielke T, Berninghausen O, Becker T, Schulten K, Beckmann R, van der Sluis EO. 2011. Cryo-EM structure of the ribosome-SecYE complex in the membrane environment. *Nature Structural & Molecular Biology* **18**:614–621. doi: [10.1038/nsmb.2026](https://doi.org/10.1038/nsmb.2026), PMID: [21499241](https://pubmed.ncbi.nlm.nih.gov/21499241/)
- Freddolino PL**, Harrison CB, Liu Y, Schulten K. 2010. Challenges in protein-folding simulations. *Nature Physics* **6**:751–758. doi: [10.1038/nphys1713](https://doi.org/10.1038/nphys1713)
- Goh BC**, Perilla JR, England MR, Heyrana KJ, Craven RC, Schulten K. 2015. Atomic modeling of an immature retroviral lattice using molecular dynamics and mutagenesis. *Structure* **23**:1414–1425. doi: [10.1016/j.str.2015.05.017](https://doi.org/10.1016/j.str.2015.05.017), PMID: [26118533](https://pubmed.ncbi.nlm.nih.gov/26118533/)
- Grayson P**, Tajkhorshid E, Schulten K. 2003. Mechanisms of selectivity in channels and enzymes studied with interactive molecular dynamics. *Biophysical Journal* **85**:36–48. doi: [10.1016/S0006-3495\(03\)74452-X](https://doi.org/10.1016/S0006-3495(03)74452-X), PMID: [12829462](https://pubmed.ncbi.nlm.nih.gov/12829462/)
- Gumbart J**, Schreiner E, Trabuco LG, Chan K, Schulten K. 2011. Viewing the mechanisms of translation through the computational microscope. In: Frank J (Ed). *Molecular Machines in Biology*. p 142–157. doi: [10.1017/cbo9781139003704.010](https://doi.org/10.1017/cbo9781139003704.010)
- Hsin J**, Gumbart J, Trabuco LG, Villa E, Qian P, Hunter CN, Schulten K. 2009. Protein-induced membrane curvature investigated through molecular dynamics flexible fitting. *Biophysical Journal* **97**:321–329. doi: [10.1016/j.bpj.2009.04.031](https://doi.org/10.1016/j.bpj.2009.04.031), PMID: [19580770](https://pubmed.ncbi.nlm.nih.gov/19580770/)
- Humphrey W**, Dalke A, Schulten K. 1996. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**:33–38. doi: [10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5)
- Jiang W**, Phillips JC, Huang L, Fajer M, Meng Y, Gumbart JC, Luo Y, Schulten K, Roux B. 2014. Generalized scalable multiple copy algorithms for molecular dynamics simulations in NAMD. *Computer Physics Communications* **185**:908–916. doi: [10.1016/j.cpc.2013.12.014](https://doi.org/10.1016/j.cpc.2013.12.014), PMID: [24944348](https://pubmed.ncbi.nlm.nih.gov/24944348/)
- Jo S**, Kim T, Iyer VG, Im W. 2008. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **29**:1859–1865. doi: [10.1002/jcc.20945](https://doi.org/10.1002/jcc.20945)
- Jolley CC**, Wells SA, Fromme P, Thorpe MF. 2008. Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. *Biophysical Journal* **94**:1613–1621. doi: [10.1529/biophysj.107.115949](https://doi.org/10.1529/biophysj.107.115949), PMID: [17993504](https://pubmed.ncbi.nlm.nih.gov/17993504/)
- Karplus M**. 1990. Molecular dynamics: applications to proteins. In: Rivail J.-L (Ed). *Modelling of Molecular Structures and Properties, Volume 71 of Studies in Physical and Theoretical Chemistry*. Amsterdam: Elsevier Science Publishers. p 427–461.

