Title

CryoFold: Determining protein structures and data-guided ensembles from cryo-EM density maps

Journal

Matter, 4(10)

ISSN

2590-2393

Authors

Shekhar, Mrinal
Terashi, Genki
Gupta, Chitrak
et al.

Publication Date

2021-10-01

DOI

10.1016/j.matt.2021.09.004

Peer reviewed

HHS Public Access

Author manuscript

Matter. Author manuscript; available in PMC 2022 October 06.

Published in final edited form as:

Matter. 2021 October 06; 4(10): 3195–3216. doi:10.1016/j.matt.2021.09.004.

CryoFold: determining protein structures and data-guided ensembles from cryo-EM density maps

Mrinal Shekhar[1], **Genki Terashi**[2], **Chitrak Gupta**[3,4], **Daipayan Sarkar**[2,3], **Gaspard Debussche**[5], **Nicholas J. Sisco**[3,6], **Jonathan Nguyen**[3,4], **Arup Mondal**[9], **John Vant**[3,4], **Petra Fromme**[3,4], **Wade D. Van Horn**[3,6], **Emad Tajkhorshid**[1], **Daisuke Kihara**[2,8], **Ken Dill**[7], **Alberto Perez**[9], **Abhishek Singharoy**[3,4]

[1]Center for Biophysics and Quantitative Biology, Department of Biochemistry, NIH Center for Macromolecular Modeling and Bioinformatics, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, USA

[2]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA

[3]The School of Molecular Sciences, Arizona State University, Tempe, AZ 85287, USA

[4]The Biodesign Institute Center for Structural Discovery, Arizona State University, Tempe, AZ 85281, USA

[5]Department of Mathematics and Computer Sciences, Grenoble INP, 38000 Grenoble, France

[6]The Biodesign Institute Virginia G. Piper Center for Personalized Diagnostics, Arizona State University, Tempe, AZ 85281, USA

[7]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, United States

[8]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

[9]Chemistry Department, Quantum Theory Project, University of Florida, Gainesville, Florida, 32611, USA

Abstract

Cryo-electron microscopy (EM) requires molecular modeling to refine structural details from data. Ensemble models arrive at low free-energy molecular structures, but are computationally expensive and limited to resolving only small proteins that cannot be resolved by cryo-EM. Here, we introduce CryoFold - a pipeline of molecular dynamics simulations that determines ensembles of protein structures directly from sequence by integrating density data of varying sparsity at 3–5 Å resolution with coarse-grained topological knowledge of the protein folds. We present six examples showing its broad applicability for folding proteins between 72 to 2000 residues, including large membrane and multi-domain systems, and results from two EMDB competitions. Driven by data from a single state, CryoFold discovers ensembles of common low-energy models together with rare low-probability structures that capture the equilibrium distribution of proteins constrained by the density maps. Many of these conformations, unseen by traditional methods, are experimentally validated and functionally relevant. We arrive at a set of best practices for data-guided protein folding that are controlled using a Python GUI.

1 Introduction

Cryo-electron microscopy (cryo-EM) is a powerful tool for determining the structures of biomolecules. It serves a niche – such as large complexes or membrane proteins or molecules that are not easily crystallizable – that traditional methods, such as X-ray diffraction, electron or neutron scattering, or NMR often cannot handle. Routine cryo-EM structure determination has a number of components: the experiment produces raw data in the form of single-particle images, correction and processing of this data recovers an electrostatic potential map (henceforth referred to as a density map), and finally molecular modeling is required to determine structures from the map. Currently, there are two broad classes of methods for molecular modeling. First, established algorithms for refining X-ray structures, such as Phenix, Coot, or REFMAC are often used, even for ensemble determination [1]. They offer complete models with data of 2 Å or better resolution [2]. A challenge arises because the Cryo-EM datasets are commonly of lower resolution, reflecting a broad diversity of underlying conformations. Second, *integrative approaches* leverage data from multiple experimental sources [3–5], identifying consensus structures compatible with the different datasets. The challenge here is that cryo-EM data is heterogeneous, meaning that some parts of a protein structure are well-determined by the data while others are more poorly defined [6].

The uneven resolution of the datasets poses a need for extensive conformational sampling of the computational models, and identifying the most biophysically relevant conformations from the poorly resolved regions. The span of this biophysically relevant conformational search space is large and grows non-linearly with system size [7]. Multi-model approaches have been recently conceived to interpret sub-5 Å structural data with atomistic ensembles [8]. These methods focus on sampling the conformational space, constrained by the knowledge of only the experimentally observed states. However, ensembles constructed around a known conformation are not representative of the thermodynamic state of a protein [9]. Their truncated sampling offers an incomplete estimate of the number of structures and the corresponding ensemble of states that a protein can assume during equilibrium. Thus, a substantial portion of the conformational space that contributes to the heterogeneity of the observed data remains unresolved [10].

Here, we describe CryoFold, a multiphysics algorithm that derives equilibrium ensembles of folded protein structures from cryo-EM data. Illustrated in Fig. 1, CryoFold is a combination of three methods: **(1)** MAINMAST [11], MAINchain Model trAcing from Spanning Tree – a method that generates the trace of the connected peptide chain when provided with EM data, **(2)** ReMDFF [12], Resolution exchange Molecular Dynamics Flexible Fitting – a MD method for refining protein conformations from electron-density maps, and **(3)** MELD [13, 14], Modeling Employing Limited Data – a Bayesian engine that can work from insufficient data to accelerate the MD sampling of rare events, such as those needed for protein folding. Starting with density maps of resolution 5.0 Å and better, first, MAINMAST is employed to derive a chain trace of C_α atoms. Then we use this trace as a template to iterate between MELD and ReMDFF. While MELD explores a large conformational space, visiting multiple plausible secondary structures consistent with the MAINMAST template, ReMDFF simulations refine the protein backbone and sidechain conformations to fit to the

density map for each one of the assumed secondary structures (Fig. 2A) [15]. Taken alone, ReMDFF fits models into density features, but fails to explore the variations in secondary structures [12]. MELD addresses this issue by partial folding, unfolding and reformation of secondary structures [13, 14], integrating the coarse physical insights (CPI) available on web-servers [14, 16] and that from MAINMAST's initial trace. For example, based on sequences, CPI includes specific fractions of hydrophobic contacts, β-strand pairing and secondary structures required to minimize protein frustration (see Methods) [13]. Taken together, a hybrid iterative MELD-ReMDFF approach allows the determination of a set of complete all-atom models from sequence information merged with available structural data of varying coarseness, and infer a subset of models that best matches with the target dataset. Also, MELD with only CPI is limited to folding small soluble proteins of up to 100 residues. The data-guidance from ReMDFF allows MELD to fold larger structures, with at least 10-fold more residues, inside CryoFold. For intermediate to low-resolution data (less than 5 Å) wherein C-alpha tracing is unreliable, the MAINMAST step can be avoided. Nonetheless, if successful, the search template derived from backbone tracing almost always accelerates convergence of CryoFold.

The guidance from experimental data allows CryoFold to derive transmembrane systems and asymmetric multi-protein complexes. More importantly, unlike homology models, the free energy description of folded and unfolded populations accessible to MELD enables the clustering of structures into distinct metastable states. Thus, starting with the structural data from a particular protein conformation, CryoFold predicts on one hand, the energetically favorable ensemble of structures that are consistent with the data, while on the other hand, discovers multiple new low-energy protein states distal to the fitted model. Going beyond the determination of a stationary structure or local fluctuations in its vicinity [17, 18], CryoFold offers a collective interpretation of the major equilibrium conformations. This ensemble view of the data comes from MELD, wherein the 3D map-fitted search models from MAINMAST are first translated into a set of hundreds of high-dimensional structural restraints. The molecular ensembles are then generated by exchanging between these sets of restraints. The generalized ensemble methodology allows knowledge from the cryo-EM data to be imposed as an average boundary condition that all the resulting models follow either partially or entirely, rather than as a single holonomic boundary restraint traditionally imposed in real-space refinements. Thus, model populations are determined that either completely or partially satisfy the data. As ReMDFF iteratively improves the consistency between the search model and the density data, the set of MELD restraints derived from the fitted models improve. The ensemble generated with these data-guided restraints reveals simultaneously the most-likely set of refined structures, as well as molecular dynamics underlying the protein's conformational heterogeneity. Any structure from the MELD ensemble can be refined using ReMDFF to derive models consistent with the experimental data. However, to ensure rapid convergence, structures from the MELD ensembles are clustered based on their correlation coefficient (CC) relative to the experimental data. The one with the highest map-model correlation (also validated using EMRinger scores) is refined employing ReMDFF.

The states discovered by CryoFold scan the equilibrium protein ensemble either by visiting diverse manifestations of the same structural data or by visiting new energy basins,

structures from which are verified against orthogonal NMR, X-ray crystallography or cryo-EM datasets. All existing X-ray [1], NMR [19] or Cryo-EM [18] ensemble refinement tools focus on interpretation of a chosen dataset via locally restrained model construction. Extending this paradigm, CryoFold enables the generation of global ensembles, encompassing a considerably higher dimensional conformational space, that we cluster and re-refine against multiple independent datasets. The multitude of structures so resolved enables the generation of "molecular movies" directly from experimental data, wherein structures are seen transitioning between multiple energy states [20].

We report data-guided structural ensembles for six different examples here, for proteins from 72 to 618 residues, extending to heterogeneous multi-protein complexes of up to 2000 residues, and across both soluble and membrane systems. CryoFold overcomes the sampling limitations of traditional MD predictions, producing high-quality structural models: it convergences to solution(s) starting with partially folded models from MELD, and iteratively refining soluble and transmembrane structures with consistently > 90% favored backbone and sidechain statistics, and high EMRinger scores [21]. The results are independent of the initial estimated conformation and consistent with physics and stereochemistry, highlighted through results in 2016 and 2019 EMDB competitions. The hybrid protocol is available through a python-based graphical user interface with a video tutorial and list of best practices.

2 Results

We describe six systems, chosen to highlight the pros and cons of the component methods in the CryoFold pipeline. Three are soluble proteins, with varying degrees of local resolution in the density maps. One is from the 2019 EMDB competition challenge, in which data on the same protein was provided at three different resolutions [22]. These examples bring to light how MELD-ReMDFF recover correct ensembles when the MAINMAST predictions are challenged, and vice versa. One is a large asymmetric multi-protein complex that allowed us to test how big a structure CryoFold could handle. And, one is a transmembrane system, to see if MELD's implicit-solvent model would be adequate for ensemble determination in the membrane environment. At any given resolution, the accuracy of CryoFold ensemble predictions depends on: **(1)** quality of C_α traces by MAINMAST, **(2)** variations in secondary structure within the MELD ensemble, and **(3)** convergence of ReMDFF. A set of best practices required for controlling these dependencies, and regulating the size of the data-guided ensembles is outlined in the Methods.

Proof of principle on a small known protein

In this case, we began with a synthetic map of ubiquitin, a small 72-residue protein. Ubiquitin is a good test system because, on the one hand, it is small enough to fold computationally, and yet on the other hand its experimental folding time is in the millisecond range, so it is hard to fold by brute force MD simulation [23], and even, to a lesser extent, by the MELD approach [14]. From the known X-ray crystal structure of ubiquitin, we generated a synthetic density map at 3.0 Å resolution, and asked if CryoFold could correctly recover the X-ray structure.

