Cryofold 2.0: Cryo-EM Structure Determination with MELD

Liwei Chang,[†] Arup Mondal,[†] Justin L. MacCallum,[‡] and Alberto Perez[*,†]

†*Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, FL 32611, USA.*

‡*Department of Chemistry, University of Calgary, Calgary AB T2N 1N4, Canada*

E-mail: perez@chem.ufl.edu

Abstract

Cryo-electron microscopy data is becoming more prevalent and accessible at higher resolution levels, leading to the development of new computational tools to determine the atomic structure of macromolecules. However, while existing tools adapted from X-ray crystallography are suitable for the highest-resolution maps, new tools are needed for lower-resolution levels and to account for map heterogeneity. In this paper, we introduce CryoFold 2.0, an integrative physics-based approach that combines Bayesian inference and the ability to handle multiple data sources with the molecular dynamics flexible fitting (MDFF) approach to determine the structures of macromolecules using cryo-EM data. CryoFold 2.0 is incorporated in the MELD (Modeling Employing Limited Data) plugin, resulting in a more computationally efficient and accurate pipeline than running MELD or MDFF alone. The approach requires fewer computational resources and shorter simulation times than the original Cryofold, and minimizes manual intervention. We demonstrate the effectiveness of the approach on eight different systems, highlighting its various benefits.

Introduction

Cryo-electron microscopy (cryo-EM) is now a well-established experimental technique for determining the structures and assemblies of biomolecules in atomistic detail. As of February 2023, the number of cryo-EM structures deposited in the Protein Data Bank (PDB, https://www.rcsb.org)[1,2] has surpassed those deposited using nuclear magnetic resonance (NMR). This progress can be attributed to advancements in both hardware and software components for data collection, image processing, and 3D map reconstruction. These innovations have enabled researchers to determine single-particle structures at subnanometer resolution.[3] When the resolution is better than 2 Å, structure-building workflows originally developed for X-ray crystallography, such as Coot,[4] Phenix,[5] CNS[6] or REFMAC,[7] can be applied to determining atomic positions in the map.[8] For samples with resolutions higher than 3.5 Å, recent developments in deep learning-based tools, such as DeepTracer[9]and ModelAngelo,[10] have been instrumental to reduce the need for demanding human intervention.

Despite the significant progress made in pushing the resolution limit of cryo-EM, a considerable portion of the cryo-EM maps in the Electron Microscopy Data Bank (EMDB, `www.emdatabank.org`) have lower resolutions and lack structural models.[11] Additionally, the resolution heterogeneity that reflects diverse structural ensembles presents a challenge for traditional approaches developed for more homogeneous samples (e.g. from X-ray crystallography). Several physics-based approaches combine molecular simulations with electron density maps to identify the structures and other properties of the system. For instance, molecular dynamics flexible fitting (MDFF),[12] self-guided Langevin dynamics,[13] correlation-driven Molecular Dynamics,[14] and normal mode molecular dynamics (NMMD)[15] have been implemented in popular MD packages to help in the fitting and refinement of molecular models guided by cryo-EM density maps.

CryoFold[16] was recently introduced as a computational pipeline synergizing the MAINMAST,[17] ReMDFF,[18] and MELD[19] approaches to overcome the limitations of each individual program. CryoFold begins with MAINMAST, which generates an initial backbone

model of a protein given an electron density map. ReMDFF is then employed to fit the all atom structure to the cryo-EM map, refining backbone and sidechain orientations. However, ReMDFF cannot correct misfolded secondary structures. In the third step, we select an initial model from ReMDFF and calculate existing contacts between residues which are then provided as a noisy dataset to MELD – since MELD is not directly aware of the cryoEM density map. In the fourth step, MELD samples through multiple partial unfolding and refolding events that satisfy different interpretations of the contacts and is able to sample alternative secondary structures. MELD ensmbles are then analyzed to identify the models with highest cross correlation coefficient to the experimental density map and MDFF is used to refine the agreement of the model with the density map. Thus, Cryofold requires several iterations between MELD and MDFF to arrive at the best agreement with the density map.

CryoFold[16] has demonstrated its ability to build high-resolution structures for both small and large protein systems and in the 2019 EMDataResourse Challenge competition.[20] However, one challenge of the CryoFold pipeline is the need of human intervention in changing format between MAINMAST, MDFF and MELD, as well as choosing guiding information and its uncertainty to use in the MELD stage. The lack of integration limits the usability and performance of the pipeline. Several recent papers[21–23] highlight the need for integrated platforms in the community to reduce problems associated with formatting and software interoperability. The integrative modeling platform (IMP),[24] is a prime example of an integrative approach.

Here, we introduce CryoFold 2.0, which builds upon the same principles as CryoFold but integrates the functionality of ReMDFF directly into MELD, resulting in a single, integrated platform for solving biomolecular structures with cryo-EM maps. This new approach offers several advantages over the earlier version. First, we have eliminated the need for human intervention between the different stages of CryoFold. Second, all components are now aware of the cryo-EM density map. And third, this integration is compatible with MELD's philosophy to combine other sources of data in regions where cryo-EM data is limited such

as crosslinking mass spectroscopy.[25] MELD has already shown promising results in solving biomolecular structures using different sources of data, such as sparsely labeled NMR samples[26,27] or solid-state NMR Paramagnetic Relaxation Enhancements data.[28] We have tested CryoFold 2.0 on eight different systems using diverse starting models to highlight the benefits of the current approach.