- Klauda JB**, Venable RM, Freites JA, O'Connor JW, Tobias DJ, Mondragon-Ramirez C, Vorobyov I, MacKerell AD, Pastor RW. 2010. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *The Journal of Physical Chemistry B* **114**:7830–7843. doi: [10.1021/jp101759q](https://doi.org/10.1021/jp101759q), PMID: [20496934](https://pubmed.ncbi.nlm.nih.gov/20496934/)
- Kovacs JA**, Yeager M, Abagyan R. 2008. Damped-dynamics flexible fitting. *Biophysical Journal* **95**:3192–3207. doi: [10.1529/biophysj.108.132357](https://doi.org/10.1529/biophysj.108.132357), PMID: [18586844](https://pubmed.ncbi.nlm.nih.gov/18586844/)
- Kucukelbir A**, Sigworth FJ, Tagare HD. 2014. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11**:63–65. doi: [10.1038/nmeth.2727](https://doi.org/10.1038/nmeth.2727), PMID: [24213166](https://pubmed.ncbi.nlm.nih.gov/24213166/)
- Kuzmanic A**, Pannu NS, Zagrovic B. 2014. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nature Communications* **5**:3220. doi: [10.1038/ncomms4220](https://doi.org/10.1038/ncomms4220), PMID: [24504120](https://pubmed.ncbi.nlm.nih.gov/24504120/)
- Langer G**, Cohen SX, Lamzin VS, Perrakis A. 2008. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols* **3**:1171–1179. doi: [10.1038/nprot.2008.91](https://doi.org/10.1038/nprot.2008.91), PMID: [18600222](https://pubmed.ncbi.nlm.nih.gov/18600222/)
- Leaver-Fay A**, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YE, Fleishman SJ, Corn JE, Kim DE, Lyskov S, et al. 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology* **487**:545–574. doi: [10.1016/B978-0-12-381270-4.00019-6](https://doi.org/10.1016/B978-0-12-381270-4.00019-6), PMID: [21187238](https://pubmed.ncbi.nlm.nih.gov/21187238/)
- Leschziner AE**, Nogales E. 2007. Visualizing flexibility at molecular resolution: analysis of heterogeneity in single-particle electron microscopy reconstructions. *Annual Review of Biophysics and Biomolecular Structure* **36**:43–62. doi: [10.1146/annurev.biophys.36.040306.132742](https://doi.org/10.1146/annurev.biophys.36.040306.132742), PMID: [17201674](https://pubmed.ncbi.nlm.nih.gov/17201674/)
- Li X**, Mooney P, Zheng S, Booth CR, Braunfeld MB, Gubbens S, Agard DA, Cheng Y. 2013. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**:584–590. doi: [10.1038/nmeth.2472](https://doi.org/10.1038/nmeth.2472), PMID: [23644547](https://pubmed.ncbi.nlm.nih.gov/23644547/)
- Li Q**, Wanderling S, Paduch M, Medovoy D, Singharoy A, McGreevy R, Villalba-Galea CA, Hulse RE, Roux B, Schulten K, Kossiakoff A, Perozo E. 2014. Structural mechanism of voltage-dependent gating in an isolated voltage-sensing domain. *Nature Structural & Molecular Biology* **21**:244–252. doi: [10.1038/nsmb.2768](https://doi.org/10.1038/nsmb.2768), PMID: [24487958](https://pubmed.ncbi.nlm.nih.gov/24487958/)
- Liao M**, Cao E, Julius D, Cheng Y. 2013. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**:107–112. doi: [10.1038/nature12822](https://doi.org/10.1038/nature12822), PMID: [24305160](https://pubmed.ncbi.nlm.nih.gov/24305160/)
- Lindert S**, McCammon JA. 2015. Improved cryoEM-guided iterative molecular dynamics–rosetta protein structure refinement protocol for high precision protein structure prediction. *Journal of Chemical Theory and Computation* **11**:1337–1346. doi: [10.1021/ct500995d](https://doi.org/10.1021/ct500995d), PMID: [25883538](https://pubmed.ncbi.nlm.nih.gov/25883538/)
- Liu C**, Xiong Y. 2014. Electron density sharpening as a general technique in crystallographic studies. *Journal of Molecular Biology* **426**:980–993. doi: [10.1016/j.jmb.2013.11.014](https://doi.org/10.1016/j.jmb.2013.11.014), PMID: [24269527](https://pubmed.ncbi.nlm.nih.gov/24269527/)
- López-Blanco JR**, Chacón P. 2013. iMODFIT: efficient and robust flexible fitting based on vibrational analysis in internal coordinates. *Journal of Structural Biology* **184**:261–270. doi: [10.1016/j.jsb.2013.08.010](https://doi.org/10.1016/j.jsb.2013.08.010), PMID: [23999189](https://pubmed.ncbi.nlm.nih.gov/23999189/)
- Lu P**, Bai XC, Ma D, Xie T, Yan C, Sun L, Yang G, Zhao Y, Zhou R, Scheres SH, Shi Y. 2014. Three-dimensional structure of human γ -secretase. *Nature* **512**:166–170. doi: [10.1038/nature13567](https://doi.org/10.1038/nature13567), PMID: [25043039](https://pubmed.ncbi.nlm.nih.gov/25043039/)
- Matthies D**, Dalmás O, Borgnia MJ, Dominik PK, Merk A, Rao P, Reddy BG, Islam S, Bartesaghi A, Perozo E, Subramaniam S. 2016. Cryo-EM structures of the magnesium channel CorA reveal symmetry break upon gating. *Cell* **164**:747–756. doi: [10.1016/j.cell.2015.12.055](https://doi.org/10.1016/j.cell.2015.12.055), PMID: [26871634](https://pubmed.ncbi.nlm.nih.gov/26871634/)
- McGreevy R**, Singharoy A, Li Q, Zhang J, Xu D, Perozo E, Schulten K. 2014. xMDFF: molecular dynamics flexible fitting of low-resolution X-ray structures. *Acta Crystallographica Section D Biological Crystallography* **70**:2344–2355. doi: [10.1107/S1399004714013856](https://doi.org/10.1107/S1399004714013856)
- McGreevy R**, Teo I, Singharoy A, Schulten K. 2016. Advances in the molecular dynamics flexible fitting method for cryo-EM modeling. *Methods* **100**:50–60. doi: [10.1016/j.ymeth.2016.01.009](https://doi.org/10.1016/j.ymeth.2016.01.009), PMID: [26804562](https://pubmed.ncbi.nlm.nih.gov/26804562/)
- Milazzo AC**, Cheng A, Moeller A, Lyumkis D, Jacovetty E, Polukas J, Ellisman MH, Xuong NH, Carragher B, Potter CS. 2011. Initial evaluation of a direct detection device detector for single particle cryo-electron microscopy. *Journal of Structural Biology* **176**:404–408. doi: [10.1016/j.jsb.2011.09.002](https://doi.org/10.1016/j.jsb.2011.09.002), PMID: [21933715](https://pubmed.ncbi.nlm.nih.gov/21933715/)
- Murshudov GN**, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA. 2011. REFMAC 5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D Biological Crystallography* **67**:355–367. doi: [10.1107/S0907444911001314](https://doi.org/10.1107/S0907444911001314)
- Neutze R**, Brändén G, Schertler GF. 2015. Membrane protein structural biology using X-ray free electron lasers. *Current Opinion in Structural Biology* **33**:115–125. doi: [10.1016/j.sbi.2015.08.006](https://doi.org/10.1016/j.sbi.2015.08.006), PMID: [26342349](https://pubmed.ncbi.nlm.nih.gov/26342349/)
- Noble AJ**, Zhang Q, O'Donnell J, Hariri H, Bhattacharya N, Marshall AG, Stagg SM. 2013. A pseudoatomic model of the COPII cage obtained from cryo-electron microscopy and mass spectrometry. *Nature Structural & Molecular Biology* **20**:167–173. doi: [10.1038/nsmb.2467](https://doi.org/10.1038/nsmb.2467), PMID: [23262493](https://pubmed.ncbi.nlm.nih.gov/23262493/)
- Orzechowski M**, Tama F. 2008. Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophysical Journal* **95**:5692–5705. doi: [10.1529/biophysj.108.139451](https://doi.org/10.1529/biophysj.108.139451), PMID: [18849406](https://pubmed.ncbi.nlm.nih.gov/18849406/)
- Phillips JC**, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K. 2005. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26**:1781–1802. doi: [10.1002/jcc.20289](https://doi.org/10.1002/jcc.20289)
- Rosenthal PB**, Henderson R. 2003. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of Molecular Biology* **333**:721–745. doi: [10.1016/j.jmb.2003.07.013](https://doi.org/10.1016/j.jmb.2003.07.013), PMID: [14568533](https://pubmed.ncbi.nlm.nih.gov/14568533/)