We found that only two MELD-ReMDFF iterations were needed to give a model having a root mean square difference or RMSD of 2.53 Å from the crystal structure (PDB id: 1UBQ, see Table S1). Starting from a random coil, the overall ubiquitin topology was already recovered in the first iteration (Fig. 2B). The ensembles are reminiscent of the *F'* and *F1* intermediates of ubiquitin folding [23]. Secondary structure refinement of the small fourth β-strand and 3_{10} helical loop yielded a fully folded state after the second iteration. Notably, MELD alone was unable to recover the helical loop (seen in MELD Step 1) despite sampling key folding intermediates seen in 2D-NMR [24]. The synthetic density data reinforced such key secondary structural information during ReMDFF. This yielded more accurate CPI or coarse physical information for the next iteration of MELD, and subsequently secondary structure restraints for the next round of ReMDFF. Altogether, the proof of principles example demonstrates that the new data-driven pipeline is capable of attaining multiple equilibrium states that the too narrow ensembles in ReMDFF [12] or the too extensive ensembles in MELD [13] cannot individually achieve.

Test on a soluble lipoprotein with a uniformly high-resolution data

Francisella lipoprotein Flpp3 is a 108 amino acids long membrane-interacting protein that serves as a target for drug development against tularemia[25]. In this case, we had two datasets: one at high resolution (1.8 Å) from our Serial Femtosecond X-ray (SFX) crystallography experiments of Flpp3, PDB: 6PNY (See Supplementary Information and [26]), and another truncated at low resolution (5.0 Å). The point of this test was to see if we could use the low-resolution data to achieve the high-resolution structure. For both sets, we used MAINMAST [11] to introduce the $C_{\pm}$ traces as constraints for MELD (Fig. 3A,B). Convergent ensembles derived from this MAINMAST-guided MELD step were then refined by ReMDFF to improve the sidechains until the density was resolved with models of reliable geometry. MAINMAST's spanning tree algorithm alone cannot offer any reliable sidechain geometry, it just places the $C_{\pm}$ atoms in the density. We found that one iteration of the MELD-ReMDFF cycle following MAINMAST sufficed to resolve an all-atom model of Flpp3 from the SFX density, with accurate secondary and tertiary structure assignments, and sidechain packing (structural statistics summarized in Table S2). At 5 Å resolution MAINMAST produced low quality backbone traces (Fig. 3B). Remarkably, even these low quality C_{α} traces were enough for MELD-ReMDFF to successfully produce models comparable to our high-resolution refinements. After two MELD-ReMDFF iterations, the best structure obtained was within 2.29 Å RMSD from the SFX model. The overall ensemble from the second MELD-ReMDFF iteration (inset in graph of Fig. 3B) also samples a narrower range of RMSD and global correlation coefficient (CC) values showing the convergence towards a set of conformations in good agreement with the cryo-EM data. However, for the same RMSD relative to the known target model, we find structures covering a broad range of CCs within the MELD ensembles. This breadth of the CC values corresponding to the models in the lowest RMSD window, which is reproducible across all the following examples, confirms that while CryoFold focuses on the possible best fit, the collection of data-guided structures concomitantly accounts for uncertainty about the best fit.

The MELD-only predictions modelled the β-sheets accurately; however, they failed to accurately converge on all helices (Supplementary Fig. S1). For example, a 4-turn helix

was underestimated to contain only 2-3 turns. Similar to the ubiquitin example, but now using empirical X-ray maps, guidance by the density in CryoFold recovered these turns in both the high and low resolution cases (Tables S2 and S3). Thus, the Flpp3 test further demonstrates that the CryoFold trio of methods gives accurate structures for longer chains than is otherwise possible with any one of these methods.

Here, we are also able to test an important aspect of physics-based structure determination, namely whether we can generate meaningful conformational ensembles, not just single average structures. The quality of the CryoFold ensembles is assessed against a set of 20 NMR models of Flpp3 [25] by looking at the conformation of key residues (Y83,K35 and D4) responsible for binding tularemia drugs (Fig. S2A) [26]. Upon projecting the ensemble of 50 lowest-energy CryoFold structures onto a space defined by the distance between Y83-K35 & Y83-D4, where *closed Flpp3* is represented by (Y83-K35 <5.0 Å & Y83-D4 > 10.0 Å), and *open Flpp3* implies (Y83-K35 >10.0 Å & Y83-D4 < 5.0 Å) [26], all the major conformational states seen in the NMR experiments have been recovered (Fig. S2B). Thus, extending beyond the prediction of a single stationary structure, the cluster of low-energy conformations predicted by CryoFold captures both the open and closed conformations, starting only with the 6PNY data from the closed state.

The classification of structural ensembles based on projections onto the distance space requires *a priori* knowledge of the structural features of all the major states in the ensemble. In an alternate scheme that does not require such knowledge, the models were classified based on their Rosetta-energy and RMSD relative to the crystal structure [27]. Rosetta is chosen as a benchmark due to its use of energy functions analogous to the CHARMM or AMBER force fields in MDFF/ReMDFF and MELD. In this energy space, the ensemble of 50 Flpp3 structures derived from CryoFold at 1.8 Å resolution recovered only a minimum number of the states observed in NMR (Fig. S3). This limitation is explained by our recent studies showing that very high-resolution data of 1-3 Å poses stiff data constraints that make it entropically unfavorable for an MD simulation to overcome and explore states that are not strictly defined by the data [28]. Consequently, the sampling becomes highly localized to only one state and the ensemble is overpopulated with similar structures, cutting down on conformational diversity. Rosetta-EM visited almost all the NMR states, potentially benefiting from its Monte Carlo sampling scheme, but still using a 20-fold larger ensemble size than used in CryoFold. In contrast, for the 5.0 Å regime, CryoFold produces a markedly better performance with the 50-model ensemble overlapping with the majority of NMR intermediates and the final SFX solution, as well as consistently determining structures with lower energy than Rosetta-EM. Therefore, the extended sampling benefits of CryoFold are more apparent in fuzzier datasets. Here, a broader segment of the protein folding funnel is accessed by MELD, recovering models even from the poor initial guesses generated by MAINMAST(Fig. S4).

Taken together, the ubiquitin and Flpp3 examples establish CryoFold as an enhanced sampling tool for resolving multiple metastable states of proteins with > 100 residues, guided only by a single experimental dataset at 3-5 Å. Instead of individually determining multi-model interpretations for the 1 SFX and 20 NMR datasets, CryoFold allowed the generation of all the druggable *vs.* non-druggable models as part of a single ensemble

driven by just one piece of information. In the absence of the NMR knowledge, even a multi-model refinement of the high-resolution SFX data would have produced a narrow ensemble, artifactually suggesting that this protein is rigid (Figs. S2 and S3). The multi-state equilibrium ensemble generation in CryoFold removes such assumptions and brings to light the dynamic nature of this protein in addition to resolving the experimental structures.

Test on soluble domains of a membrane protein with heterogeneous-resolution data

We look at the cytoplasmic domain of a large trans-membrane protein, TRPV1, a heat-sensing ion channel (592 amino acids long). The point of this test is that the data is highly heterogeneous, with experimental density maps ranging between 3.8 to 5.0 Å [29], as determined by Resmap. Furthermore, TRPV1 has two apo-structures deposited in the RCSB database, one with moderately resolved transmembrane helices and cytoplasmic domains (pdb id: 3J5P, EMDataBank: EMD-5778), and another with highly-resolved transmembrane helices (pdb id:5IRZ, EMDataBank: EMD-8118) but with the cytoplasmic regions, particularly the β-strands, less locally resolved than in 3J5P. CryoFold was employed to regenerate these unresolved segments of the cytoplasmic domain from the heterogeneous lower-resolution data of 5IRZ. We compare the CryoFold model to the reported 3J5P structure (Fig. 4), where these domains are much better resolved showing clear patterns of β-sheets. The final model was observed to be at an RMSD of 3.41 Å with a CC of 0.74 relative to 5IRZ. The same model with some loops removed for consistency with the EMD-5778 density produced an RMSD of 2.49 Å and CC of 0.73 with respect to the reported 3J5P model. Taken together, models derived from the CryoFold refinement of 5IRZ capture in atomistic details the highly resolved features of this density, yet without compromising with the mid-resolution cytoplasmic areas where it performs as well as the 3J5P model (Table S4).

TRPV1 was part of the 2016 Cryo-EM modeling challenge where only ReMDFF was used [30]. Presented in Table S5, our updated CryoFold model of TRPV1 (model no. 4), represents the top - 20% of the submissions with > 90% Ramachandran favored statistics, and an EMRinger score of 2.54. This model is now refined over the originally reported structure with a score of 1.75, and our previous submission at 2.25. Comparisons with the results from other methods is provided in the supplement, and summarized in the Discussions. The improvement is attributed solely to the higher-quality β-sheet models that is now derived from the enhanced sampling obtained by running MELD and ReMDFF in tandem. Starting with a random coil as search model (Fig. 4B), the recovery of these β-sheets is highly improbable with the limited conformational space that MDFF visits. In fact, MAINMAST and MDFF combined also could not resolve the cytoplasmic region of TRPV1 [31]. Addressing this issue, MELD invokes a multi-replica temperature exchange scheme, wherein at high replica indices it samples many distinct structures that have short lifetimes [13]. At the lower-temperature replica a stronger coupling with the data is achieved, and these structures are folded into a smaller number of long-lived clusters, each with varying degrees of native contacts and secondary structure (Fig.S5). The 5.0 Å local resolution of TRPV1's soluble region is the fuzziest density feature that CryoFold is tasked to resolve. Consequently, refinement of the TRPV1 β-sheets required MELD to sample the broadest structural funnel among all the chosen examples (Fig. S4). It focuses initial models starting

within an RMSD span of 10–25 Å from the target down to refined structures displaying four classes of low-energy topologies (Fig. S5). Thus, unlike MDFF (or ReMDFF), MELD allows for a much broader search of structural motifs hidden within the same density features. In Fig. 4D we see the conformational diversity of the refined ensemble coming from MELD. When these methods are combined inside CryoFold, both the backbone and sidechain geometries are refined against the target 5IRZ density to find the most probable set of conformations (80% of the ensemble population) that capture the TRPV1's labile β-strands.

An analysis of the less probable CryoFold ensembles reveals partial unfolding of the β-strands in the soluble domains of TRPV1 with around 3-4% of the structures presenting incomplete β-sheets, akin to the model originally submitted with 5IRZ (Fig. S5C). Partial unfolding of these regions have not been attributed to any functional implications in TRPV1, though some peripheral evidence of functional advantages from unfolding exist in TRPV3 channels [32]. The β-strands and loops from the soluble domains form the inter-protomer interface within the tertrameric channel. Secondary structural changes at these interfaces, triggers coupling between cytoplasmic and transmembrane domains, priming the channel for opening. Such changes, though rare, are indeed apparent in our MELD assignments. Therefore, the ensemble of structures and not merely the most probable model that CryoFold offers, opens the door to analyzing a number of distinct folded and unfolded conformations, all of which can contribute to the same density map with different weights [22, 33]. Also evident from the TRPV1 case study, we can generate such atomistic ensembles with data of low local resolution, yet with accuracy commensurate to structures derived from higher resolution density maps.