Methods

Modeling Employing Limited Data (MELD). MELD uses a Bayesian inference approach to accelerate physics-based molecular simulations guided by external ambiguous and noisy information, while simultaneously determining the best interpretation of the data compatible with the physics model. Under the Bayes theorem formulation,

$$p(x \mid D) \sim p(D \mid x)p(x), \tag{1}$$

the prior probability of a conformation x, $p(x)$, is the Boltzmann probability distribution determined by the selected force field such as Amber ff14SBside[29,30] + ff99SB[31] with a generalized Born implicit solvent model,[32] $e^{-E_{Amber}(x)/kT}$. The external information enhances Molecular dynamics sampling in regions compatible with the data (e.g., distances between pairwise atoms). The likelihood of the guiding data D given x, $p(D \mid x)$, is proportional to $e^{-E_r(x)/kT}$, where T is the temperature and $E_r(x)$ is the restraint energy from MELD. To avoid kinetic traps, MELD uses a flexible Hamiltonian and temperature replica exchange molecular dynamics (H,T-REMD) scheme as the sampling engine.[33,34] Additionally, MELD can selectively activate only a subset of the restraints based on the ranking of their restraint energies calculated from each sampled conformation given external information. MELD employs GPU-accelerated OpenMM[35] for efficient computation.

Grid force from cryo-EM data. After processing the raw 2D images obtained from cryo-EM data, a 3D density grid map is generated which can be used to fit the atomistic

structure in various ways. One popular approach for this employs molecular dynamics to achieve a flexible fitting (MDFF) of the model into the density map. In traditional MDFF, an additional potential from the density map is generated for fitting the initial structure using molecular dynamics as follows:

$$V_{\mathrm{EM}}(\mathbf{r}) = \begin{cases} \zeta\left(1 - \frac{\Phi(\mathbf{r})-\Phi_{\mathrm{thr}}}{\Phi_{\mathrm{max}}-\Phi_{\mathrm{thr}}}\right), & \Phi(\mathbf{r}) \geq \Phi_{\mathrm{thr}} \\ \zeta, & \Phi(\mathbf{r}) < \Phi_{\mathrm{thr}} \end{cases} \tag{2}$$

where $\Phi(\mathbf{r})$ is the density value at position $\mathbf{r}$, Φ_{thr} is the threshold for the density dataset to exclude solvent data with low density values. ζ is a scale factor to control the strength of density map potential and also defines a flat potential for the solvent region. $\Phi_{\mathrm{max}} = \max(\Phi(\mathbf{r}))$, which is designed to drive atoms into the high density region with low potential energy. The total energy U_{t} of the system is given by $U_{\mathrm{t}} = U_{\mathrm{ff}} + U_{\mathrm{EM}} + U_{\mathrm{add}}$, where $U_{\mathrm{EM}} = \sum_i w_i V_{\mathrm{EM}}(\mathbf{r}_i)$ with w_i usually being the mass of atom i and U_{add} can be additional restraints such as secondary structure restraints to prevent overfitting to low resolution regions. For low resolution maps, the density potential energy surface is smooth, an ensemble of conformations can be sampled from MD with the density map potential. For high resolution maps, however, the data describes structural features near atomistic level that can cause the initial structure to be stuck in a local rather than global minima of the energy surface. This problem can be alleviated by replica exchange sampling with density maps at different resolutions. Following the resolution exchange methodology of Singharoy *et al.*,[18] the potential energy at $\mathbf{r}$ can be expressed by

$$V_{\mathrm{EM}}(\mathbf{r}) = \sum_n c_n G(\mathbf{r}; \mathbf{r}'_n, \sigma'_n) \tag{3}$$

where $\mathbf{r}'_n$ and σ'_n are the centers (points on density map grid) and width of Gaussian components, c_n is the weighting factor. By applying a Gaussian blur kernel with width σ, the potential map becomes

$$V_\sigma(\mathbf{r}) = \sum_n c_n \exp\left(-\frac{\|\mathbf{r} - \mathbf{r}'_n\|^2}{2(\sigma^2 + \sigma'^2_n)}\right) \tag{4}$$

where c_n is the grid potential given by the density map data and $\sigma_n'^2$ is the blurring scale. This strategy is well-suited to the MELD methodology in that the cryo-EM density map can serve as an additional restraint source and the resolution exchange can be easily handled by the Hamiltonian replica exchange scheme. At certain times **t** during simulations, the systems on replica **i**, **j** with blurring scale σ_i, σ_j, respectively, are exchanged under the Metropolis acceptance criterion:

$$p\left(\mathbf{x}_i, \sigma_i, \mathbf{x}_j, \sigma_j\right) = \min\left(1, \exp\left(\frac{-U\left(\mathbf{x}_i, \sigma_j\right) - U\left(\mathbf{x}_j, \sigma_i\right) + U\left(\mathbf{x}_i, \sigma_i\right) + U\left(\mathbf{x}_j, \sigma_j\right)}{k_B T}\right)\right), \quad (5)$$

where k_B is the Boltzmann constant, $U\left(\mathbf{x}, \sigma\right)$ is the total potential energy of the system at position $\mathbf{x}$ determined by the density map with blurring scale σ.

Thes updated MELD package with the implementation of grid force for structure modeling with cryo-EM data is open-source and available for download at `https://github.com/maccallumlab/meld`.

Simulation protocols and benchmark systems. In this study, we performed cryo-EM guided simulation for eight systems. The Amber ff14SBside force field with a Generalized Born implicit solvent model were choosed for parameterizing the system and the Langevin integrator implemented in OpenMM with friction coefficient 1 ps$^{\check{}1}$ was used to run the simulation. For adenylate kinase (ADK), carbon monoxide dehydrogenase (CODH), MAJIN, CdiA and Cdil, we generated the synthetic density map using the target structure at varying resolutions. For SARS-CoV-2 spike protein, NSP2 and ATPase NSF, the corresponding experimental maps were used during simulation. The simulation details for each system are summarized in Table S1.