- Schröder GF**, Brunger AT, Levitt M. 2007. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* **15**:1630–1641. doi: [10.1016/j.str.2007.09.021](https://doi.org/10.1016/j.str.2007.09.021), PMID: [18073112](https://pubmed.ncbi.nlm.nih.gov/18073112/)
- Schröder GF**, Levitt M, Brunger AT. 2010. Super-resolution biomolecular crystallography with low-resolution data. *Nature* **464**:1218–1222. doi: [10.1038/nature08892](https://doi.org/10.1038/nature08892), PMID: [20376006](https://pubmed.ncbi.nlm.nih.gov/20376006/)
- Sener M**, Hsin J, Trabuco LG, Villa E, Qian P, Hunter CN, Schulten K. 2009. Structural model and excitonic properties of the dimeric RC-LH1-PufX complex from *Rhodobacter sphaeroides*. *Chemical Physics* **357**:188–197. doi: [10.1016/j.chemphys.2009.01.003](https://doi.org/10.1016/j.chemphys.2009.01.003), PMID: [20161332](https://pubmed.ncbi.nlm.nih.gov/20161332/)
- Shaikh TR**, Gao H, Baxter WT, Asturias FJ, Boisset N, Leith A, Frank J. 2008. SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nature Protocols* **3**:1941–1974. doi: [10.1038/nprot.2008.156](https://doi.org/10.1038/nprot.2008.156)
- Singharoy A**, Polavarapu A, Joshi H, Baik MH, Ortoleva P. 2013. Epitope fluctuations in the human papillomavirus are under dynamic allosteric control: a computational evaluation of a new vaccine design strategy. *Journal of the American Chemical Society* **135**:18458–18468. doi: [10.1021/ja407489r](https://doi.org/10.1021/ja407489r), PMID: [24199651](https://pubmed.ncbi.nlm.nih.gov/24199651/)
- Singharoy A**, Venkatakrishnan B, Liu Y, Mayne CG, Lee S, Chen CH, Zlotnick A, Schulten K, Flood AH. 2015. Macromolecular crystallography for synthetic abiological molecules: combining xMDF and PHENIX for structure determination of cyanostar macrocycles. *Journal of the American Chemical Society* **137**:8810–8818. doi: [10.1021/jacs.5b04407](https://doi.org/10.1021/jacs.5b04407), PMID: [26121416](https://pubmed.ncbi.nlm.nih.gov/26121416/)
- Stone JE**, Gullingsrud J, Grayson P, Schulten K. 2001. A system for interactive molecular dynamics simulation. In: Hughes JF, Séquin CH (Eds). *ACM Symposium on Interactive 3D Graphics*. New York: ACM SIGGRAPH. p 191–194. doi: [10.1145/364338.364398](https://doi.org/10.1145/364338.364398)
- Stone JE**, Phillips JC, Freddolino PL, Hardy DJ, Trabuco LG, Schulten K. 2007. Accelerating molecular modeling applications with graphics processors. *Journal of Computational Chemistry* **28**:2618–2640. doi: [10.1002/jcc.20829](https://doi.org/10.1002/jcc.20829)
- Stone JE**, McGreevy R, Isralewitz B, Schulten K. 2014. GPU-accelerated analysis and visualization of large structures solved by molecular dynamics flexible fitting. *Faraday Discussion*. **169**:265–283. doi: [10.1039/C4FD00005F](https://doi.org/10.1039/C4FD00005F), PMID: [25340325](https://pubmed.ncbi.nlm.nih.gov/25340325/)
- Sugita Y**, Okamoto Y. 1999. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314**:141–151. doi: [10.1016/S0009-2614\(99\)01123-9](https://doi.org/10.1016/S0009-2614(99)01123-9)
- Suhre K**, Navaza J, Sanejouand Y-H. 2006. NORMA : a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. *Acta Crystallographica Section D Biological Crystallography* **62**:1098–1100. doi: [10.1107/S090744490602244X](https://doi.org/10.1107/S090744490602244X)
- Tama F**, Miyashita O, Brooks III CL. 2004. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *Journal of Structural Biology* **147**:315–326. doi: [10.1016/j.jsb.2004.03.002](https://doi.org/10.1016/j.jsb.2004.03.002)
- Tanner DE**, Chan KY, Phillips JC, Schulten K. 2011. Parallel generalized born implicit solvent calculations with NAMD. *Journal of Chemical Theory and Computation* **7**:3635–3642. doi: [10.1021/ct200563j](https://doi.org/10.1021/ct200563j), PMID: [22121340](https://pubmed.ncbi.nlm.nih.gov/22121340/)
- Topf M**, Baker ML, Marti-Renom MA, Chiu W, Sali A. 2006. Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. *Journal of Molecular Biology* **357**:1655–1668. doi: [10.1016/j.jmb.2006.01.062](https://doi.org/10.1016/j.jmb.2006.01.062), PMID: [16490207](https://pubmed.ncbi.nlm.nih.gov/16490207/)
- Topf M**, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. 2008. Protein structure fitting and refinement guided by cryo-EM density. *Structure* **16**:295–307. doi: [10.1016/j.str.2007.11.016](https://doi.org/10.1016/j.str.2007.11.016), PMID: [18275820](https://pubmed.ncbi.nlm.nih.gov/18275820/)
- Trabuco LG**, Villa E, Mitra K, Frank J, Schulten K. 2008. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**:673–683. doi: [10.1016/j.str.2008.03.005](https://doi.org/10.1016/j.str.2008.03.005), PMID: [18462672](https://pubmed.ncbi.nlm.nih.gov/18462672/)
- Trabuco LG**, Villa E, Schreiner E, Harrison CB, Schulten K. 2009. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **49**:174–180. doi: [10.1016/j.ymeth.2009.04.005](https://doi.org/10.1016/j.ymeth.2009.04.005), PMID: [19398010](https://pubmed.ncbi.nlm.nih.gov/19398010/)
- Trabuco LG**, Schreiner E, Gumbart J, Hsin J, Villa E, Schulten K. 2011. Applications of the molecular dynamics flexible fitting method. *Journal of Structural Biology* **173**:420–427. doi: [10.1016/j.jsb.2010.09.024](https://doi.org/10.1016/j.jsb.2010.09.024), PMID: [20932910](https://pubmed.ncbi.nlm.nih.gov/20932910/)
- Unger VM**. 2001. Electron cryomicroscopy methods. *Current Opinion in Structural Biology* **11**:548–554. doi: [10.1016/S0959-440X\(00\)00260-8](https://doi.org/10.1016/S0959-440X(00)00260-8)
- Velazquez-Muriel JA**, Valle M, Santamaría-Pang A, Kakadiaris IA, Carazo JM. 2006. Flexible fitting in 3D-EM guided by the structural variability of protein superfamilies. *Structure* **14**:1115–1126. doi: [10.1016/j.str.2006.05.013](https://doi.org/10.1016/j.str.2006.05.013), PMID: [16843893](https://pubmed.ncbi.nlm.nih.gov/16843893/)
- Villa E**, Sengupta J, Trabuco LG, LeBarron J, Baxter WT, Shaikh TR, Grassucci RA, Nissen P, Ehrenberg M, Schulten K, Frank J. 2009. Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis. *PNAS* **106**:1063–1068. doi: [10.1073/pnas.0811370106](https://doi.org/10.1073/pnas.0811370106), PMID: [19122150](https://pubmed.ncbi.nlm.nih.gov/19122150/)
- Wade RH**. 1992. A brief look at imaging and contrast transfer. *Ultramicroscopy* **46**:145–156. doi: [10.1016/0304-3991\(92\)90011-8](https://doi.org/10.1016/0304-3991(92)90011-8)
- Wickles S**, Singharoy A, Andreani J, Seemayer S, Bischoff L, Berninghausen O, Soeding J, Schulten K, van der Sluis EO, Beckmann R. 2014. A structural model of the active ribosome-bound membrane protein insertase YidC. *eLife* **3**:e03035. doi: [10.7554/eLife.03035](https://doi.org/10.7554/eLife.03035), PMID: [25012291](https://pubmed.ncbi.nlm.nih.gov/25012291/)
- Williams L**. 1983. Pyramidal parametrics. In: *Proceedings of the 10th Annual Conference on Computer Graphics and Interactive Techniques*. New York: ACM. p 1–11. doi: [10.1145/800059.801126](https://doi.org/10.1145/800059.801126)