Tests on apoferritin at three different resolutions from the 2019 EMDB modeling challenge

The EMDB competition is a community-wide effort to assess the limits of structure prediction using cryo-EM data. Here we were tasked to determine the structure of a 174-residue apoferritin monomer using data at 1.8, 2.3 and 3.1 Å resolution. Following an initial tracing by MAINMAST on the monomeric map, it took two iterations for CryoFold to arrive at the final model for the first two resolutions, and three iterations for the third map. In total 13 teams participated in the 2019 competition that focused primarily on ab-initio structure determination, and all the results are reported on the EMDB website. CryoFold (team 73) models were independently assessed to be of high accuracy (Fig. S6 (scale labeled in green)), specifically for three different categories of scores: Reference-free, EM-map and target-structure scores. The results were robust over the narrow range of resolutions tested, earning us the top rank for multiple entries [22]. Comparability with respect to the target structures is almost always very high, as also reflected in commensurately high Fourier Shell Coefficient (FSC = 0.5) and the correlation coefficient with the experimental map. Another noticeable strength is the strong EMRinger scores of the MD-based refinement, very similar to ReMDFF's performance in the 2016 competition [30]. A relatively new measure to evaluate mainchain geometry and to identify areas of probable secondary structure based on C-Alpha geometry, called CaBLAM [34] also found the CryoFold models to be favorable. One limitation however, is the increased number of Ramachandran outliers observed in the CryoFold and MDFF determined structures, which implicates the assumptions of classical

CHARMM-type force fields[30]. Our recently developed neural network potentials have already been useful to circumvent this issue [35].

Test on a large multi-chain protein complex with mid-resolution data

A grand challenge for cryo-EM is to determine structures of multi-chain complexes. Symmetry is used wherever possible, e.g., in viruses or homo-oligomeric membrane proteins [32]. However, most protein-protein or protein-nucleic acid complexes are asymmetric. Our test here is whether CryoFold could obtain the structure in an asymmetric complex. We focused on ATP synthase. Recently Murphy et al. reported 30 distinct conformations of this motor at 2.7-4.5 Å resolution [36]. A majority of these structures contain rotating conformations of the so called transmembrane *c*-ring. For simplicity, we have removed this *c*-ring (the transmembrane problem will be addressed in the next section) and chose to model specific ATP synthase conformations that do not contribute to the rotation of the ring. Therefore, we started refining PDB ID: 6RET that contains 31 chains resolved at 4.3 Å, which is one of the lower resolution densities wherein the *c*-ring is in a non-rotatable conformation.

Similar to the Flpp3 and TRPV1 cases, here the ensemble computed by CryoFold correctly captured the low-lying states of the multi-chain system in addition to the target 6RET conformation. For example, seven of the thirty models reported by Murphy et al. which include overall deformations of the 2000-residue system without rotation of the *c*-ring were represented well in the CryoFold ensemble. Using RMSD matrices (Fig. S7A), these structures were clustered in 4 distinct states (States I: 6RET; II: 6RDQ, 6RDR; III: 6RDW, 6RDX; and IV: 6RDK, 6RDL). Remarkably, all these four states are identifiable in an RMSD matrix of 2200 MELD structures within CryoFold (Fig. 5B). States II, III and IV from MELD are initially at backbone RMSD 7.6, 12.0 and 8.4 Å from 6RET, respectively (Fig: S7B). After ReMDFF refinements, structures are consistent with the experimental models from Murphy et al. for states II, III and IV which were refined to RMSD values of 2.1, 2.8, and 1.8 Å, respectively (Fig. 5C, S7C and S8C). Beyond sampling the rare secondary structural changes, seen in the first few examples, here MELD visits states separated by variations in tertiary structures of the protein-protein interfaces (Fig. S9). A simple multi-model ensemble from ReMDFF of the individual density maps completely misses the existence of the other states. Therefore, starting with an ensemble of structures generated to resolve 6RET, the inter-state exchange promoted by MELD's enhanced sampling of the interface contacts [37], allowed ReMDFF to resolve three more conformations of ATP synthase consistent with 6RDQ, 6RDW and 6RDK (Tables: S6 and S7).

A key biophysical outcome that we make from the CryoFold ensembles of ATP synthase is the flexibility of this motor's peripheral stalk domains. Specifically, the OSCP hinge (chain P) assumes a number of distinct open and closed conformations with an RMSD of 3.3-6.4 Å (Fig. 5D) relative to the hinge from 6RET. The elastic coupling in ATP synthase has remained a topic of contention in the bioenergy community with crystallographers claiming minimum flexibility of the stalk regions [38], in sharp contrast to single-molecule observations of "power-strokes" that originate from deformations of the stalk [39]. Within the CryoFold

ensembles incorporating all the states I-IV, we see that the central stalk is in fact less flexible than the peripheral stalk with an RMSD ranging between 2.4-3.8 Å relative to 6RET. So, our results show that most of the elastic coupling in *polytomella* ATP synthase comes from the flexibility of the peripheral stalk, rather than the central stalk. Going beyond the knowledge derived from stationary models, our resolution of structural ensembles exchanging between four low-energy states clearly suggests stalk deformability, and adds credence to the power-stoke mechanism of ATP turnover.

Tests on soluble and membrane domains of a large ion channel with mid-resolution data

A second major challenge in structural ensemble determination arises from the modeling of complete transmembrane protein systems, including structure of both the soluble and TM domains. The refinement becomes particularly daunting for CryoFold, as MELD simulations fail to capture structural changes from explicit protein-membrane interactions [13]. Consequently, the accuracy of the model will depend on the structural information available from the map, and less on the fidelity of the physical interactions that underscore MELD.

Addressing this challenge, CryoFold was employed to model a monomer from the pentameric Magnesium channel CorA, containing 349 residues, at 3.80 Å resolution [40] (pdb id: 3JCF, EMDataBank: EMD-6551) (Figs. 6 and S10). An initial topological prediction of the channel was obtained by flexibly fitting of a linear polypeptide onto the Cα trace obtained from the cryo-EM density using MAINMAST. These traces were already within 6.0 Å of the target C$_\alpha$ conformation in 3JCF, providing high-confidence coarse-grained information for MELD to operate. Leveraging the MAINMAST trace, MELD was used to perform local conformational sampling, regenerating most of the secondary structures. Such local refinement requires a narrow sampling of the folding funnel (Fig. S4). The model with the highest correlation coefficient to the map was then refined using ReMDFF, resulting in models which were at 2.90 Å RMSD to the native state. Even though this model possessed high secondary structure content of 76%, substantial unstructured regions remained both in the cytoplasmic and the transmembrane regions, warranting a further round of refinement. In the subsequent MELD-ReMDFF iteration, the resulting models were re-refined to 2.60 Å RMSD from the native state and final CC of 0.84 with the map. The CryoFold models were also comparable in geometry to that deposited in the database (Fig. 6 and Table S8).

We find that starting with high-quality chain traces, CryoFold ensembles can indeed be guided to model helical membrane segments even in an implicit solvent environment. The β-sheet rich soluble domains are concomitantly refined from lower resolution features of the same map. Seen in Fig. 6B, the uncertainty in the ensemble is broader in the soluble domains, which, similar to TRPV1 are verified to be more flexible and engage in Magnesium-mediated pore opening [40]. Convergence of the refined MELD ensemble is further indicated by tighter clusters of structures with systematic improvement in CCs and RMSD values relative to the target Fig. 6C across three rounds of ReMDFF MELD iterations. CryoFold therefore overcomes MELD's traditional weaknesses, and going beyond the limited convergence radius and over-fitting artifacts of flexible fitting methodologies [41], establishes MD simulation as a data-guided ensemble determination tool for transmembrane proteins and their complexes.

3 Discussion

The systems presented here have been chosen as challenging problems to the methods that constitute CryoFold. We have not over-optimized any aspect of the protocol to fit one problem, rather complemented the uncertainties and weakness of one method with the strengths of another. This approach is akin to the consensus methods that are known to improve performance over single methods in blind prediction challenges [42]. In light of the current results, it is expected that a selected combination of methods within CryoFold's plug-and-play protocol will enable the resolution of novel protein folds (Fig. S11) from density data, where the individual methods will potentially fail.

The probability of a structure contributing to one or more subsets of data, or the converse, is determined by their energies derived from the all-atom force field (GBneck2 and ff14SB). Several subsets of data can contribute to the same metastable state or different – and some might be incompatible with the force field leading to very low populations. Therefore, the ratio of refined structures populating multiple metastable states maintains the same ratio of Boltzmann weights between these states as in the unbiased force field, while still agreeing with one or more subset of data. This unique facet of MELD allows the determination of thermodynamic averages, such as relative binding free energies[37]. Within CryoFold, since thermodynamic averages are not our focus, we have focused only on the construction of the data-guided ensembles and, using ReMDFF, extraction of the best possible single-structure representation of the data.

How much data we enforce is set as a prior (explained at lengths in the SI). If the uncertainty prior is set too low, MELD sampling is compromised and we cannot identify structures consistent with the data. If it is set too high, the lower replicas will increase in restraint energies, creating difficulty in the identification of the biologically relevant metastable states and deforming their conformations. Iteration between ReMDFF and MELD produces new sets of contact maps that gives rise to better priors and faster convergence to the relevant metastable states. Thus, during these iterations the coarse physical information or CPI derived from the experimental data is used as an average quantity, arising from different subsets of the contact data and affecting the refined models with varying degrees of uncertainty, rather than all the contacts being enforced on a single model. Only in the final iteration, the agreement with the experimental data is enforced. Here, we used ReMDFF to find a single model that is best fitted into the density map. Cryofold offers therefore, both the single best data-guided structure as well as an associated ensemble, where all the data is not satisfied by a single structure.

