Development of Automated Methods for Modeling Ligands in Cryo-Electron Microscopy Data

Andrew Muenks

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Frank DiMaio, Chair

David Shechner

Ning Zheng

Program Authorized to Offer Degree:

Biochemistry

University of Washington

Abstract

Development of Automated Methods for Modeling Ligands in Cryo-Electron Microscopy Data

Andrew Muenks

Chair of the Supervisory Committee:

Frank DiMaio

Department of Biochemistry

Advances in cryo-electron microscopy (cryoEM) and deep-learning guided protein structure prediction have expedited structural studies of protein complexes. However, methods for accurately modeling ligand conformations are lacking. For my doctoral thesis, I developed computational methods to automatically determine both ligand conformation and identity in medium- to low- resolution cryoEM maps. These methods utilize both a small molecule force field in Rosetta and information in cryoEM data. First, a ligand fitting protocol EMERALD accurately predicts ligand conformations along with surrounding side chains in maps as low as 6 Å local resolution. Then, I further expand the capabilities of EMERALD to produce small molecule models similar to deposited models with 20 or more torsion angles. Finally, libraries of common ligands and lipids are screened through EMERALD to assign identities to ligand blobs of density. Together, these automatic tools can fit into cryoEM modeling pipelines for determination of protein-ligand complexes.

List of Figures

Acknowledgements

Scientific progress is impossible without the collaboration, mentorship, and community of others. First, I would like to thank Frank DiMaio for advising me during my thesis. You allowed me to work independently but were always generous with your time when needed, creating a work environment where I could make mistakes and learn at my own pace. I appreciate the flexibility I had when working which allowed me to spend more time with my family. I cannot think of a better space for completing my thesis work.

I would also like to thank my other mentors in science. My committee members — Neil King, David Veesler, Ning Zheng, and David Shechner — all provided valuable feedback and suggestions for my thesis work. Additionally, thank you to my undergraduate mentors Lesa Beamer, Christopher Lee, and Kyle Stiers, who introduced me to computational structural biology and helped me navigate towards graduate school in the first place.

I am fortunate to have fantastic lab mates who made me look forward to coming into the lab. The second generation of lab members taught me so much, whether it be Dan Farrell demystifying Rosetta, Guangfeng Zhou showing me the ropes of GALigandDock, Gabi Reggiano sharing code for processing cryoEM maps, or Gabi or Carson Adams providing valuable advice for navigating graduate school. With these four, along with members Marisa Brandys, Ryan McHugh, Jacob North, Davi An, and Michael Wiem, I always knew I had someone to join me at departmental events or take a break over an Aladdin's gyro or pitcher of Rainier at the College Inn. My fondest memories of graduate school were those surrounded by my lab mates, and I hope to continue our relationships for the rest of our lives.

Several others in the Biochemistry Department have been instrumental. Thank you to Erin Kirschner for keeping me on track during my studies. Thanks to my fellow DEI Committee

members and the UAW 4121 representatives for making the university more livable for myself and others. I'm happy to have interacted with all biochemistry grad students during my PhD, with special emphasis to my 2018 cohort, Chloe Adams for being my blood donation buddy, Justin Applegate for always being down for a concert, and Sam Zepeda for collaboration in this thesis and the manuscript from which Chapter 2 is derived [1].

Much of this work was funded by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1762114. This funding would not have been attainable without the mentorship of Liangcai Gu during my first quarter of graduate school. This work has also been funded in part by a collaboration with Thermo Fisher Scientific and Nvidia.

Finally, I would like to thank all my friends and family for their support during my thesis. Thank you to my parents for their constant encouragement, despite biochemistry and earning a PhD being foreign to our home community. To my wife Madeline, thank you for traveling halfway across the country with me, only for us to travel back, only to be long-distance for two long years. While the world changed so much around us, our love and joy remained constant.