Results

Cryo-EM data drives global conformational transitions in adenylate kinase and carbon monoxide dehydrogenase. We first examined the efficacy of the grid force fea-

ture in MELD on two widely used systems, namely ADK and CODH. For both systems, two conformations that exhibit global conformational changes between a closed and open state have been crystallized (Table S1). To assess the fitting process for ADK, a simulated density map of the open conformation was generated at 5 Å resolution using Chimera.[36] The closed conformation was then fitted into the density map as the starting conformation (Fig. 1A). To prevent overfitting in the density map, the secondary structure restraints were also enforced in addition to the grid force from density map.[18] The fitting process was evaluated by calculating the RMSD and correlation coefficient (C.C.) between the simulation and target conformation, which converged to a structure less than 1 Å against the native after 0.4 ns (Fig. 1A). In the case of CODH, we generated a 3 Å synthetic density map of the open form using the same tool. Directly fitting the closed conformation into this density map would lead to trapping in local minima due to the increased ruggedness of the energy landscape derived from the high-resolution density map (see Fig. 1B inset).[18] We adopted the resolution exchange strategy to fit the closed conformation, using eight replicas for which the density maps had an incremental Gaussian blurring (scaled from 0 to 2, see Methods). The simulation quickly converged from the closed form to within 1 Å of the open conformation.

Refinement using Hamiltonian and temperature replica exchange improves the local structure. A good starting conformation as in the above two cases is not always available for performing structure fitting against cryo-EM data. Recent advances in structure prediction largely enriches the protein structure database, however, these methods are typically not capable of generating accurate predictions in regions lacking co-evolution information or adopting alternative conformations. We present results for a set of systems to demonstrate the benefit of combining temperature and resolution exchange with cryo-EM data to improve local structure refinement. Two systems that require fitting the flexible regions into the density map are shown in Fig. 2. In the case of the SARS-CoV-2 Spike protein, the starting conformation comes from the top-ranked structure prediction from AlphaFold.

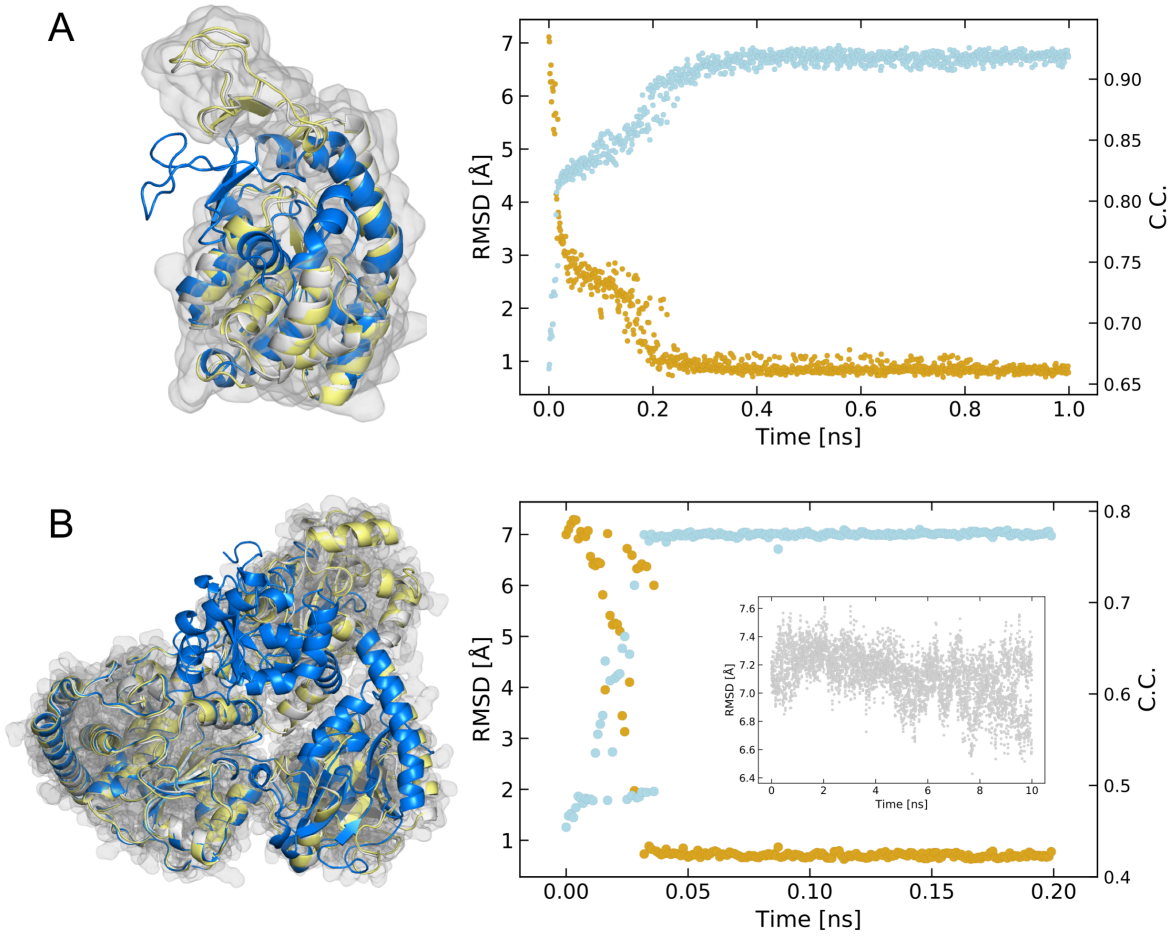

Figure 1: Global conformational transition of (A) adenylate kinase and (B) carbon monoxide dehydrogenase between a closed (blue) and open (yellow) conformation driven by the synthetic density map. Left: starting conformation (blue), fitted conformation (yellow) and native (grey). Right: (A) Time evolution of RMSD (yellow) and cross correlation (blue) between simulation and native structure; (B) Comparison of flexible fitting simulation with density map at only the original resolution (inset) and the lowest replica of resolution exchange simulation with a series density maps of incremental resolutions.