While CryoFold appears promising for obtaining biomolecular structures from cryo-EM, we are aware of some limitations. First, its success depends upon the correctness of the initial trace generated by MAINMAST. It is not clear when and whether the MD tools can recover from a wrong chain trace, particularly for resolving the transmembrane systems. We do not expect that sequence assignment in MAINMAST model is perfect. Therefore, we use the MELD-MDFF protocol for refinement. When an initial map has low resolution resulting in a lower quality MAINMAST trace, we de-weight the accompanying contact information. For this reason, we chose to perform a refinement of Flpp3 protein at a 5 Å resolution,

wherein indeed, the MAINMAST alignment was incorrect. Here, the MAINMAST contacts were employed with softer restraints inside MELD relative to the 1.8 Å case. Protein folding (primarily driven by force fields) via the replica-exchange sampling inside MELD recovered the all-atom structure from the initial misalignments to one that is commensurate to the higher-resolution model. If the protein segments are small (within 115 residues) misfolding errors coming from incorrect sequence alignments can be corrected by the force fields, ensuring local refinements. But if such errors become global, physics simulations will find it difficult to handle the problem. Thus, unlike Flpp3, repeating the CorA refinement with a misaligned MAINMAST trace resulted in unreliable models. Deep learning tools such as Deep-Tracer offer a tangible alternative to the MAINMAST traces for providing templates for ensemble generation at sub-4 Å resolution. Second, we do not have a good implicit membrane model to use in the MELD simulations and the use of explicit solvent would require many replicas, seeking more resources than currently available. Thus, by relying solely on the information coming from the density map we impose positional restraints and focus sampling on the transmembrane domains. Third, as with any MD simulation of biomolecules, the force fields are still not perfect and larger structures will be a challenge for the searching and sampling, even with an accelerator such as MELD. Finally, in our current approach, MELD is the most computationally limiting, requiring between one and ten days of sampling with 30 GPUs for the systems studied. These resources might be prohibitive for single lab resources but accessible through supercomputing resources available to academic researchers. Future research will aim at reducing the computational expense required for CryoFold. The computing need will be particularly pressing in multi-chain systems where map segmentation becomes an additional issue that we have not addressed during ensemble generation. Fortunately, MAINMAST has a recently developed multi-segment version which naturally lends to our pipeline [43], and MD simulations have been historically successful in modeling multi-subunit systems [9]. Thus, scaling CryoFold with segmented-MAINMAST offers a viable way forward for data-guided ensemble generation of large protein complexes.

Despite the aforementioned limitations, CryoFold has been compared to popular structure determination protocols. Barring the Flpp3 case at 1.8 Å, CryoFold was always found to offer higher quality models, but more importantly a diverse range of structures consistent with the expected biophysics. While for TRPV1 and CorA, other available multi-model protocols converged to structures with unphysical overlap between the β-strands (Fig. S5 and S12), a multi-protein refinement for ATP synthase could not be reproduced using standard resources, though individual chain refinements were achieved and are reported in Fig. S14. A key benefit of this work, justifying the need for intense computations, is the ability to capture ensembles rather than single structures. Consequently, we identify conformations that are close to the native structure, but also some alternative meta-stable states that are favored by the combination of force field and data. An important question follows – are these structures really relevant or just spurious? To this end, we have validated using NMR and cryo-EM experiments that in addition to the narrow set of models consistent with one density map, there exists orthogonal states that are observed both in the experiments in CryoFold refinements. These orthogonal structures sampled by MELD are indeed leveraged in biological functions, as found in the open→close transition in Flpp3, secondary structure-induced pore opening in the TRPV channels or flexibility of

the peripheral stalks in elastic coupling of the ATP synthase example, and yet behooves resolution by the limited sampling capacity of brute-force MD or Monte Carlo sampling used in stationary structure determination. Also, deep learning tools (e.g. AlphaFold) though have championed single protein structure prediction, their role in the prediction of ensemble dynamics for multi-domain systems is yet to be determined, keeping the physics approaches still the first choice.

Finally, evident from the 2016 and 2019 EMDB competition results, heterogeneous map resolutions affect the completeness of all the ensuing models. While a significant number of modelers prefer to truncate the more dynamic regions, MDFF offers a way to quantify uncertainty of the dynamic regions with root mean square deviations from an average model [12], and to correlate the inherent flexibility of complete protein models with the local resolution of density maps. Now, inside CryoFold, the flexible regions are even more thoroughly sampled by MELD offering the possibility of seeking hidden states in these fuzzy regions. Altogether, we present the first MD based methodology for data-guided protein folding and ensemble refinement, bridging the strengths from two distinct areas of Biophysics. The implementation is semi-automated, and manual fitting is completely avoided. However, the user will require to control the Input/Output between the three methods, and optimize the default parameters as required. Detailed in the Methods and in the SI, we have provided a GUI to facilitate this stage.

4 Conclusions

Structures, dynamics and function are interlinked. We often concentrate on a set of tools to determine structures from data and then use alternate computational techniques to determine dynamics between these metastable structures to ultimately elucidate biological functions. By leveraging the parallel algorithms with techniques such as CryoEM that capture multiple states (but an unknown number of them) computations can go beyond single structures to establish molecular dynamics directly from data. CryoFold is a first step in that direction.

5 Materials and Methods

The data-guided fold and fitting paradigm presented herein combines three real-space refinement methodologies, namely MELD, MAINMAST and ReMDFF. In what follows, these three formulations are articulated individually and the readers are referred to the original publications for details. Then, we outline the hybridization of the methods to provide a molecular dynamics-based *de novo* structure determination tool, CryoFold. Details of the setup for each individual system, as well as, an outline of the computational resources required is outlined in Supplementary Information to showcase the different contexts in which CryoFold can operate (see Table S9).

MELD

Modeling Employing Limited Data (MELD) employs a Bayesian inference approach (eq. (1)) to incorporate empirical data into MD simulations[13, 14]. The bayesian prior $p(\vec{x})$ comes from an atomistic force field (ff14SB sidechain, ff99SB backbone) and an implicit solvent model (Generalized born with neck correction, gb-neck2) [44, 45]. The likelihood $p(\vec{D} \mid \vec{x})$,

representing a bias towards known information, determines how well do the sampled conformations agree with known data, D. $p(\vec{D})$ refers to the likelihood of the data, which we take as a normalization term that can typically be ignored. Taken together,

$$\overbrace{p(\vec{x} \mid \vec{D})}^{\text{posterior}} = \frac{p(\vec{D} \mid \vec{x})p(\vec{x})}{p(\vec{D})} \sim \overbrace{p(\vec{D} \mid \vec{x})}^{\text{likelihood}}\overbrace{p(\vec{x})}^{\text{prior}}. \quad (1)$$

MELD is designed to handle data with one or more of these features: sparsity, noise and ambiguity. Brute-force use of such data leads to incorrect models[46] as not all the data is compatible with the native state. MELD addresses the refinement of low-resolution data by enforcing only a fraction (x%) of this data at every step of the MD simulation. Although x is kept fixed, the subset of data chosen to bias the simulation keeps changing with the simulation steps in a deterministic way. For a given structure all the data is evaluated, sorted according to their energy penalty and the x% with lowest energy guide the simulation until the next step. The data is incorporated as flat-bottom harmonic restraints $E(r_{ij})$ for evaluating the likelihood ($p(\vec{D} \mid \vec{x})$).

$$E(r_{ij}) = \begin{cases} \frac{1}{2}k(r_1 - r_2)(2r_{ij} - r_1 - r_2) & \text{if } r_{ij} < r_1 \\ \frac{1}{2}k(r_{ij} - r_2)^2 & \text{if } r_1 \le r_{ij} < r_2 \\ 0 & \text{if } r_2 \le r_{ij} < r_3 \\ \frac{1}{2}k(r_{ij} - r_3)^2 & \text{if } r_3 \le r_{ij} < r_4 \\ \frac{1}{2}k(r_4 - r_3)(2r_{ij} - r_4 - r_3) & \text{if } r_4 \le r_{ij}, \end{cases} \quad (2)$$

When these restraints are satisfied they do not contribute to the energy or forces, contributing for flat bottom region of eq. 2 and (Fig. S13). When the restraints are not satisfied they add energy penalties and force biases to the system – guiding it to regions that satisfy a subset of the data, or conformational envelopes. MELD is available as a plugin on the MD simulation platform OpenMM. Details of MELD implementation and ensemble generation are provided in Supplementary methods: Description of MELD and Data-guided ensemble generation.

MAINMAST

MAINchain Model trAcing from Spanning Tree (MAINMAST) is a *de novo* modeling program that directly builds protein main-chain structures from an EM map of around 4-5 Å or better resolutions[11]. MAINMAST automatically recognized main-chain positions in a map as dense regions and does not use any known structures or structural fragments. The procedure of MAINMAST consists of mainly four steps (Fig. S15). In the first step, MAINMAST identifies local dense points (LDPs) in an EM map by mean shifting algorithm. All grid points in the map are iteratively shifted by a gaussian kernel function and then merged to the clusters. The representative points in the clusters are called LDPs.

In the second step, all the LDPs are connected by constructing a minimum spanning tree (MST). It is found that the most edges in the MST covers the main-chain of the protein structure in EM map[11]. In the third step, the initial tree structure (MST) is refined iteratively by the so-called tabu search algorithm. This algorithm attempts to explore a large search space by using a list of moves that are recently considered and then forbidden. In the final step, the longest path of the refined tree is aligned with the amino acid sequence of the target protein. This process assigns optimal Cα positions of the target protein on the path and evaluates the fit of the amino acid sequence to the longest path in a tree. MAINMAST is now available as a plugin on Chimera. Details of MAINMAST implementation are provided in Supplementary methods: Description of MAINMAST.

Traditional MDFF

The protocol for molecular dynamics flexible fitting (MDFF) has been described in detail[15]. Briefly, a potential map V_{EM} is generated from the cryo-EM density map, given by

$$V_{\mathrm{EM}}(\mathbf{r}) = \begin{cases} \zeta\left(1 - \dfrac{\Phi(\mathbf{r}) - \Phi_{\mathrm{thr}}}{\Phi_{\max} - \Phi_{\mathrm{thr}}}\right) & \text{if } \Phi(\mathbf{r}) \geq \Phi_{\mathrm{thr}}, \\ \zeta & \text{if } \Phi(\mathbf{r}) < \Phi_{\mathrm{thr}}. \end{cases} \tag{3}$$

where $\Phi(\mathbf{r})$ is the biasing potential of the EM map at a point $\mathbf{r}$, 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 $\Phi_{\max} = \max(\Phi(\mathbf{r}))$.

A search model is refined employing MD, where the traditional potential energy surface is modified by V_{EM}. The density-weighted MD potential conforms the model to the EM map, while simultaneously following constraints from the traditional force fields.

ReMDFF

While traditional MDFF works well with low-resolution density maps, recent high-resolution EM maps have proven to be more challenging. This is because high-resolution maps run the risk of trapping the search model in a local minimum of the density features. To overcome this unphysical entrapment, resolution exchange MDFF (ReMDFF) employs a series of MD simulations. Starting with $i = 1$, the ith map in the series is obtained by applying a Gaussian blur of width σ_i to the original density map. Each successive map in the sequence $i = 1, 2, \ldots L$ has a lower σ_i (higher resolution), where L is the total number of maps in the series ($\sigma_L = 0$ Å). The fitting protocol assumes a replica-exchange approach described in details[12] and illustrated in Fig. S16. At regular simulation intervals, 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 $p(\mathbf{x}_i, \sigma_i, \mathbf{x}_j, \sigma_j) =$

$$\min\left(1, \exp\left(\frac{-V(\mathbf{x}_i, \sigma_j) - V(\mathbf{x}_j, \sigma_i) + V(\mathbf{x}_i, \sigma_i) + V(\mathbf{x}_j, \sigma_j)}{k_B T}\right)\right) \tag{4}$$

where k_B is the Boltzmann constant, $V(\mathbf{x}, \sigma)$ is the instantaneous total energy of the configuration $\mathbf{x}$ within a fitting potential map of blur width σ. Thus, ReMDFF fits the search model to an initially large conformational space that is shrinking over the course

of the simulation towards the highly corrugated space described by the original MDFF potential map. Both MDFF and ReMDFF are available as plugins on VMD. Details of ReMDFF implementation are provided in Supplementary methods: Description of Resolution exchange MDFF.