Chapter 1. Introduction

Proteins are molecular machines that perform nearly all functions within a cell. Central to the activity of proteins are their interactions with ligands, which can act as substrates or regulators of the protein. Analyzing these protein-ligand interactions with three-dimensional molecular models are vital to understanding the functions, regulation, and mechanisms of proteins. Traditionally, X-ray crystallography has been the method of choice for macromolecular structure determination. While models with high-resolution are achievable with X-ray crystallography, the specific requirements necessary for protein crystallization preclude the use of the method for large protein complexes, protein samples with heterogeneity, and membrane proteins.

Recent advancements in both microscope hardware and computational processing have led to cryo-electron microscopy (cryoEM) emerging as a mainstream method for biomolecular structure determination. Unlike X-ray crystallography, protein crystallization is unnecessary for cryoEM as protein sample is flash frozen in a thin layer of vitreous ice and then imaged with an electron microscope. While in ideal cases cryoEM data approaches atomic resolutions [2, 3, 4], most structures determined by cryoEM are in the 3–5 Å resolution range. At these resolutions, model building is time consuming, error prone, and often ambiguous. To assist this process, methods have been developed to automatically build de novo polypeptide chains into EM data [5-8], and with the advent of protein structure prediction via deep learning, high-quality starting models can oftentimes be obtained from sequence information alone [9, 10, 11]. While these methods help build protein models into cryoEM density, tools for automatic fitting of small molecule ligands into cryoEM data are limited. Given the widespread adoption of cryoEM in

academia and in industry to support translational studies of drug targets, the ability to accurately model ligand-bound structures is paramount.

There are numerous automated tools from X-ray crystallography for modeling small molecule ligands [12-15]. However, their methodology is unproven for use in interpreting all but the highest resolution cryoEM maps. Traditional ligand fitting methods rely on shape and topological features of density maps to match [12] or build [13, 14, 15] the ligand into density. But as resolution decreases below 3 Å, the topological features these methods rely on become less defined, and their success rates fall below 20% [15, 16]. While these software packages have been updated to consider cryoEM data, the updates focus on protein modeling without reported updates to small molecule modeling [17, 18] or focus on small molecule refinement instead of automatic model building [19].

Along with ligand fitting, several tools have been developed for identifying ligands from an X-ray crystallography density map. Like with fitting, the preceding methodology relies heavily on shape features of unmodeled ligand density blobs by using ratios of cross-correlations [20] or comparing topological features between skeletonized density and ligand atoms [21]. The reliance on shape features leads to worse performance as resolution decreases, with the methods unproven for ligand identification at resolutions lower than 2.5 Å [21]. Machine learning has been leveraged to use map and ligand features beyond simple shape and topology for ligand identification [22], but still has decreasing returns at lower resolutions and is only recommended for high-resolution cryoEM.

Along with map features, chemical force fields have provided an energetic approach to accurately fit ligands into their respective density for cryoEM. Two approaches — GemSpot [23] and MDFF [24] — utilize the ligand-docking software GLIDE to model ligands into cryoEM

data. However, both require user input in either selecting models during the protocol or choosing a starting configuration, limiting the automation and applicability of these approaches. Another EM ligand fitting tool, ChemEM, recently utilized its own force field and difference maps to model small molecules [25]. But, it was benchmarked on a limited dataset and only considers side chain flexibility during refinement, possibly leading to poorer docked ligand conformations. Difference maps have also been utilized to identify potential blobs of the EM map to consider for modeling [26]. However, tools to suggest ligand identities of density blobs remain non-existent for EM data.

The protein modeling software Rosetta recently incorporated a new small molecule force field, RosettaGenFF, which accurately models the energetics of arbitrary biomolecules in a manner balanced against Rosetta's protein force field [27]. Alongside the force field, a genetic-algorithm (GA) optimization method allowing for full receptor side chain flexibility, GALigandDock, was created to effectively search the conformational space needed to determine a minimum energy docked conformation. Combining the energy model and docking protocol yielded superior ligand docking accuracy compared to other state-of-the-art methods when benchmarked on standardized datasets.