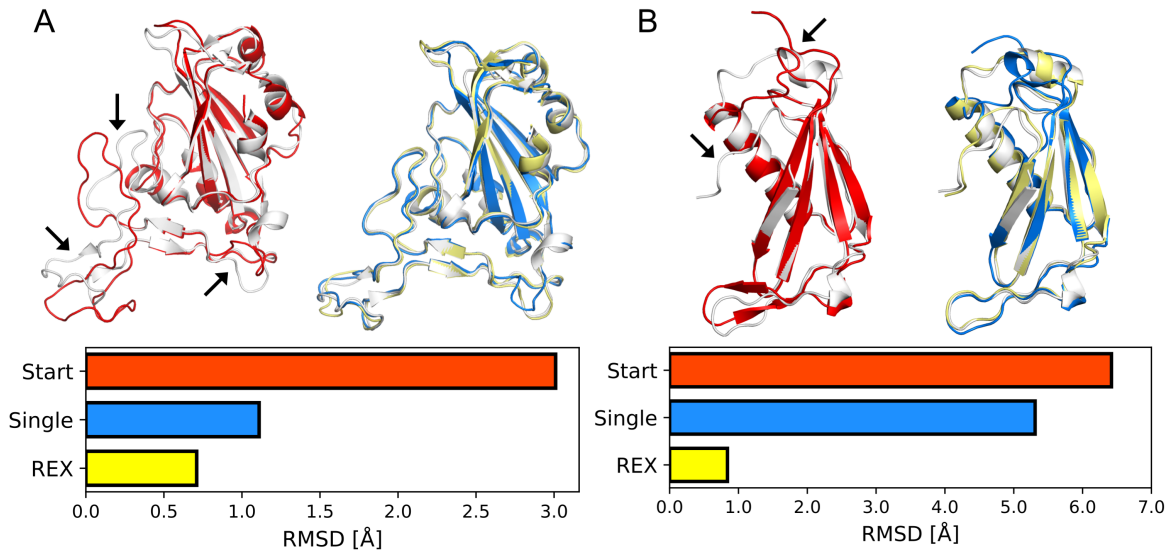

Figure 2: Structure representation (top) and RMSD measure (bottom) of local structure refinement improvement for SARS-CoV-2 spike protein (A) and MAJIN (B) with the starting conformation (red), refined structure with a single replica (blue) and replica exchange based refinement (yellow).

It is clear that direct fitting with the experimental density map can already effectively pull most of the regions into the native conformation, while we can further obtain structures within 1 Å to the deposited model by fitting with temperature and resolution exchange (Fig. 2A). For the MAJIN system, we generated the starting conformation by running simulations at high temperature with positional restraints applied to regions with well-defined secondary structures in the native form. Direct fitting with the synthetic density map of the native structure tends to force regions far from native to the closest region with low potential energy. Better result are obtained by promoting larger exploration of the energy landscape using replica exchange simulations with varying resolutions and temperatures (Fig. 2B).

We also tested the protocol on two other systems, CdiA and CdiI, which require large conformational changes (e.g., α-helix to β-strand transition). Both systems served as prediction targets in the NMR-assisted prediction category of the CASP13.[26,37] The starting conformations were generated by running simulations at high temperature with positional restraints applied to part of the protein. The helical region becomes partially disordered for the first system, and a register-shifted hairpin needs to be corrected (Fig. 3A left). Correct