CryoFold (MELD-MAINMAST-ReMDFF) protocol and best practices

Illustrated in Fig. 1, the CryoFold protocol begins with MELD computations, which guided by backbone traces from MAINMAST yields folded models. These models are flexibly fitted into the EM density by ReMDFF to generate refined atomistic structures.

1. First, information for the construction of Bayesian likelihood is derived from secondary structure predictions (PSIPRED), which were enforced with a 70% confidence. This percentage of confidence offers an optimal condition for MELD to recover from the uncertainties in secondary structure predictions[16]. For membrane proteins, this number can be increased to 80% when the transmembrane motifs are well-defined helices. MELD extracts additional prior information from the MD force field and the implicit solvent model (see eq.1).

2. In the second step, any region determined with high accuracy will be kept in place with cartesian restraints imposed on the Cα during the MELD simulations. This way, the already resolved residues can fluctuate about their initial position.

3. In the third step, contact restraints (e.g. distance between the Cα traces of MAINMAST) are derived. The threshold value of density chosen for MAINMAST chain-tracing is 0.5-1.0. A second important MAINMAST parameter is the number of iterations. If the chain length >115 residues, it requires between 1000-5000 iterations to converge. For smaller protein segments (<100 residues), up to 500 iterations suffice.

   The application of MAINMAST allows construction of pairwise interactions as MELD-restraints directly from the EM density features. Together with the cartesian restraints of step 2, these MAINMAST-guided distance restraints are enforced via flat-bottom harmonic potentials (see eq. 2) to guide the sampling of a search model; notably, the search model is either a random coil or manifests some topological features when created by fitting the coil to the Cα trace with targeted MD. Depending upon the stage of CryoFold refinements, only a percent of the cartesian and distance restraints need be satisfied. The cartesian restraints are often localized on the structured regions, while the distance restraints typically involve regions that are more uncertain (e.g loop residues).

   The two parameters that are relevant in going from MAINMAST models to MELD setup is the threshold Cα-Cα distance to consider and the number of contacts to trust. At lower resolution, we set a higher distance threshold (e.g. 8$\AA$) and reduce the per cent of contacts to trust (e.g (55%). Ultimately, after running MELD simulations, the agreement with the density map and violations on the number of restraints can provide a good estimate of the quality of the initial assumptions. If large number of violations are detected, the percentage of trusted contacts should decrease or the distance threshold increase.

4. Fourth, a Temperature and Hamiltonian replica exchange protocol (H,T-REMD) is employed (using data from steps 1 to 3) to accelerate the sampling of low-energy conformations in MELD[13, 14], refining the secondary-structure content of the model. The Hamiltonian is changed by changing the force constant applied to the restraints. Simulations at higher replica indexes have higher temperatures and lower (vanishing) force constants so sampling is improved. At low replica indices, temperatures are low and the force constants are enforced at their maximum value (but only a certain per cent of the restraints, the ones with lower energy, are enforced). See SI for details for individual applications. Simulations of 30 ns per replica with 15 to 25 replicas are routinely applied to construct the conformational ensembles.

5. Fifth, the correlation coefficient of the H,T-REMD-generated structures with the EM-density is employed as a metric to select the best model for subsequent refinement by ReMDFF (Fig. S16). Resolution exchange across 5 to 11 maps with successively increasing Gaussian blur of 0.5 Å (σ in eq. 4) sufficed to improve the correlation coefficient and structural statistics. The model with the highest EMringer score forms the starting point of the next round of MELD simulations, where now the contact information come from the ReMDFF models. Thereafter, another round ReMDFF is initiated, and this iterative MELD-ReMDFF protocol continues until the δCC between two consecutive iterations is <0.1.

   ReMDFF employs secondary structure (or ssrestraints) to avoid over-fitting of structures into the density maps. In CryoFold, these constraints are employed starting from the second iteration of the MELD-ReMDFF cycle, only after the first MELD step is complete, wherein secondary structure from PSI and MAINMAST data are translated in all-atom structures. The gscales parameter ranges between 0.1–0.3 in earlier MELD-ReMDFF iterations till the topology information in MELD converges. In subsequent iterations when the map resolution is between 3–4.5 Å, the temperature is brought to 80 K and in the final step the gscale is increased to 1.0 to enable sidechain refinements. For maps lower than 5 Å resolution, only backbone fitting is performed.

   As more iteration cycles between MELD and ReMDFF are done, the contact distance threshold and percentage of data to trust increases. At the last stages of refinement Cα-Cα thresholds of 6 Å and percentages as high as 80% are used. The decisions are based on the agreement between ReMDFF models and the CryoEM map.

Throughout different rounds of iterative refinement, the structures from ReMDFF are used as seeds in new MELD simulations. At the same time, distance restraints from the ReMDFF model are updated and the pairs of residues present in those contact interactions are enforced at different accuracy levels. As expected, the more rounds of refinement we do, the higher the accuracy levels for the contacts is achieved in CryoFold. In going through this procedure, the ensembles produced get progressively narrower as we increase the amount of restraints enforced. The discussed parameters can be conveniently set in the GUI. A video tutorial and

the description of this pipeline encompassing Chimera, OpenMM and VMD is provided in Supplementary methods: Graphical User Interface.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

AS and CG acknowledge start-up funds from the SMS and CASD at Arizona State University, CAREER award by NSF-MCB 1942763 and the resources of the OLCF at the Oak Ridge National Laboratory, which is supported by the Office of Science at DOE under Contract No. DE-AC05-00OR22725, made available via the INCITE program. ET laboratory is supported by NIH (P41GM104601); ET, AS and MS acknowledge NIH (R01GM098243-02). This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (Awards OCI-0725070 and ACI-1238993) and the state of Illinois. KD and AP appreciate support from a PRAC computer allocation supported by NSF Award ACI1514873, support from NIH Grant GM125813 and the Laufer Center. AP appreciates start-up support from the University of Florida. DK acknowledges support from the National Institutes of Health (R01GM123055), the National Science Foundation (MCB1925643, DMS1614777, CMMI1825941), and the Purdue Institute of Drug Discovery. WVH acknowledges NIH (R01GM112077).

References

1. Burnley BT, Afonine PV, Adams PD & Gros P Modelling dynamics in protein crystal structures by ensemble refinement. eLife 1, e00311 (2012). URL 10.7554/eLife.00311. [PubMed: 23251785]
2. Terwilliger TC, Adams PD, Afonine PV & Sobolev OV A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps. Nature methods 15, 905 (2018). [PubMed: 30377346]
3. Rout MP & Sali A Principles for Integrative Structural Biology Studies. Cell 177, 1384–1403 (2019). [PubMed: 31150619]
4. Ray W, Kudryashev Y-R, Lia M, Egelmana X, Basler EH, Yifan MC, David B & Frank D De novo protein structure determination from near-atomic-resolution cryo-em maps. Nature Methods 12, 335 (2015). [PubMed: 25707029]
5. Zhou W, Fiorin G, Anselmi C, Karimi-Varzaneh HA, Poblete H, Forrest LR & Faraldo-Gómez JD Large-scale state-dependent membrane remodeling by a transporter protein. eLife 8, e50576 (2019). URL 10.7554/eLife.50576. [PubMed: 31855177]
6. Wang X & Boudker O Large domain movements through the lipid bilayer mediate substrate release and inhibition of glutamate transporters. eLife 9, e58417 (2020). URL 10.7554/eLife.58417. [PubMed: 33155546]
7. Frank J & Ourmazd A Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. Methods 100, 61–67 (2016). URL https://linkinghub.elsevier.com/retrieve/pii/S1046202316300251. [PubMed: 26884261]
8. Fraser JS, Lindorff-Larsen K & Bonomi M What will computational modeling approaches have to say in the era of atomistic cryo-em data? Journal of Chemical Information and Modeling 60, 2410–2412 (2020). URL 10.1021/acs.jcim.0c00123. 10.1021/acs.jcim.0c00123. [PubMed: 32090567]
9. Goh BC, Hadden JA, Bernardi RC, Singharoy A, McGreevy R, Rudack T, Cassidy CK & Schulten K Computational methodologies for real-space structural refinement of large macromolecular complexes. Annu. Rev. Biophys 45, 253–278 (2016). [PubMed: 27145875]
10. Cossio P & Hummer G Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. Journal of Structural Biology 184, 427–437 (2013). URL http://www.sciencedirect.com/science/article/pii/S1047847713002712. [PubMed: 24161733]
11. Terashi G & Kihara D De novo main-chain modeling for EM maps using MAINMAST. Nature Communications 9, 1618 (2018). URL 10.1038/s41467-018-04053-7.

12. Singharoy A, Teo I, McGreevy R, Stone JE, Zhao J & Schulten K Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. eLife 10.7554/eLife.16105 (2016).
13. MacCallum JL, Perez A & Dill KA Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. Proceedings of the National Academy of Sciences of the United States of America 112, 6985–6990 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26038552 https://www.ncbi.nlm.nih.gov/pmc/PMC4460504/. [PubMed: 26038552]
14. Perez A, MacCallum JL & Dill KA Accelerating molecular simulations of proteins using Bayesian inference on weak information. Proceedings of the National Academy of Sciences of the United States of America 112, 11846–11851 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26351667 https://www.ncbi.nlm.nih.gov/pmc/PMC4586851/. [PubMed: 26351667]
15. Trabuco LG, Villa E, Mitra K, Frank J & Schulten K Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. Structure 16, 673–683 (2008). [PubMed: 18462672]
16. Jones SR, Gainetdinov RR, Hu X-T, Cooper DC, Wightman RM, White FJ & Caron MG Loss of autoreceptor functions in mice lacking the dopamine transporter. Nature Neurosci. 2, 649–655 (1999). [PubMed: 10404198]
17. Bonomi M, Camilloni C, Cavalli A & Vendruscolo M Metainference: A Bayesian inference method for heterogeneous systems. Science Advances 2, e1501177–e1501177 (2016). URL http://advances.sciencemag.org/content/2/1/e1501177.full-text.pdf+html. [PubMed: 26844300]
18. Herzik MA, Fraser JS & Lander GC A multi-model approach to assessing local and global cryo-em map quality. Structure 27, 344–358.e3 (2019). URL http://www.sciencedirect.com/science/article/pii/S0969212618303642. [PubMed: 30449687]
19. Lange OF, Lakomek N-A, Farès C, Schröder GF, Walter KFA, Becker S, Meiler J, Grubmüller H, Griesinger C & de Groot BL Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. Science 320, 1471–1475 (2008). URL https://science.sciencemag.org/content/320/5882/1471. https://science.sciencemag.org/content/320/5882/1471.full.pdf. [PubMed: 18556554]
20. Dashti A, Shekhar M, Hail DB, Mashayekhi G, Schwander P, Georges A. d., Frank J, Singharoy A & Ourmazd A Functional pathways of biomolecules retrieved from single-particle snapshots. bioRxiv (2019). URL https://www.biorxiv.org/content/early/2019/01/14/291922. https://www.biorxiv.org/content/early/2019/01/14/291922.full.pdf.
21. Baradaran R, Berrisford JM, Minhas GS & Sazanov LA Crystal structure of the entire respiratory complex I. Nature 494, 443–448 (2013). [PubMed: 23417064]
22. Lawson CL et al. Outcomes of the 2019 emdataresource model challenge: validation of cryo-em models at near-atomic resolution. Nature Methods 18, 156–164 (2021). URL https://www.biorxiv.org/content/early/2020/06/15/2020.06.12.147033. [PubMed: 33542514]
23. Piana S, Lindorff-Larsen K & Shaw DE Atomic-level description of ubiquitin folding. Proceedings of the National Academy of Sciences 110, 5915–5920 (2013).
24. Schanda P, Forge V & Brutscher B Protein folding and unfolding studied at atomic resolution by fast two-dimensional nmr spectroscopy. Proceedings of the National Academy of Sciences 104, 11257–11262 (2007). URL https://www.pnas.org/content/104/27/11257. https://www.pnas.org/content/104/27/11257.full.pdf.
25. Zook J et al. NMR Structure of Francisella tularensis Virulence Determinant Reveals Structural Homology to Bet v1 Allergen Proteins. Structure (London, England : 1993) 23, 1116–1122 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26004443 https://www.ncbi.nlm.nih.gov/pmc/PMC4835214/.
26. Zook J et al. XFEL and NMR Structures of Francisella Lipoprotein Reveal Conformational Space of Drug Target against Tularemia. Structure 28, 540–547.e3 (2020). URL https://linkinghub.elsevier.com/retrieve/pii/S0969212620300460. [PubMed: 32142641]
27. Leelananda SP & Lindert S Using nmr chemical shifts and cryo-em density restraints in iterative rosetta-md protein structure refinement. Journal of Chemical Information and Modeling 60, 2522–2532 (2020). URL 10.1021/acs.jcim.9b00932., 10.1021/acs.jcim.9b00932. [PubMed: 31872764]