- Wriggers W.** 2010. Using situs for the integration of multi-resolution structures. *Biophysical Reviews* **2**:21–27. doi: [10.1007/s12551-009-0026-3](https://doi.org/10.1007/s12551-009-0026-3), PMID: [20174447](https://pubmed.ncbi.nlm.nih.gov/20174447/)
- Wu X, Subramaniam S, Case DA, Wu KW, Brooks BR.** 2013. Targeted conformational search with map-restrained self-guided Langevin dynamics: application to flexible fitting into electron microscopic density maps. *Journal of Structural Biology* **183**:429–440. doi: [10.1016/j.jsb.2013.07.006](https://doi.org/10.1016/j.jsb.2013.07.006), PMID: [23876978](https://pubmed.ncbi.nlm.nih.gov/23876978/)
- Zhao G, Perilla JR, Yufenyuy EL, Meng X, Chen B, Ning J, Ahn J, Gronenborn AM, Schulten K, Aiken C, Zhang P.** 2013. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **497**:643–646. doi: [10.1038/nature12162](https://doi.org/10.1038/nature12162), PMID: [23719463](https://pubmed.ncbi.nlm.nih.gov/23719463/)
- Zhao M, Wu S, Zhou Q, Vivona S, Cipriano DJ, Cheng Y, Brunger AT.** 2015. Mechanistic insights into the recycling machine of the SNARE complex. *Nature* **518**:61–67. doi: [10.1038/nature14148](https://doi.org/10.1038/nature14148), PMID: [25581794](https://pubmed.ncbi.nlm.nih.gov/25581794/)