In this dissertation, the docking power of RosettaGenFF and GA optimization were leveraged to overcome the challenges of modeling small molecules at near-atomic resolution. CryoEM density data was integrated with the physically realistic force field of RosettaGenFF to create RosettaEMERALD (EM Maps ERoded for Automatic Ligand Docking) for robust ligand modeling into cryoEM maps with no user input during the protocol. The performance of EMERALD was evaluated on ligand-bound protein structures deposited in the EMDB [28] and compared our results to their respective deposited structures and high-resolution crystal

structures when available. Expansions of EMERALD were developed to consider the special modeling cases of ion-mediated small molecule interactions and ligands with large numbers of rotatable bonds. Finally, an identification tool was created that provides probabilities of identity for EMERALD-docked models for an unmodeled region of density. Together, these tools allow automated ligand model determination and can fit into microscopists' structure determination pipelines.

Chapter 2. Ligand fitting into cryoEM data

2.1 EMERALD: a ligand docking protocol for cryoEM modeling

Methodology for cryoEM ligand modeling must first be able to accurately fit small molecules in the EM map. Our solution to this problem is EMERALD, illustrated in Figure 2.1. GALigandDock places ligands in a protein pocket by iteratively refining a pool of 100 conformations, selecting the best 100 models at each generation using predicted energy. To enable this method to use cryoEM density, two changes were integral: density-guided initial ligand placement, and the use of density in model selection at each round.

To ensure high-quality ligand conformations in the ligand pool, randomly perturbed ligand conformers were aligned into unmodeled density to generate half of the initial pool (Fig. 2.1b). Voxels in the density map within 10 Å of the center of mass of the ligand but greater than 2.5 Å from an atom in the receptor were detected. The detected voxels were pared down with a modified erosion algorithm from previously described methods [14, 29] so that only voxels with strong signal in the map remain. Once eroded, the remaining voxels represented a pseudo-atomic skeleton used for ligand conformer alignment.

Blobs of density are often noncontiguous and difficult to separate from noise at lower resolution. To account for the low resolution, the erosion algorithm was performed in two successive steps with increasing strictness on erosion. On the first pass, peaks in the density were detected and eroded only considering voxels sharing a face with each other. This maintains connections between density blobs that may be noncontiguous. The remaining voxels were clustered into potential skeleton networks by separating groups of voxels that are 3 Å away from another group. Only the largest network of voxels was chosen for further erosion to eliminate noisy voxels. The largest group of voxels underwent a second, stricter erosion that considered all voxels that share a face or edge with each other, leading to a pseudo-atomic skeleton.

The skeleton was used during initial ligand conformer generation of the genetic algorithm to populate a starting pool with conformers that already fit into the density. Small molecules were randomly translated and perturbed in the binding pocket and half of the small molecules in the initial pool were "aligned" to the skeleton. For alignment, the ligands were centered on the center of mass of the skeleton, and then atom-skeleton point pairs were determined. The shortest distance of an atom-skeleton pair while searching over all pairs was found, and this search was repeated until either all atoms or all skeleton points had a unique pairing. For the coordinates in each atom-skeleton pair, a "topped out" harmonic function (Eq. 1) was used to restrain ligand atoms:

$$E_{ij} = 36(1 - e^{-x_{ij}^2/9}) \quad \text{(Eq. 1)}$$

where E is an energy penalty applied and x is the distance in Angstroms between the atom-skeleton pair. The ligand is aligned into the density over 2 stages of energy minimization with 20 and 15 short rounds of minimizations with the atom-skeleton restraints updated after each round. While half of the initial ligands were aligned to the density as described above, the other half of

the population were selected from the top 50 models of 5000 random ligand conformations to ensure diversity in the initial population. All side chains within 5 Å plus the radius of the ligand of the initial ligand center of mass were also considered for optimization.