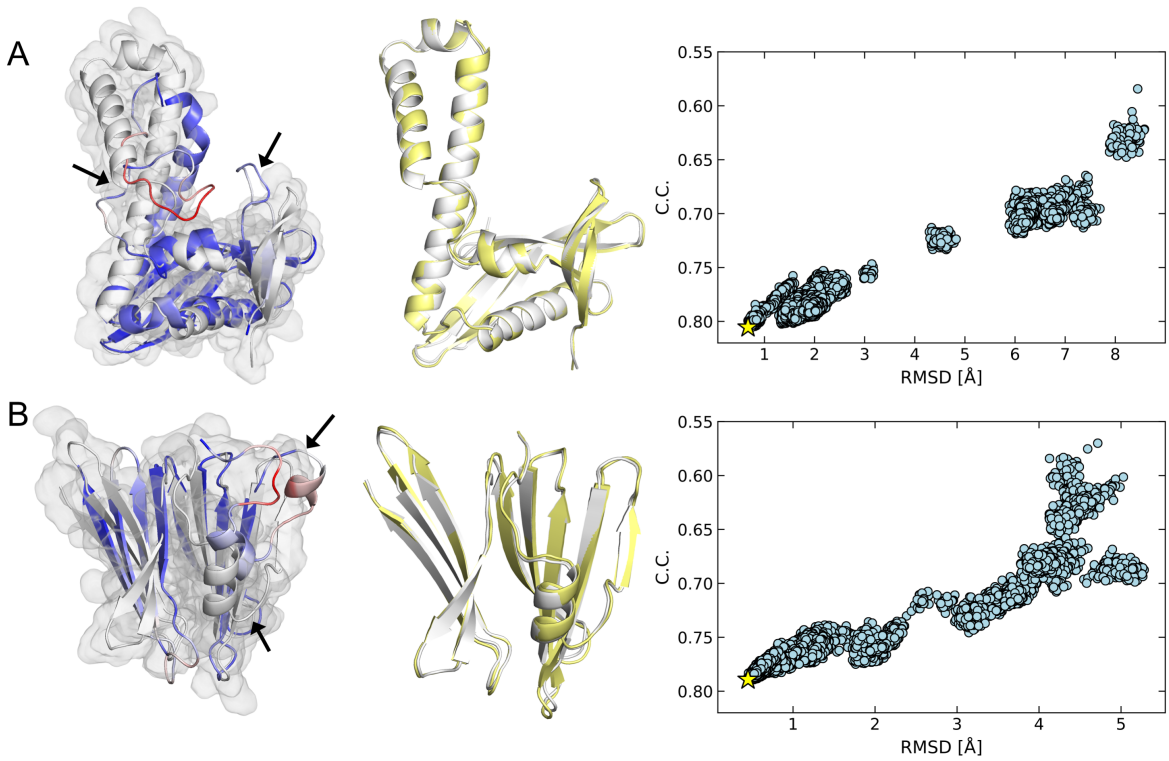

Figure 3: Structure representation (left) and RMSD-C.C. measure (right) of local structure refinement for CdiA (A) and CdiI (B). The starting conformation is colored from blue (low C_α displacement) to red (high C_α displacement) aligned with the native structure (grey). The refined structure is chosen based on the cross-correlation score (yellow, labeled as a star in the right panel).

determination of the second system requires a structural shift in one of the β-sheet layers and an α-helix needs to be converted to a β-strand (Fig. 3B left). Not surprisingly, direct fitting with the density map can only sample narrow ensembles near the starting conformation, resulting in poorly fitted structures (see Fig. S1,2). However, we obtained sub-1 Å structures by applying temperature and resolution exchange in both cases (Fig. 3 center and right panels).

Integrative structure determination from cryo-EM with MELD and AI tools. Cryo-EM density maps can be heterogeneous with varying resolutions in different parts of the system. For regions solved with high resolutions (e.g. less than 3 Å), automatic structure-building tools such as DeepTracer[9] and ModelAngelo[10] can be used to directly predict the

atomic positions. However, such methods are currently not reliable when the resolution decreases. Integrated approaches are usually necessary at this stage to synergize information from several sources to build high-quality structural models. We first demonstrate it in the ATP-bound N-ethylmaleimide sensitive factor (NSF). The cryo-EM map of ATP-bound NSF was initially solved at 4.2 Å with local resolution varying from 4.0 to 8.0 Å, and the structure depicting a six-fold symmetry.[38] The top-ranked model from AF matches the overall structure of a single domain well but requires further refinement to be high accuracy (Fig. 4A). Using it as the starting conformation in the resolution exchange enhanced fitting approach can quickly improve the agreement with the density map, however, the short helix indicated in Fig. 4C largely deviates from the native conformation because it was trapped in the density of a neighboring domain. With the premise of automatic structure-building tools, we used ModelAngelo with the density map and sequence as input to obtain additional structural information. We then extracted positional information by aligning the AF prediction with the fragment output from ModelAngelo using TM-align[39] (Fig. 4B). Combining the positional restraints from the predicted fragments with the density map potential further improves the structural agreement with the native structure and reduces simulation time (Fig. 4C,D).

The SARS-CoV-2 protein NSP2 serves as a final example of the resolution heterogeneity in cryo-EM data. The structural model deposited for the density map was built with DeepTracer and AF predictions, followed by a series of refinement steps.[40] Here, we first run ModelAngelo with the density map, which performs well on the high-resolution region but is not able to provide any information at the C-terminal domain (Fig. 5A). It has been shown that AF can predict subregions of NSP2 well, including the C-terminal region. To combine the structural information from ModelAngelo and AF, we converted their structure output into templates for AF (Fig. S3). Although the C-terminal domain remains distant from other parts in the template, AF can predict a structure resembling the deposited model. The resulting structure was exploited as the initial conformation to perform density map