Appendix 1

1 Theory of cascade MDFF

The MDFF potential is given by

$$U_{EM}(\mathbf{R}) = \sum_i w_i V_{EM}(\mathbf{r}_i), \quad (7)$$

where \mathbf{R} is the system configuration, i labels the atoms in the structure, w_i and \mathbf{r}_i are weight and position, respectively, assigned to atom i , and V_{EM} is the potential map obtained from the EM density map.

Subjecting the potential map V_{EM} to a Gaussian blur increases its widths in a manner described as follows. The Gaussian blur process produces a potential map V_σ from the potential map V_{EM} through convolution with a normalized Gaussian of specified width σ :

$$V_\sigma(\mathbf{r}) = \int d\mathbf{r}' G(\mathbf{r}; \mathbf{r}', \sigma) V_{EM}(\mathbf{r}'), \quad (8)$$

$$G(\mathbf{r}; \mathbf{r}', \sigma) = A(\sigma) \exp\left(-\frac{\|\mathbf{r} - \mathbf{r}'\|^2}{2\sigma^2}\right), \quad (9)$$

$$A(\sigma) = \frac{1}{(2\pi\sigma^2)^{3/2}}. \quad (10)$$

One can characterize the resolution of $V_{EM}(\mathbf{r})$ explicitly by assuming that it can be written as a sum of Gaussians.

$$V_{EM}(\mathbf{r}) = \sum_n c_n G(\mathbf{r}; \mathbf{r}'_n, \sigma'_n), \quad (11)$$

where c_n are weighting factors, \mathbf{r}'_n and σ'_n are, respectively, the centers and widths of the component Gaussians. Substituting the above expression into **Equation 8** yields

$$V_\sigma(\mathbf{r}) = \int d\mathbf{r}' G(\mathbf{r}; \mathbf{r}', \sigma) \sum_n c_n G(\mathbf{r}'; \mathbf{r}'_n, \sigma'_n) \quad (12)$$

$$= \sum_n c_n A(\sigma) A'_n(\sigma'_n) \int d\mathbf{r}' \exp\left(-\frac{\|\mathbf{r} - \mathbf{r}'\|^2}{2\sigma^2} - \frac{\|\mathbf{r}' - \mathbf{r}'_n\|^2}{2\sigma_n'^2}\right), \quad (13)$$

Evaluation of the above expression gives

$$V_\sigma(\mathbf{r}) = \sum_n C_n \exp\left(-\frac{\|\mathbf{r} - \mathbf{r}'_n\|^2}{2(\sigma^2 + \sigma_n'^2)}\right), \quad (14)$$

$$C_n = c_n A(\sigma) A'_n(\sigma'_n) (2\pi)^{3/2} \left(\frac{\sigma\sigma'_n}{\sqrt{\sigma^2 + \sigma_n'^2}}\right)^3 \quad (15)$$

$$= c_n [2\pi(\sigma^2 + \sigma_n'^2)]^{-3/2}. \quad (16)$$

Note that setting $\sigma = 0$ A in (**Equations 14–16**) recovers the expression for the initial map in **Equation 11**, i.e. $V_0(\mathbf{r}) = V(\mathbf{r})$. On the other hand, increasing σ results in an increase in the

widths of the component Gaussians, and leads in turn to an increase in the range of the MDFF force $\mathbf{F}_\sigma(\mathbf{r})$, where

$$\mathbf{F}_\sigma(\mathbf{r}) = -\nabla V_\sigma(\mathbf{r}) \quad (17)$$

$$= \sum_n \frac{C_n}{\sigma^2 + \sigma_n^2} \exp\left(-\frac{\|\mathbf{r} - \mathbf{r}'_n\|^2}{2(\sigma^2 + \sigma_n^2)}\right) (\mathbf{r} - \mathbf{r}'_n). \quad (18)$$