28. Vant JW, Sarkar D, Streitwieser E, Fiorin G, Skeel R, Vermaas JV & Singharoy A Data-guided multi-map variables for ensemble refinement of molecular movies. The Journal of Chemical Physics 153, 214102 (2020). URL 10.1063/5.0022433. 10.1063/5.0022433. [PubMed: 33291927]
29. Kucukelbir A, Sigworth FJ & Tagare HD Quantifying the local resolution of cryo-EM density maps. Nat. Methods 11, 63–65 (2014). [PubMed: 24213166]
30. Wang Y, Shekhar M, Thifault D, Williams CJ, McGreevy R, Richardson J, Singharoy A & Tajkhorshid E Constructing atomic structural models into cryo-em densities using molecular dynamics- pros and cons. Journal of Structural Biology 204, 319–328 (2018). URL http://www.sciencedirect.com/science/article/pii/S1047847718301990. [PubMed: 30092279]
31. Terashi G & Kihara D De novo main-chain modeling with mainmast in 2015/2016 em model challenge. Journal of Structural Biology 204, 351–359 (2018). URL http://www.sciencedirect.com/science/article/pii/S1047847718301710. [PubMed: 30075190]
32. Zubcevic L, Hsu AL, Borgnia MJ & Lee S-Y Symmetry transitions during gating of the trpv2 ion channel in lipid membranes. eLife 8, e45779 (2019). [PubMed: 31090543]
33. Abriata LA & Peraro MD Will Cryo-Electron Microscopy Shift the Current Paradigm in Protein Structure Prediction? Journal of chemical information and modeling 60, 2443–2447 (2020). [PubMed: 32134661]
34. Williams CJ et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Science 27, 293–315 (2018). URL http://doi.wiley.com/10.1002/pro.3330. [PubMed: 29067766]
35. Vant JW, Lahey S-LJ, Jana K, Shekhar M, Sarkar D, Munk BH, Kleinekathöfer U, Mittal S, Rowley C & Singharoy A Flexible Fitting of Small Molecules into Electron Microscopy Maps Using Molecular Dynamics Simulations with Neural Network Potentials. Journal of Chemical Information and Modeling 60, 2591–2604 (2020). URL https://pubs.acs.org/doi/10.1021/acs.jcim.9b01167. [PubMed: 32207947]
36. Murphy BJ, Klusch N, Langer J, Mills DJ, Yildiz Ö & Kühlbrandt W Rotary substates of mitochondrial ATP synthase reveal the basis of flexible F 1 -F o coupling. Science 364, eaaw9128 (2019). URL https://www.sciencemag.org/lookup/doi/10.1126/science.aaw9128. [PubMed: 31221832]
37. Morrone JA, Perez A, MacCallum J & Dill KA Computed binding of peptides to proteins with meld-accelerated molecular dynamics. Journal of Chemical Theory and Computation 13, 870–876 (2017). [PubMed: 28042966]
38. Rubinstein JL Structure of the mitochondrial ATP synthase by electron cryomicroscopy. The EMBO Journal 22, 6182–6192 (2003). URL http://emboj.embopress.org/cgi/doi/10.1093/emboj/cdg608. [PubMed: 14633978]
39. Martin JL, Ishmukhametov R, Spetzler D, Hornung T & Frasch WD Elastic coupling power stroke mechanism of the F 1 -ATPase molecular motor. Proceedings of the National Academy of Sciences 115, 5750–5755 (2018). URL http://www.pnas.org/lookup/doi/10.1073/pnas.1803147115.
40. Matthies D et al. Cryo-EM Structures of the Magnesium Channel CorA Reveal Symmetry Breaks upon Gating. Cell 164, 747–756 (2016). URL 10.1016/j.cell.2015.12.055. [PubMed: 26871634]
41. DiMaio F, Song Y, Li X, Brunner MJ, Xu C, Conticello V, Egelman E, Marlovits TC, Cheng Y & Baker D Atomic-accuracy models from 4.5 Å cryo-electron microscopy data with density-guided iterative local refinement. Nat. Methods 12, 361–365 (2015). [PubMed: 25707030]
42. Wilson EA, Hirneise G, Singharoy A & Anderson KS Total predicted mhc-i epitope load is inversely associated with population mortality from sars-cov-2. Cell Reports Medicine 2, 100221 (2021). URL https://www.sciencedirect.com/science/article/pii/S2666379121000379. [PubMed: 33649748]
43. Terashi G, Kagaya Y & Kihara D Mainmastseg: Automated map segmentation method for cryo-em density maps with symmetry. Journal of Chemical Information and Modeling 60, 2634–2643 (2020). URL 10.1021/acs.jcim.9b01110. 10.1021/acs.jcim.9b01110. [PubMed: 32197044]
44. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE & Simmerling C ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. Journal

of Chemical Theory and Computation 11, 3696–3713 (2015). URL 10.1021/acs.jctc.5b00255. [PubMed: 26574453]

45. Nguyen H, Roe DR & Simmerling C Improved Generalized Born Solvent Model Parameters for Protein Simulations. Journal of Chemical Theory and Computation 9, 2020–2034 (2013). URL 10.1021/ct3010485. [PubMed: 25788871]
46. Goh BC, Hadden JA, Bernardi RC, Singharoy A, McGreevy R, Rudack T, Cassidy CK & Schulten K Computational methodologies for real-space structural refinement of large macromolecular complexes. Annual Review of Biophysics 45, 253–278 (2016). URL 10.1146/annurev-biophys-062215-011113. 10.1146/annurev-biophys-062215-011113.
47. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC & Ferrin TE UCSF Chimera - A visualization system for exploratory research and analysis. J. Comp. Chem 25, 1605–1612 (2004). [PubMed: 15264254]

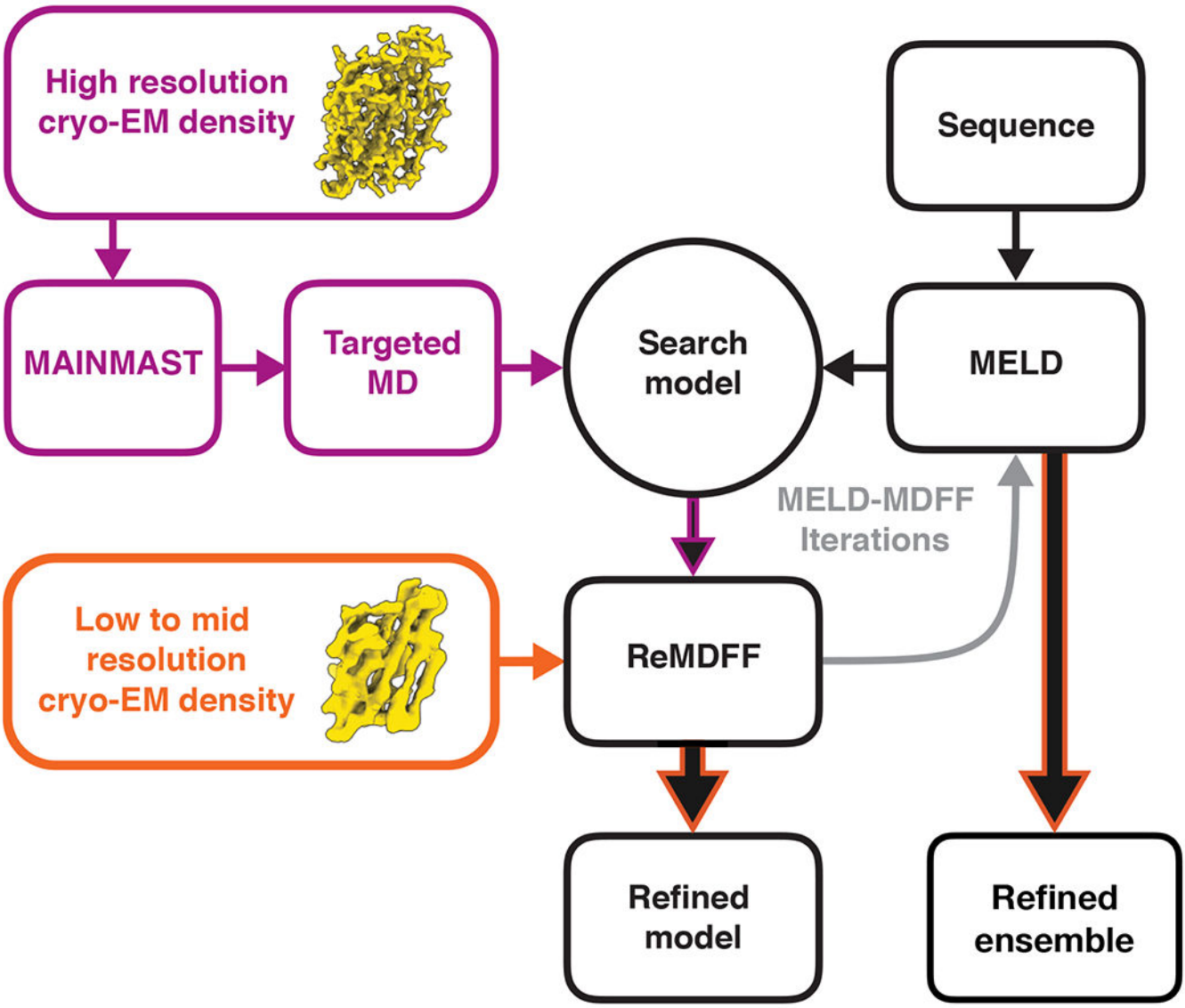

Figure 1: An overview of the CryoFold protocol.