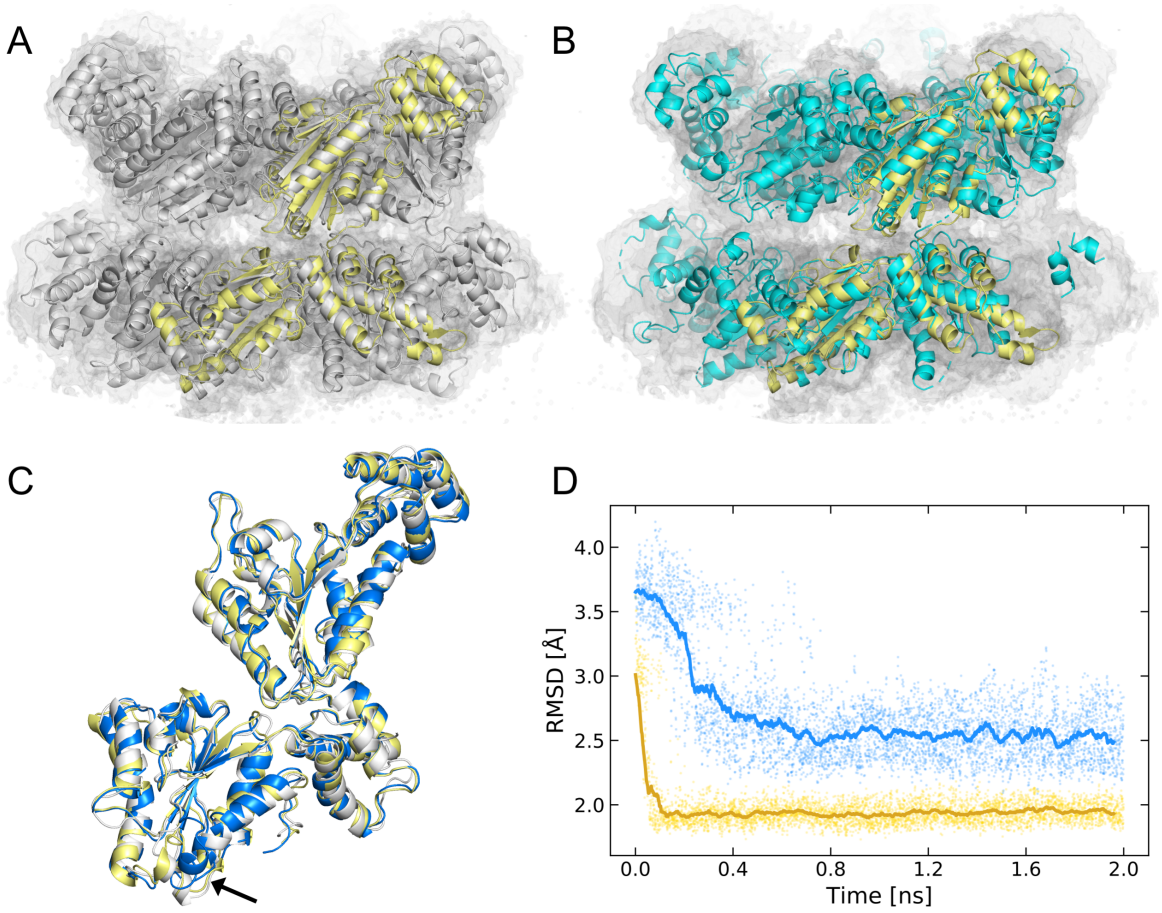

Figure 4: Refinement of ATPase NSF (native structure and cryo-EM data are shown in grey) starting from the AF prediction (yellow in A, B) with fragment predictions from ModelAngelo (B, cyan). The structures of the highest cross-correlation from (D) are shown in blue (refinement without fragment information) and yellow (refinement with fragment information) together with the native domain in (C).

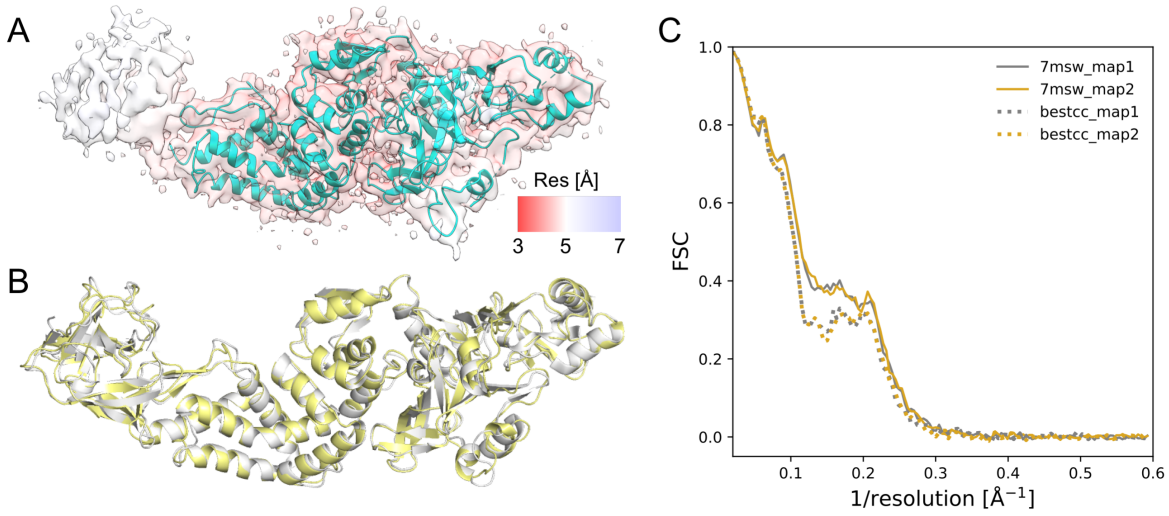

Figure 5: (A) Cryo-EM full map of NSP2 colored by local resolution with fragment prediction from ModelAngelo traced inside. (B) Refined structure with the highest cross-correlation (yellow) aligned with native structure (grey). (C) FSC curve of simulated density map from both structures in (B) against two experimental half maps.

fitting with MELD. In addition to the density map potential, we also used the positional restraints derived from the structure output of ModelAngelo as guidance. It is clear that the representative structure after fitting matches the density map well compared to the deposited model (Fig. 5C), and they differ mostly in the low-resolution regions (Fig. 5B).

Discussion and Conclusion

Advances in experimental and computational approaches to structural biology yield insight into the structures of proteins and their assemblies in larger complexes. The establishment of task forces and data deposition repositories such as the pdb-dev (`https://pdb-dev.wwpdb.org/`) is further promoting the homogenization, transparency, and transferability of data across multiple laboratories. Blind competition studies, including CASP,[41] CAPRI,[42] SAMPL,[43] and others, have contributed to increased insight into methodological advances, the current state of the field, and provided an unbiased look at areas of significant progress. Likewise, the establishment of the cryo-EM structure determination challenge (`https://challenges.emdataresource.org/`) presents a great opportunity to gauge the advances

in determining structures from cryo-EM maps. In the 2019 EMDataResource challenge, the CryoFold approach emerged as one of the top performing groups in determining the structures for different systems given the initial density maps.[20]

The success of CryoFold stemmed from combining three pipelines: MAINMAST,[17] MDFF,[18] and MELD[19] into the structure determination. MAINMAST proposed initial amino acid positions based on the density map, and the resulting contacts were used in MELD to guide sampling. MDFF was then employed to fit the structures rapidly into the density map, and by iterating through MELD and MDFF, optimal structures were obtained. However, in the original CryoFold, the different programs did not integrate seamlessly, necessitating the reformatting of outputs from one program to be used as inputs for the next. Moreover, the MELD stage did not directly utilize the density map, resulting in the exploration of numerous regions of conformational space that would have been inaccessible. Sali's group[44] has already highlighted that the interoperability and compatibility of various computational software platforms remain significant obstacles in structure determination efforts, particularly in integrative or hybrid approaches.