At each iteration, the population of ligand conformers along with their surrounding flexible side chains are further optimized against the sum of a weighted density correlation and the RosettaGenFF energy (Fig. 2.1c). The ligand population and nearby side chains were optimized over 10 generations of a genetic algorithm using default parameters in GALigandDock and a scoring function with a high electron density score weight of 100 to evaluate a ligand's fit into density. Finally, to minimize the energy of the models, the top 20 ligand conformers at the end of the GA were further optimized along with nearby macromolecule atoms using a cartesian minimization in Rosetta (Fig. 2.1d). The full protocol generates a structure in 30–120 minutes, depending on the size of the ligand and the cryoEM map, with the population initialization making up the largest share of EMERALD's completion time.

2.2 Preparation for docking

2.2.1 Curation of dataset

All single-particle EMDB entries with an associated ligand bound structure at 6 Å nominal resolution or better as of September 03, 2021 were obtained. Given the specificity of trying to model ions and glycans, structures with only these types of ligands were excluded from the dataset. Additionally, the set had several cases with small molecules in proximity. To simplify the docking situation, entries with 2 or more ligands within the binding pocket as defined in our docking protocol were also eliminated from the set. To only have entries with fully modeled macromolecule-ligand complex models that fit the EM density well, structures with a density correlation below 0.4, or which left large regions of density unmodeled, were

dropped. When considering the first instance of a unique ligand for each EMDB entry, there were a total of 1704 total cases to process for docking. Only cases with 25 or fewer torsion angles were analyzed as the search space of ligands with more torsions becomes difficult to fully explore during a GA. This, along with losing cases from inherent failure during ligand processing, left 1053 cases to analyze.

2.2.2 File Preparation

For accurate ligand docking, small molecules need proper protonation states and partial charges. However, the protonation state assigned can depend on the protonation assignment method. To determine the most likely protonation state for a small molecule, we calculated protonation states with 3 assignment tools — phenix.elbow [30], openbabel [31], and dimorphite [32] — and selected the protonation state assigned with 2 or more methods. If there was no agreement or failures during the assignment, the phenix.elbow assignment was used for modeling. SDF files of the first instance of each unique ligand-entry pair were downloaded from the PDB and used for input. For processing with phenix.elbow, all possible hydrogen atoms were added to the SDF file using openbabel, and then hydrogen atoms were removed to the final protonation assignment using phenix.elbow. To generate the protonation state with openbabel, the hydrogen atoms were simply added to the downloaded SDF file at a pH of 7.4. Instead of adding protons to a structure, dimorphite protonated small molecules as SMILES strings, which were then converted to a structure via openbabel. All 3 protonation assignment methods agreed for 794 instances with 2 methods agreeing for 157 cases (Fig. 2.2).

With the protonation state assigned, a mol2 file with AM1-BCC partial charges was generated with antechamber [33, 34]. Finally, a Rosetta specific parameters file was created for each ligand. Receptors were cleaned by eliminating non-macromolecular atoms in the PDB file

and replacing modified residues with their unmodified correspondent. The ligand to be docked was added to its position in the deposited structure and randomly translated 0.0-2.0 Å in any direction before docking.