For a high-resolution density map (data-rich case), backbone tracing is performed using MAINMAST to determine Cα positions, and a random coil is fitted to these positions using targeted MD. This fitted protein model is subjected to the next MELD-ReMDFF cycles as a search model. For a low or medium resolution density map (data-poor case), a search model is constructed from primary sequence using MELD. This search model is fitted into the density map using ReMDFF. The ReMDFF output is fed back to MELD for the next iteration, and the cycle continues until convergence. The last iteration of the cycle yeilds a refined model and refined ensemble.

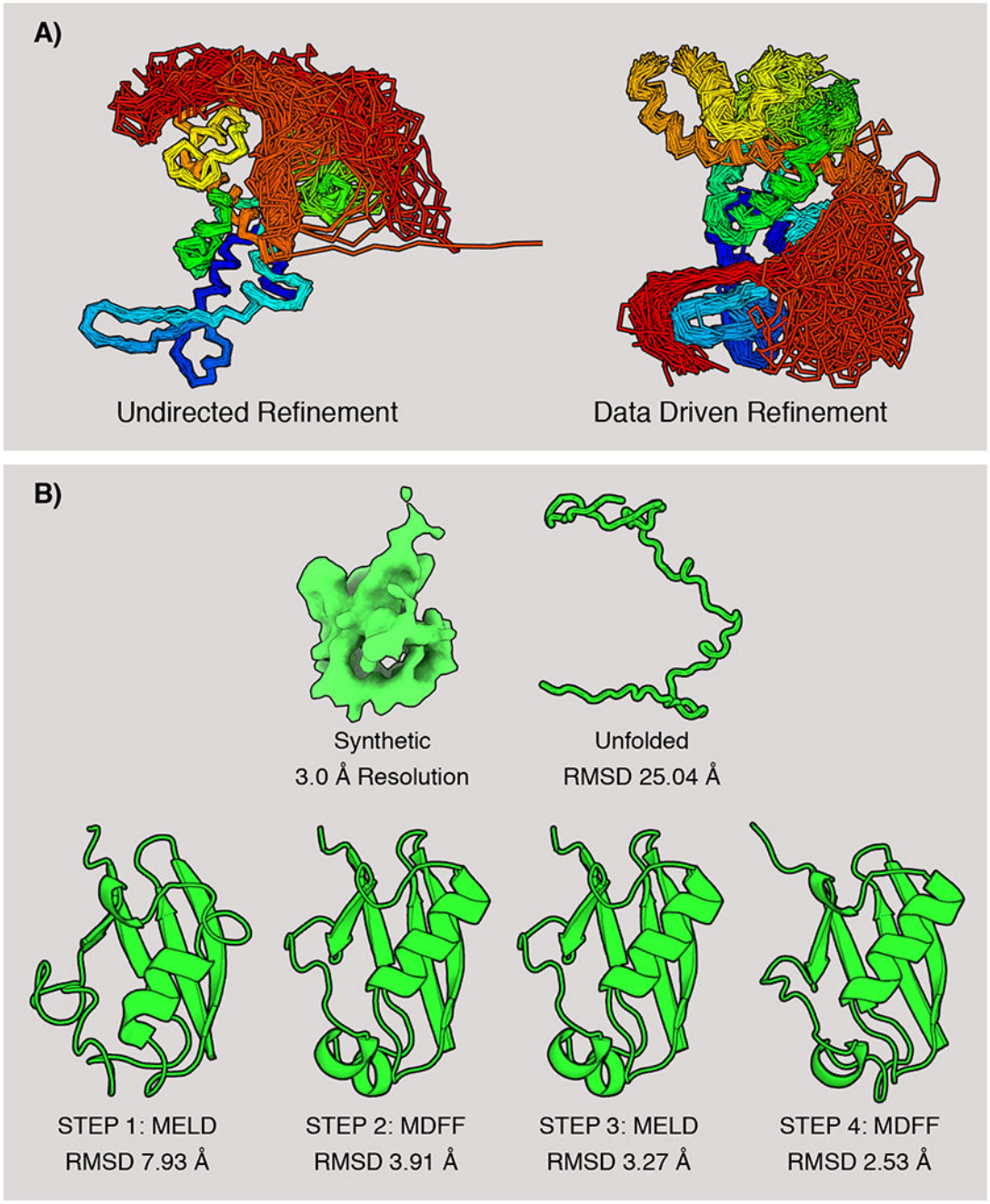

Figure 2: Ensemble models for TRPV1 and the refinement protocol for ubiquitin.

(A) Ensemble refinement with CryoFold showcased for the soluble domain of TRPV1. Several conformations from the TRPV1 ensemble are superimposed; color coding from blue (N-terminal) to red (C-terminal). In a MELD-only simulation, a soluble loop (indicated in red) artifactually interacted with the transmembrane domains. Following the data-guidance from ReMDFF, this loop interacted with the soluble domains and a more focused ensemble is derived that agrees with the density map. **(B)** Stages of the refinement protocol for a test case, ubiquitin. The initial model is an unfolded coil. MELD was used to generate 50 search

models from just the amino acid sequence, and no usage of the density map data. Then, these models were rigid-fitted into the density map using Chimera[47], and ranked based on their global correlation coefficient. ReMDFF refined the best rigid-fitted model even further. The ReMDFF model with the highest correlation coefficient (CC) to the density map served as a template for the subsequent iteration with MELD. In two consecutive MELD-ReMDFF iterations the RMSD of the folded model relative to the crystal structure (1UBQ) attenuated from 25.04Å to 2.53Å. The RMSD for unlabeled Cα-Cα pairing, reflecting that fit of atoms to density maps do not depend on the labels of the residues, changes from 3.18 Å (step 1) → 1.99 Å (step 2) → 1.54 Å (step 3) → 1.28 Å (step 4). However, unlike all-atom RMSD, such estimates are less sensitive to topological correctness of the model as poor connectivity can still reflects in low deviations from the standard.

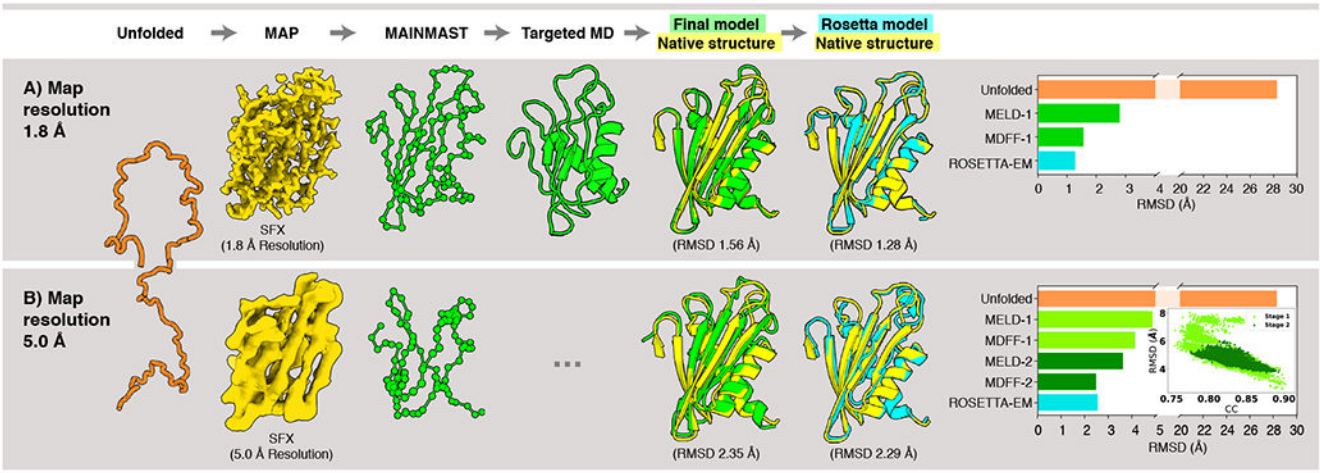

Figure 3: Hybrid structure determination of Flpp3.

(A) High-resolution density map at 1.8 Å resolution. An unfolded structure was used as the initial model. A SFX density map at 1.8 Å resolution was employed to generate the Cα position (green spheres) using MAINMAST, and the initial model was fitted into these positions by targeted MD. The resulting structure (green cartoon model) was then subjected to MELD-ReMDFF refinement. This procedure yielded a structure with RMSD of 1.56 Å relative to the native SFX structure (yellow). The global CC of this structure is 0.83 and Molprobity score is 0.93 with 94.34% Ramachandran favoured backbones and 98.78% favoured sidechains (Table S2). The Rosetta-EM model (cyan) has an RMSD of 1.28 Å with respect to the SFX structure. **(B)** Lower-resolution density map at 5 Å resolution. An initial Cα trace in the map was computed using MAINMAST. Subsequent MELD-ReMDFF refinement resulted in a structure (green cartoon model) with an RMSD of 2.29 Å from the SFX structure (yellow) (Table S3). The best Rosetta-EM model has (cyan) an RMSD of 2.35 Å to the SFX structure. Bar plots depict the evolution of RMSD of the CryoFold models with each subsequent MELD-ReMDFF refinement. The inset of the bar plot in panel B is an RMSD vs global CC scatter plot for the first and second cycle of MELD refinements shown in lime green and dark green, respectively.

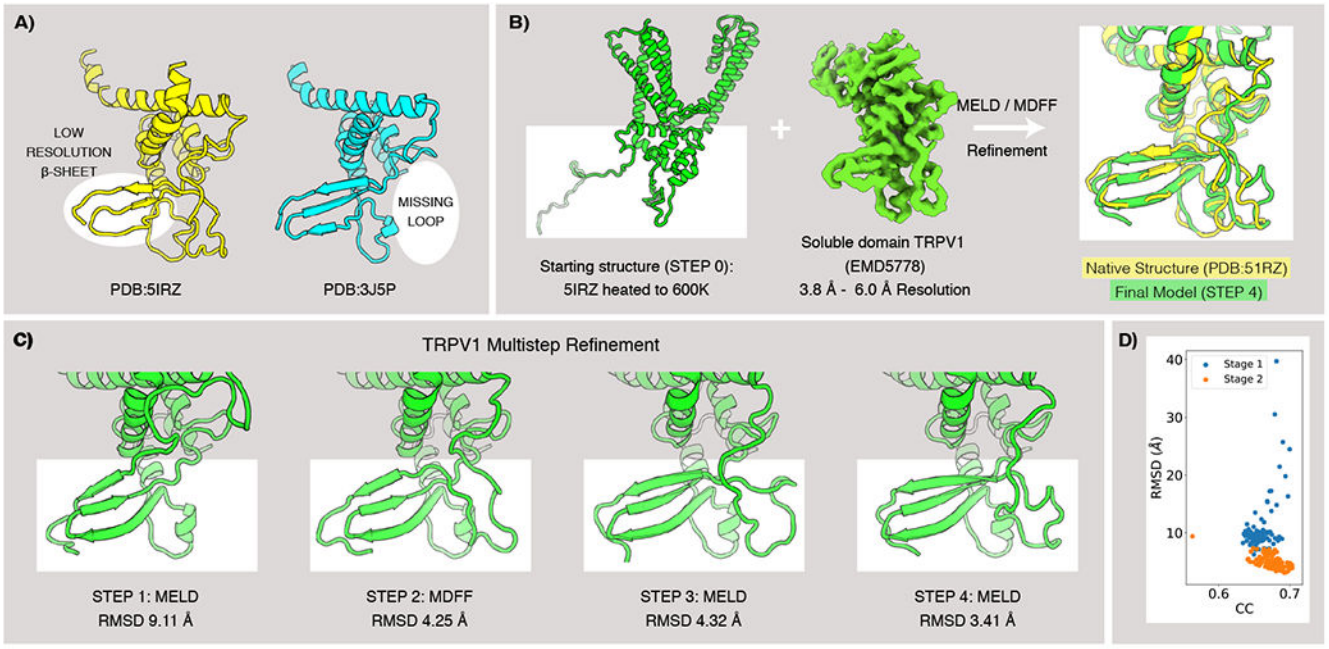

Figure 4: Modeling of the soluble domain of TRPV1.