2 MD simulation setup

Unless otherwise stated, all simulations reported in the present study used the following MD parameters. MDFF simulations were run in vacuum, maintained at 300 K via a Langevin thermostat and employed a time step of 1 fs. Secondary structure restraints were imposed using NAMM's Extra Bonds function to prevent loss of secondary structure due to strong MDFF guiding forces.

MD simulations were prepared using CHARMM-GUI (Jo *et al.*, 2008) and run under NPT conditions, maintained at 303.15 K temperature and 1 atm pressure using a Langevin thermostat and barostat. All systems were parametrized using the CHARMM36 force field (Klauda *et al.*, 2010; Best *et al.*, 2012). Structures were solvated in explicit water (TIP3P model) boxes, with at least 15 Å separation between structure and water box boundaries. Particle-mesh Ewald electrostatics was used and the time step was 2 fs. For simulations of the reported structures of β -galactosidase, backbone atoms were held fixed while a minimization over 1000 time steps was performed, followed by 32-ns and 40-ns equilibration for the 3.2-Å and 2.2-Å structures, respectively. Following the equilibration step, production runs of 30 ns were performed for both structures.

In the case of TRPV1, the channel was embedded in a membrane of standard lipid composition POPE, POPC, POPG at ratio 2:1:1. Initial runs involved minimization over 1000 steps and 1-ns equilibration of the lipid tails, with all other atoms fixed. In the following MD run, minimization was performed over 5000 steps and equilibration was performed for 3 ns, with protein backbone atoms held fixed. Finally, the entire system was equilibrated for a further 6.4 ns. During the equilibration, C-terminal residues 752 to 762 were harmonically restrained because a substantial C-terminal segment was missing in the structure.

3 Preparation of initial test structures

The initial test structure of β -galactosidase was obtained by subjecting the reported structure (Bartesaghi *et al.*, 2014) to a 4-ns MD simulation at a temperature of 300 K in vacuum and using secondary structure restraints. Trajectory frames recorded at 2-ps intervals were evaluated for backbone RMSDs with respect to the reported structure. A frame with an RMSD value of 7.6 Å (Figure 2—figure supplement 2a) and lowest global cross-correlation with respect to the reported map was picked to be the initial test structure. The structural quality measure of this model is provided in Table 2. A second search model was prepared by repeating the same protocol but at 1000 K. This search model is also characterized by an RMSD of 7.6 Å but now features a more distorted local structure as measured in terms of increased rotamer and Ramachandran outliers (Supplementary file 1F).

The initial test structure of TRPV1 was also derived from a reported structure (Liao *et al.*, 2013). In order to render the disjointed reported structure contiguous for correct structural dynamics during simulation, the missing loop region (residues 503 to 506) was added by hand. Additionally, the substantial ankyrin repeat region (residues 111 to 198) was removed because the corresponding density was missing from the map. For the purpose of testing the robustness of cMDFF and contrasting its performance with that of direct MDFF, the

structure was distorted (see **Figure 2—figure supplement 2b**) during an interactive MD (**Stone et al., 2001; Grayson et al., 2003**) simulation, subjecting residues 199 to 430 in one subunit's extramembrane domain to a series of transformations, consisting roughly of a polar angle change of 15° toward the cytoplasmic pole followed by an azimuthal rotation of 30° , so that the backbone RMSD of the transformed region was about 22 Å relative to the original structure.

4 Protocol efficiency

In the present study, the convergence of fitting results of TRPV1 and β -galactosidase for different protocols can be evaluated through RMSD value employing the reported de novo structures as reference. **Figure 2—figure supplement 3** shows the time evolution of RMSD for cMDFF, ReMDFF, and direct MDFF for the two proteins, It should be noted that the plots do not include the final refinement step, which is the same across all three protocols.

cMDFF and ReMDFF can be seen to reach similar RMSD levels, outperforming direct MDFF. ReMDFF converges more quickly than cMDFF in either example. Of the 6 replicas employed for the ReMDFF of TRPV1 two resulted in poorly fitted structures, having become trapped in density minima even after exchanging with the lowest-resolution map chosen with $\sigma = 5$ Å. All replicas can be monitored during the simulation and poorly fitted ones can be discarded by a user. It is also worth noting that the region of the TRPV1 map to which both cMDFF and ReMDFF successfully fitted the model, characterizes a diverse range of local resolutions from 4 Å to 6 Å challenging thus the conformational sampling capability of any flexible fitting technique. For the same reason, this region was avoided during Rosetta refinements of TRPV1 (**DiMaio et al., 2015**) but is addressed now via MDFF.