2.3 EMERALD docking results

2.3.1 Benchmarking on EMDB

All 1053 entries were run in triplicate and the lowest-energy model for each individual run was further analyzed for docking success. Because of a low confidence in the reference models due to their low resolution, docked models were not directly compared to their respective reference models. Instead, all reference models were relaxed into their EM density map in Rosetta using the cartesian minimization used following the genetic algorithm to resolve minor clashing or ligand strain. While an RMSD cutoff of 2 Å has traditionally been used for docking success, the lack of confidence in the low-resolution reference models and inability of RMSD to consider receptor contacts led us to divide results further by the number of residues that make hydrogen bonds with the ligand and a ligand density correlation calculated in Rosetta. These metrics were used to categorize docking results as: (**1**) matches (docked pose within 1 Å of relaxed reference model); (**2**) non-match, similar quality (>1 Å RMSD, density correlation$_{dock}$ - density correlation$_{deposited}$ > 0.025 and hydrogen bonds$_{dock}$ - hydrogen bonds$_{deposited}$ > -1); or (**3**), non-match, worse quality (>1 Å RMSD, density correlation$_{dock}$ - density correlation$_{deposited}$ < -0.025 or hydrogen bonds$_{dock}$ - hydrogen bonds$_{deposited}$ < -1).

The results of these docking trajectories are summarized in Figure 2.3. In 57% of the cases, our density-guided docking produced a top model within 1 Å RMSD (considering all non-hydrogen atoms in the ligand) of the deposited model after energy minimization (“match”, Fig. 2.3a.). There were 401 cases (38%) where EMERALD produced a model with an RMSD value

greater than 1 Å, and the model was similar or better than the deposited model in both metrics ("non-match, similar or better quality"). The smallest group belonged to 48 cases where the deposited model was not recapitulated, but the EMERALD model had a worse density fit or fewer hydrogen bonds than the deposited model ("non-match, worse quality;" 5% of cases). We found that incorporating EM data in GALigandDock is necessary for recapitulating deposited ligand structures with high success rates (Fig. 2.4), as the number of matched cases decreases when using GALigandDock without map information.

The resolution of cryoEM maps often varies from the nominal resolution of a map, so to analyze the performance of EMERALD against map resolution, we compared docking results to local resolution rather than nominal resolution. Maps with local resolution calculations were generated with MonoRes via the Xmipp software package [35]. The deposited maps were filtered with a Gaussian kernel with a sigma of 0.02 times the map dimensions. Binary masks were created using the filtered maps by keeping voxels with a value above 0.05 times the maximum voxel value in the filtered map. With the binary masks, local resolution estimate maps for all instances were created. To calculate the local resolution surrounding the modeled ligand, the local resolution of all voxels within 5 Å of the ligand were averaged. Voxels with zero local resolution values were not included in the average. Considering that ligand binding sites are often less-resolved areas of a map, the nominal resolution was reported if the calculated local resolution of a map around the ligand was better than 1 Å than the nominal resolution since an error likely occurred. With the local resolution calculated, we found modeling accuracy decreases as the local resolution of the map surrounding the ligand worsens and as the number of torsions of the ligand increases (Fig. 2.3b, c).

Because of the low resolution of the density maps, it is difficult to interpret the quality of docked poses from density fit and receptor interactions alone. To instill more confidence in docking results, we analyzed the convergence among the top ranked ligand poses across triplicate runs (Fig. 2.3d-f). The distances between atom pairs across models were calculated and results were further divided into those with 2 or more trajectories having their lowest energy models within 1 Å RMSD, within 1 Å for 60% of atoms, or within 1 Å for fewer than 60% of atoms. Of the cases that match the deposited structure, 2 or more of the trajectories converge for 81% of cases, further strengthening the quality of the matched cases (Fig. 2.3d). Moreover, only 23% of the worse-quality cases converge on the same ligand model (Fig. 2.3e). Given how well trajectory convergence agrees with these categories, it can serve as a proxy for confidence when our docked model differs from the reference model in ambiguous cases. 42% of the ambiguous cases have similar top models across our trajectories (Fig. 2.3f), giving us confidence in an alternative model to the deposited structure for those entries.

Our dataset includes 15 of the 20 cases benchmarked using GemSpot, another cryoEM ligand pipeline, with 5 cases filtered out of the dataset for being peptides or having inter-residue bonds like ion coordination. For 13 of the 15 ligands, EMERALD produced a ligand within 1 Å of the deposited structure, with 9 of those placements assessed as confident. For the other two cases, our models disagreed with the deposited model, and GemSpot also found solutions different from the deposited model in these two cases [23].