(A) TRPV1 structures deposited in 2016 (pdb 5IRZ in yellow) and in 2013 (pdb 3J5P in cyan in cartoon representation, showing the latter has a more resolved β-sheet while the former possess an additional extended loop. **(B)** The 5IRZ model was heated at 600 K using brute-force MD, while constraining the α helices. After 10 ns of simulation, this treatment resulted in a search model with the loop regions significantly deviated and the β sheets completely denatured. The search model was subjected to MELD-ReMDFF refinement. A single round of MELD regenerated most of the β-sheet from this random chain, however the 5- to 15-residue long interconnecting loops still occupied non-native positions. Subsequent ReMDFF refinement with the 5IRZ density resurrected the loop positions. One more round of the MELD and ReMDFF resulted in the further refinement of the model. The final refined model agrees well with 5IRZ. **(C)** Progress of the refinement in each step of CryoFold. MELD step 1 shows the β sheets modeled correctly, while the loops recovered in ReMDFF step 2, and refinement was complete by step 4. The approach resulted in structures with 93.75% Ramachandran favored backbones and 92.37% favored sidechains and the Molprobity score of 1.67 (Table S4). Similar to the ubiquitin example, the RMSD for unlabeled Cα-Cα pairing, changes from 2.25 Å (step 1) → 1.28 Å (step 2) → 1.15 Å (steps 3-4). **(D)** Analysis of the MELD ensembles from the first and second MELD-ReMDFF iterations. The scatter plot shows RMSD vs CC for each structure from both ensembles. The ensemble statistics significantly shifts towards models consistent with the density maps, and yet capturing deviations around the best-fitted model, concomitantly accounting for data uncertainty.

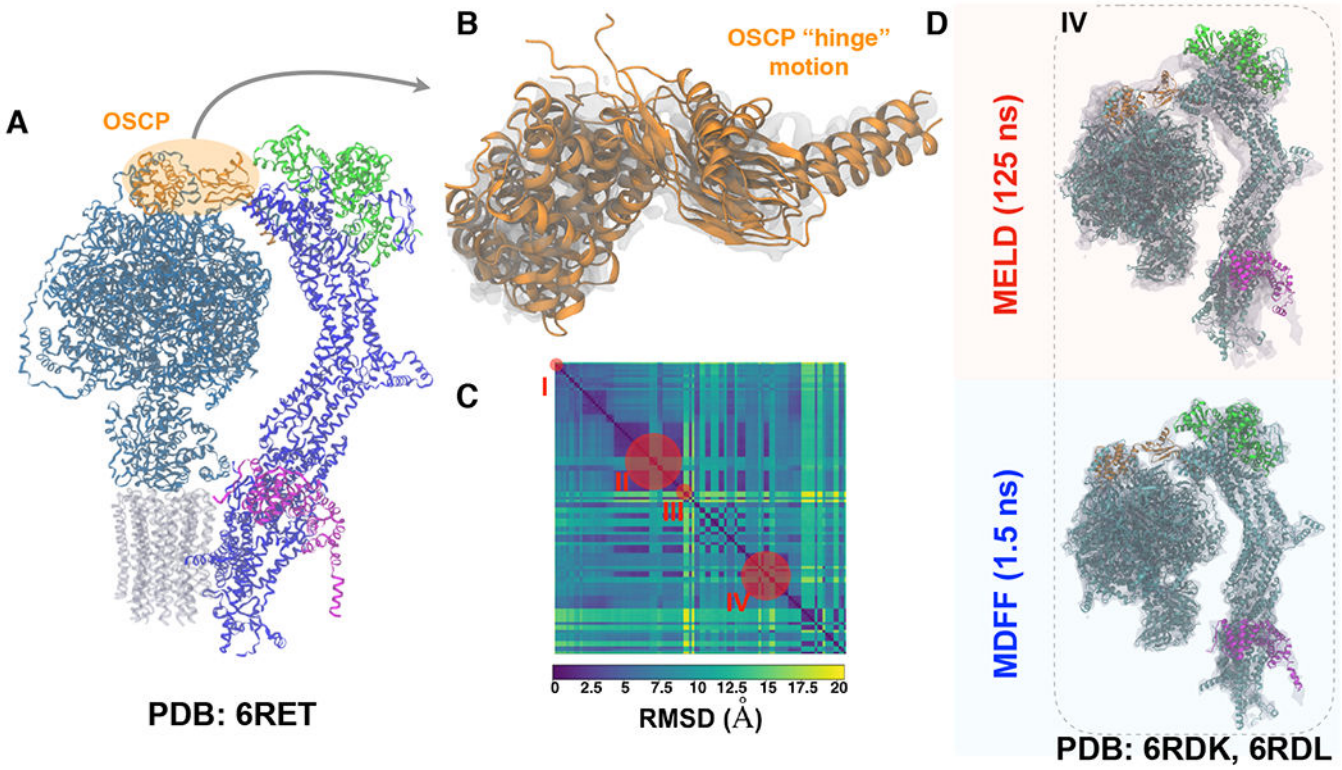

Figure 5: CryoFold samples several biologically relevant states of the soluble domain of mitochondrial F_1 - F_0 ATPsynthase.

We modeled mitochondrial F_1 - F_0 ATPsynthase starting from pdb 6RET (state I) and excluding the grey region embedded in the membrane from refinement. CryoFold samples different conformations through a hinge motion in the OSCP region (orange) connecting the arm (blue) with the rotary domains (cyan). Clustering and 2D-RMSD analysis shows Cryofold samples conformations of additional ATPsynthase states represented by pdb codes 6RDK, 6RDL (state IV). Ohter states represented by pdb codes 6RDQ, 6RDR (state II) and 6RDW, 6RDX (state III) are included in SI.

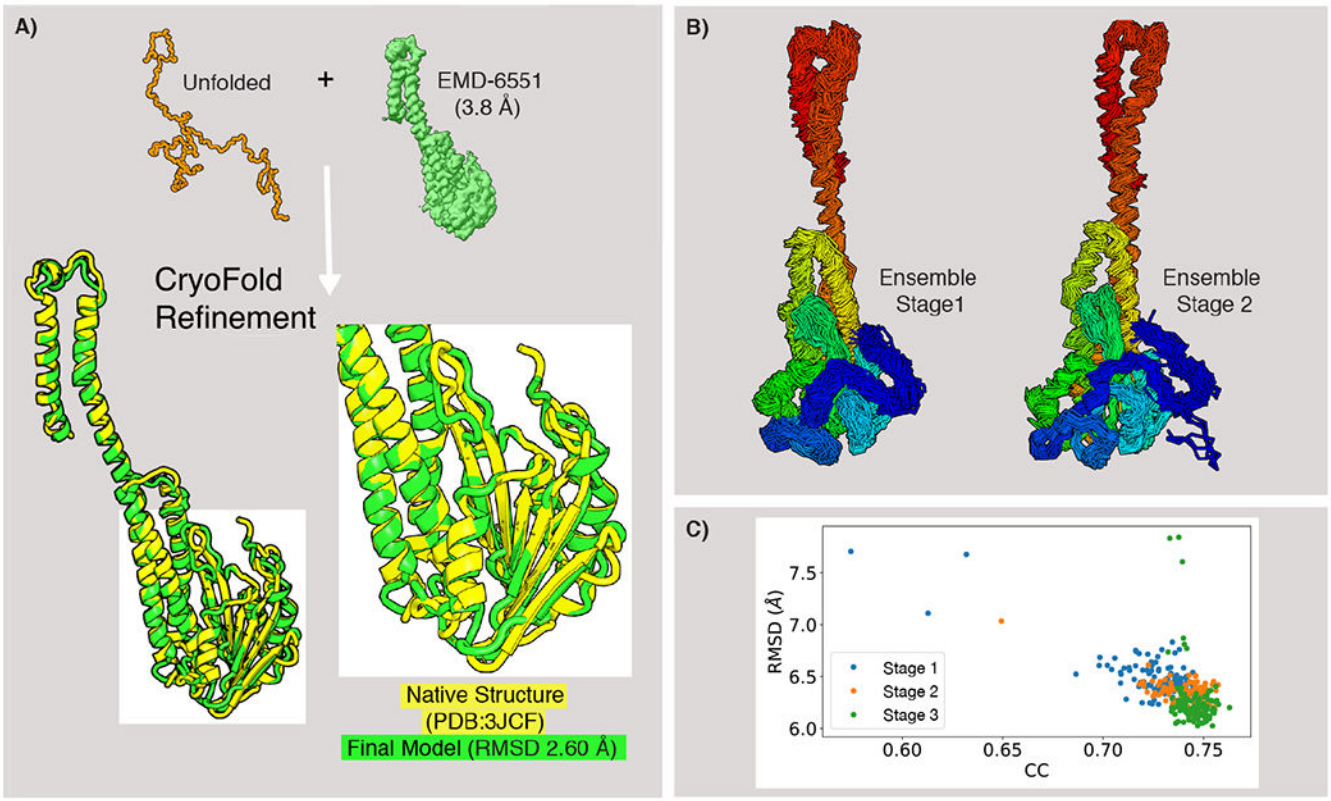

Figure 6: Modeling transmembrane Magnesium-channel CorA.

(A) The CryoFold protocol on CorA. A starts from an Cα trace based Cryo-EM density map using MAINMAST and refined through different cycles of MELD and ReMDFF produces a structure that agrees well with the reported native structure (yellow), featuring accurate beta structures. (B) CryoFold produces narrower, more constraint ensembles as we iterate through MELD/MDFF. (C) A scatter plot of RMSD vs CC derived from the MELD ensembles at every stage of three MELD-ReMDFF iterations. The end-model refined using ReMDFF of the third-stage MELD ensemble is 2.60 Å RMSD from the reported structure.