5 Global vs. local measures of fit

The cross-correlation coefficient calculated over an entire structure, termed global cross-correlation coefficient (GCC), has been a popular indicator of goodness-of-fit of a structure to a corresponding EM density map. However, averaging over the entire structure smears out potentially useful local structure information and in some cases, can be misleading (see **Figure 2—figure supplement 1**), since GCC cannot distinguish between correct and wrong assignments of residues to a given map region as long as the residues are equally well fitted.

Local measures of fit allow one to assess every part of the structure individually. For example, local cross-correlation coefficients (LCCs) (**Stone et al., 2014**) were tracked over the course of the cMDFF simulations of β -galactosidase and TRPV1 (**Figure 2—figure supplement 4**). The improvement in LCC of the majority of residues in each case lends greater confidence in the fitting result. At the same time, residues that have relatively lower LCCs can be identified and given attention (**Stone et al., 2014; McGreevy et al., 2016**).

6 Fourier shell coefficients

Fourier Shell Coefficients (FSCs) can be used as a means of evaluating quality of fit by comparing the degree of similarity between the original map and a simulated map derived from the structure to be evaluated, using the simulated map feature of VMD's MDFF package (**Trabuco et al., 2008, 2009**) and the same voxel size as the original map.

In the present study, FSC curves for fitted TRPV1 and β -galactosidase structures were calculated via the FSC operation in SPIDER (**Shaikh et al., 2008**) using a shell width of 0.5 reciprocal space units, and resolution cutoff of half the voxel size. In the case of TRPV1, both the full structure and MDFF-fitted region were evaluated. The latter was obtained by

applying a mask of the region around residues 199 to 430 in the fitted structure to the simulated map, and in the reported structure to the original map.

As a means of summarizing comparisons by FSC, other studies have used 'integrated FSCs', a numerical measure obtained by integrating under the FSC plot within a predefined resolution interval. Two integrated FSC measures, corresponding to the intervals 3.4 Å to 10 Å and 5 Å to 10 Å, were obtained in the present study and tabulated in **Table 1**, and **Supplementary file 1A and 1C**.

7 Other simulations of β -galactosidase

Beyond the simulations reported in **Table 1**, further simulations were performed to explore the performance of MDFF within the contexts of further types of analyses. The first of these simulations was a direct MDFF simulation of the reported β -galactosidase structure, fitting only backbone atoms to the 3.2-Å map. The fitted structure was compared to the 'refined de novo' structure reported in Results. It was found that EMRinger scores were lower at 2.35 when only backbone atoms were fitted, compared to 4.23 when non-hydrogen atoms were fitted. This result suggests that even if the backbone is correctly placed, the MD force fields alone (i.e. CHARMM36 (**Best et al., 2012**) here) are incapable of providing sidechain geometries consistent with the map. Refinement of the sidechains will therefore require explicit fitting to the density, above and beyond the orientations captured by the force fields alone.

In the second simulation, the resulting structure of the cMDFF simulation was subjected to an equilibrium MD simulation in explicit solvent. As shown in **Figure 2—figure supplement 6**, the equilibrium RMSD fluctuations during this simulation ranged between 3.0 Å to 3.4 Å of the starting structure. It is worth noting that these RMSD values agree well with the 3.2 Å resolution limit of the β -galactosidase map. Thus, the result indicates that uncertainties of the map resolution reflects quantitatively the structural variations of the cMDFF-fitted β -galactosidase model at the room temperature. Consequently, this model is representative of the thermodynamic ensemble that the EM map characterizes. Further analysis of the quality of the 3.2-Å β -galactosidase model is presented in the Model Validation subsection of Results.

The third set of simulations takes advantage of a unique opportunity, presented by the availability of two different maps of the same structure, at resolutions of 3.2 Å and 2.2 Å, to compare the results of fitting β -galactosidase to maps of different resolutions. The reported structures were subjected to direct MDFF simulation for 0.7 and 1 ns for the 3.2-Å and 2.2-Å models, respectively. The RMSF for each residue is calculated over consecutive 10-ps windows during the fitting. The RMSF values for all residues, including those for the PETG binding pockets (**Bartesaghi et al., 2015**) are plotted in **Figure 4—figure supplement 4**, reflecting smaller fluctuations during the fitting to the 2.2-Å map than in the 3.2-Å one. The relationship between fluctuation and map quality is examined in greater detail in Results, and imply that the RMSF of the fitted structure correlates negatively with the resolution of the corresponding map. Thus, the smaller fluctuations of the 2.2 Å structure than those of the 3.2 Å one further validate our proposed RMSF-map quality relationship of **Figure 4**.