The current version, CryoFold 2.0, incorporates the MDFF methodology into the MELD approach, allowing for the simultaneous exploitation of the benefits of both methods. This integration reduces the amount of sampling required to identify the native state since simulations are guided by the cryo-EM density map while simultaneously benefiting from MELD's ability to sample different secondary structures. This reduction affects both simulation length and the number of replicas needed in MELD. While in the initial CryoFold we typically employed 30 replicas running for hundreds of nanoseconds, the current approach uses 8 (ReMDFF) or 16 (T, ReMDFF) replicas running for up to tens of nanoseconds. Furthermore, the method can seamlessly integrate all other types of data that MELD already models, such as NOESY peaks, chemical shift perturbation, chemical crosslinking mass spectroscopy, FRET, and EPR, along with the associated noise and ambiguity in the dataset.[19] CryoFold 2.0 still necessitates an initial model that can be generated from any of the other

automatic structure building and prediction tools, such as MAINMAST,[17] DeepTracer,[9] ModelAngelo,[10] or AlphaFold,[45] among others.

In the results section, we demonstrated the efficacy of our new approach for various problems in which neither MDFF nor MELD alone would suffice. We also highlighted the advantages of utilizing a Hamiltonian exchange approach, wherein the cryo-EM density map is modified. By artificially decreasing the resolution of the density map, we reduce the frustration in the restraint energy landscape imposed by the density map. As a result, large rearrangements of side chains and backbone are more likely at lower resolutions, provided they are compatible with the force field. These conformations are then further refined at lower replicas where the resolution of the density map is increased. In contrast, such conformational changes are not feasible when using a single high-resolution map (see Fig. 2). Furthermore, results in Fig. S1 and S2 show that further improvement in sampling efficiency can be accomplished by coupling temperature and resolution exchange to the replica ladder. This combination increases the amount of high-accuracy structures sampled (see Fig. S1) and also increases the ability to sample them in cases where resolution exchange alone is not enough (see Fig. S2). Previous observations with other methods regarding overfitting to low-resolution density maps[46] also apply to CryoFold 2.0. For example, we observe that enforcing secondary structure restraints for sampling a conformational transition (e.g., open/close) using resolution maps with less than 4 Å resolution increases the quality of the resulting model.

In conclusion, by integrating the MDFF and MELD components of CryoFold into a single platform (MELD), we have increased the performance and accuracy of the method. Furthermore, this will reduce the barrier of entry to new users, and reduce the amount of human intervention needed. We believe the current version will be more suitable to address future challenges of structure determination with cryo-EM, as well as more complex systems that benefit from the utilization of multiple experimental datasets.

References

(1) Rose, P. W. et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Research* **2017**, *45*, D271–D281.

(2) Wang, Z.; Patwardhan, A.; Kleywegt, G. J. Validation analysis of EMDB entries. *Acta Crystallographica Section D* **2022**, *78*, 542–552.

(3) Hanske, J.; Sadian, Y.; Müller, C. W. The cryo-EM resolution revolution and transcription complexes. *Current Opinion in Structural Biology* **2018**, *52*, 8–15.

(4) Casañal, A.; Lohkamp, B.; Emsley, P. Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. *Protein Science* **2020**, *29*, 1055–1064.

(5) Adams, P. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 213–221.

(6) Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nature Protocols* **2007**, *2*, 2728–2733.

(7) Murshudov, G.; Skubák, P.; Lebedev, A.; Pannu, N.; Steiner, R.; Nicholls, R.; Winn, M.; Long, F.; Vagin, A. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography* **2011**, *67*, 355–367.

(8) Beton, J. G.; Cragnolini, T.; Kaleel, M.; Mulvaney, T.; Sweeney, A.; Topf, M. Integrating model simulation tools and cryo-electron microscopy. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2023**,

(9) Pfab, J.; Phan, N. M.; Si, D. DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2017525118.

(10) Jamali, K.; Kimanius, D.; Scheres, S. ModelAngelo: Automated Model Building in Cryo-EM Maps. *arXiv* **2022**,

(11) Herzik, M. A.; Fraser, J. S.; Lander, G. C. A Multi-model Approach to Assessing Local and Global Cryo-EM Map Quality. *Structure* **2019**, *27*, 344–358.e3.

(12) Trabuco, L. G.; Villa, E.; Schreiner, E.; Harrison, C. B.; Schulten, K. Molecular dynamics flexible fitting: A practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **2009**, *49*, 174–180.

(13) Wu, X.; Subramaniam, S.; Case, D. A.; Wu, K. W.; Brooks, B. R. Targeted conformational search with map-restrained self-guided Langevin dynamics: Application to flexible fitting into electron microscopic density maps. *Journal of Structural Biology* **2013**, *183*, 429–440.

(14) Igaev, M.; Kutzner, C.; Bock, L. V.; Vaiana, A. C.; Grubmüller, H. Automated cryo-EM structure refinement using correlation-driven molecular dynamics. *eLife* **2019**, *8*, e43542.

(15) Vuillemot, R.; Miyashita, O.; Tama, F.; Rouiller, I.; Jonic, S. NMMD: Efficient Cryo-EM Flexible Fitting Based on Simultaneous Normal Mode and Molecular Dynamics atomic displacements. *Journal of Molecular Biology* **2022**, *434*, 167483.

(16) Shekhar, M. et al. CryoFold: Determining protein structures and data-guided ensembles from cryo-EM density maps. *Matter* **2021**, *4*, 3195–3216.