8 MDFF refinement for TRPV1

Similar to the β -galactosidase results in **Table 1**, the refinement of TRPV1 was significantly better with cMDFF and ReMDFF than with direct MDFF, but starting with much poorer search models than what were employed for the β -galactosidase refinements. The TRPV1 refinement results are summarized in **Supplementary file 1A**: (i) RMSD of the fitted structure with respect to the reported de novo structure is 7.9 Å for direct MDFF, higher

than the 2.4 Å and 2.5 Å RMSD values for cMDFF and ReMDFF, respectively (**Figure 2—figure supplement 3b**); (ii) EMRinger scores for cMDFF and ReMDFF are 1.68 and 1.99 respectively, higher than the 1.51 score obtained for direct MDFF; (iii) MolProbity scores (**Chen et al., 2010**) are 2.4 and 2.5 for cMDFF and ReMDFF, smaller than the 7.9 score for direct MDFF, implying fewer, less severe steric clashes and fewer poor rotamers in the former than in the latter; (iv) integrated FSC (iFSC2, for the range 3.4–10 Å obtained as described in Appendix 1 - Section 6 above), attains higher values of 2.62 and 2.75 for cMDFF and ReMDFF respectively, than the 1.79 value for direct MDFF. iFSC1, corresponding to the lower resolution range of 5–10 Å is found to behave similarly to iFSC2; and (v) GCCs improve from an initial value of 0.16 to 0.50, 0.54 and 0.53 for direct, cMDFF, and ReMDFF protocols, respectively. Similarly, typical residue LCC values improve from 0 to 0.5 or higher, as shown in **Figure 2—figure supplement 4b**.

Measures of structural quality for the above fits are tabulated in **Supplementary file 1B**. Cross-validation with half-maps was also performed on the cMDFF structure, as per the β -galactosidase simulations, to ensure that it was not overfitted. As in the case of β -galactosidase, iFSC and EMRinger scores for direct and cross comparisons were similar. FSC analysis results are described in Appendix 1 - Section 11.

9 MDFF fitting of TRPV1 TM domain

Noting that the transmembrane (TM) region of the TRPV1 map is better resolved than the map for the extramembrane region, we performed 200 ps of MDFF on a truncated TM portion of the structure (residues 381 to 695) to characterize the performance of MDFF in the transmembrane region alone.

The TM region had previously been refined employing Rosetta tools (**Leaver-Fay et al., 2011**), providing an opportunity for comparison. Two MDFF simulations were performed, the first with only non-hydrogen sidechain atoms coupled to the density and harmonic restraints holding backbone atoms in the configuration of the reported structure (**Liao et al., 2013**) and another with all non-hydrogen atoms coupled to the density and the backbone restraints removed.

MDFF characteristics for fitting the isolated TM region of TRPV1 are summarized in **Supplementary file 1C**. Quality of fit measures, namely EMRinger and iFSC, for the backbone-restrained simulations were lower than those of the Rosetta-derived structure. However, MolProbity scores for the MDFF-derived structures are better than those of Rosetta. Allowing the backbone to be fitted into the map without restraints from the reported structure substantially improved the quality of fit measures so that they are comparable to those of Rosetta's, while maintaining a lower MolProbity score.

10 RMSF as a local measure of map quality

Per-residue Root mean square fluctuation (RMSF) has been found to be a good local measure of model fit, decreasing in relation to the increasing quality of fit over a cMDFF simulation, as illustrated in **Figure 4—figure supplement 1**. In addition to being a measure of fit, RMSF values can also be used for the analysis of local map quality. When monitored during MDFF refinement simulations, RMSF values serve as a good measure of local map quality, decreasing linearly with increase in local map resolution, as demonstrated in a number of test cases (**Figure 4**), including the 20S proteasome (**Figure 4—figure supplement 3**). In contrast to RMSF, other local measures of fit, such as LCC and EMRinger scores, are not necessarily good indicators of local map quality, as is evident from their poor correlation with local map resolution in **Figure 4—figure supplement 2**.

11 Cross-validation of MDFF-fitted structures

To demonstrate that the over-fitting does not occur during MDFF refinements, which is also fairly representative of cMDFF and ReMDFF refinements, the reported de novo structures of β -galactosidase and TRPV1 were each directly fitted to two half-maps (labelled 1 and 2) of the corresponding reported EM map, (*Bartesaghi et al., 2014*) for β -galactosidase and (*Liao et al., 2013*) for TRPV1. Subsequently, simulated maps were created from the fitted structures using VMD's MDFF plugin and resolution settings equivalent to the reported maps. In total, there were two simulated maps, also with labels 1 and 2 corresponding to the half-map from which the fits were obtained, for each protein. FSC plots describing the direct comparison of simulated maps with the corresponding half-maps (e.g. simulated map 1 with half-map 1) as well as the cross comparison of simulated maps with the non-corresponding half-maps (e.g. simulated map 1 with half-map 2) were created. **Figure 2—figure supplement 8** show the plots for β -galactosidase and TRPV1, respectively. The high degree of similarity between the cross comparisons as well as between cross comparisons and direct comparisons indicate a very low degree of over-fitting. In fact, iFSC values calculated for the plots (see Figures) are practically uniform. EMRinger scores for the same sets of comparisons were also calculated. For β -galactosidase, the EMRinger scores were 3.25 for simulated map 1 against half-map 1, 2.97 for simulated map 2 against half-map 2, 2.92 for simulated map 1 against half-map 2, and 2.81 for simulated map 2 against half-map 1; these numbers are fairly comparable to the EMRinger scores with the full maps as presented in **Table 1**. For TRPV1 again, the EMRinger scores were 1.43 for all comparisons. The high degree of similarity between EMRinger scores for the different comparisons corroborates the favorable conclusion drawn from the FSC calculations.