(17) Terashi, G.; Kihara, D. De novo main-chain modeling for EM maps using MAINMAST. *Nature Communications* **2018**, *9*, 1618.

(18) Singharoy, A.; Teo, I.; McGreevy, R.; Stone, J. E.; Zhao, J.; Schulten, K. Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *eLife* **2016**, *5*, e16105.

(19) MacCallum, J. L.; Perez, A.; Dill, K. A. 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* **2015-6**, *112*, 6985–6990.

(20) Lawson, C. L. et al. Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. *Nature Methods* **2021**, *18*, 156–164.

(21) Webb, B.; Viswanath, S.; Bonomi, M.; Pellarin, R.; Greenberg, C. H.; Saltzberg, D.; Sali, A. Integrative structure modeling with the Integrative Modeling Platform. *Protein Science* **2018**, *27*, 245–258.

(22) Koukos, P.; Bonvin, A. Integrative Modelling of Biomolecular Complexes. *Journal of Molecular Biology* **2020**, *432*, 2861–2881.

(23) Braitbard, M.; Schneidman-Duhovny, D.; Kalisman, N. Integrative Structure Modeling: Overview and Assessment. *Annual Review of Biochemistry* **2019**, 113–35.

(24) Russel, D.; Lasker, K.; Webb, B.; Velázquez-Muriel, J.; Tjioe, E.; Schneidman-Duhovny, D.; Peterson, B.; Sali, A. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLoS Biology* **2012**, *10*, e1001244.

(25) Schmidt, C.; Urlaub, H. Combining cryo-electron microscopy (cryo-EM) and cross-linking mass spectrometry (CX-MS) for structural elucidation of large protein assemblies. *Current Opinion in Structural Biology* **2017**, *46*, 157–168.

(26) Mondal, A.; Perez, A. Simultaneous Assignment and Structure Determination of Proteins From Sparsely Labeled NMR Datasets. *Frontiers in Molecular Biosciences* **2021**, *8*, 774394.

(27) Mondal, A.; Swapna, G.; Hao, J.; Ma, L.; Roth, M. J.; Montelione, G. T.; Perez, A. Structure determination of protein-peptide complexes from NMR chemical shift data using MELD. *bioRxiv* **2022**, 2021.12.31.474671.

(28) Perez, A.; Gaalswyk, K.; Jaroniec, C. P.; MacCallum, J. L. High Accuracy Protein Structures from Minimal Sparse Paramagnetic Solid-State NMR Restraints. *Angewandte Chemie* **2019**, *131*, 6636–6640.

(29) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **2015**, *11*, 3696–3713.

(30) Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C. Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent. *Journal of the American Chemical Society* **2014**, *136*, 13959–13962.

(31) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712 – 725.

(32) Nguyen, H.; Roe, D. R.; Simmerling, C. Improved Generalized Born Solvent Model Parameters for Protein Simulations. *Journal of Chemical Theory and Computation* **2013**, *9*, 2020 – 2034.

(33) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999**, *314*, 141–151.

(34) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *Journal of Chemical Physics* **2002-5**, *116*, 9058–9067.

(35) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D. et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology* **2017**, *13*, e1005659.

(36) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **2004**, *25*, 1605–1612.

(37) Sala, D. et al. Protein structure prediction assisted with sparse NMR data in CASP13. *Proteins: Structure, Function, and Bioinformatics* **2019**, *87*, 1315–1332.

(38) Zhao, M.; Wu, S.; Zhou, Q.; Vivona, S.; Cipriano, D. J.; Cheng, Y.; Brunger, A. T. Mechanistic insights into the recycling machine of the SNARE complex. *Nature* **2015**, *518*, 61–67.

(39) Zhang, C.; Shine, M.; Pyle, A. M.; Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods* **2022**, *19*, 1109–1115.

(40) Gupta, M. et al. CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes. *bioRxiv* **2021**, 2021.05.10.443524.

(41) Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics* **2021**, *89*, 1607–1617.

(42) Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J. E.; Vajda, S.; Vakser, I.; Wodak, S. J. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function, and Bioinformatics* **2003**, *52*, 2–9.

(43) Rizzi, A. et al. The SAMPL6 SAMPLing challenge: assessing the reliability and efficiency of binding free energy calculations. *Journal of Computer-Aided Molecular Design* **2020**, *34*, 601–633.

(44) Hancock, M.; Peulen, T.-O.; Webb, B.; Poon, B.; Fraser, J. S.; Adams, P.; Sali, A. Integration of software tools for integrative modeling of biomolecular systems. *Journal of Structural Biology* **2022**, *214*, 107841.

(45) Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, 1–11.

(46) Leelananda, S. P.; Lindert, S. Using NMR Chemical Shifts and Cryo-EM Density Restraints in Iterative Rosetta-MD Protein Structure Refinement. *Journal of Chemical Information and Modeling* **2020**, *60*, 2522–2532.

Associated content

Supporting Information

Detailed information for selected modeling systems and complementary analysis of simulations.

Data Availability

The source code can be accessed from `https://github.com/maccallumlab/meld`. A tutorial for running cryo-EM guided simulation in MELD can be found here (`http://meldmd.org/tutorial/cryofold_tutorial/cryofold.html`).

Author Information

Corresponding Author:

*perez@chem.ufl.edu

Orcid:

Liwei Chang: 0000-0001-5847-0820

Arup Mondal: 0000-0002-8970-3380

Justin L. MacCallum: 0000-0001-7917-7068

Alberto Perez: 0000-0002-5054-5338

Notes

The authors declare no competing financial interest.

Acknowledgements

The research was sponsored by the NSF Career award CHE-2235785. The authors are thankful for computational resources from the HiPerGator supercomputer at the University of Florida.

TOC Graphic