2.3.2 Validation of docked models with crystal structures

To cross-validate our results — particularly in cases where we found a different solution than the deposited model — we looked for all models with a corresponding high-resolution crystal structure. For each ligand-protein pair in the EMDB dataset, the PDB was searched for

structures solved by X-ray crystallography at 2.6 Å resolution or better with at least 50% sequence identity to the protein and containing the same ligand. Results from the PDB were filtered further to only contain entries with similar ligand binding pockets as the corresponding EM model. The crystal models were aligned to the EM models by aligning all residues within 10 Å of the ligand using matchmaker in UCSF Chimera [36]. Once aligned, the density correlations of the ligands in the crystal models were calculated in Rosetta. All entries with a pocket-aligned RMSD greater than 1.5 Å and a ligand density correlation lower than 0.1 of the EM model were discarded for being too unalike. This gave 129 ligand-bound EMDB structures with similar crystal models.

We identified a subset of 100 cases where EMERALD converged on a ligand conformation and a corresponding high-resolution crystal structure was available. The converged docked model was within 1 Å RMSD of the ligand modeled in the crystal structure for 67% of cases, while 58% of the deposited EM models were within this distance. Considering cases where the model predicted from EMERALD and the reference EM model differ, there were 6 cases where the EMERALD model was within 1 Å RMSD to the crystal structure while the EM model was not, 3 cases where the EM model was within 1 Å of the crystal structure but the EMERALD model was not, and 8 cases where both models differed from the crystal structure by more than 1 Å. Additionally in 5 of the 6 cases where our model predicts the crystal structure, our ligand model improves density correlation by at least 0.03, compared to the deposited cryoEM model.

We show docked models supported by crystal structures in Fig. 2.5 to highlight the quality of our protocol. These examples include: a) the hippocampal AMPA receptor with the antagonist MPQX [37], where our model makes additional hydrogen bond and π-stacking

interactions with the ligand, matching the crystal structure [38] (Fig. 2.5a); b) NBQX in an AMPA receptor [39], where the ligand is flipped, better matching the density, and making bidentate interactions with a nearby arginine residue (Fig. 2.5b); c) DNMDP bound to the SLFN12-PDE3A complex [40], where small changes better match the crystal structure (Fig. 2.5c); d) an ADP molecule in ClpB disaggregase [41] (Fig. 2.5d), where the phosphate groups recapitulate the crystal structure; and e) a glutamate ligand in the AMPA glutamate receptor [39], which was missing an oxygen atom in the deposited structure; when the full glutamate molecule is docked, the carboxylates are placed in a configuration matching the crystal structure [42] (Fig. 2.5e).

There were three cases where our docking protocol found a ligand different than the crystal structure, while the EM model matched the crystal structure closely. All three cases were different maps of the same system, a folate molecule bound to MERS-CoV [43, 44]. In all 3, the EMERALD model and the crystal structure only differ in the placement of a flexible arm with high B-factors in the crystallographic data (Fig. 2.6) [45]. These results lend more support for EMERALD convergence as a confidence metric, which we used to further find instances of alternate ligand conformations.

2.3.3 Alternate Poses identified by EMERALD

Even without crystal structures for reference, trajectory convergence and improved ligand density fit provide confidence in other docked poses. In the case of an antimicrobial bound multiple transferable resistance (Mtr) pump [46], our protocol converges on an ampicillin molecule that is flipped so that its phenyl group is now in a pocket of unassigned density (Fig. 2.7a, b). While the deposited model places the phenyl group sandwiched between two phenylalanine residues (Fig. 2.7a), our docked model packs the group near a cluster of

hydrophobic residues known to interact with other antibiotics [46] (Fig. 2.7b). Additionally, nearby arginine, serine, and threonine residues have been suggested to generally coordinate ligands binding to the pump [46]; our model has the carboxyl group positioned to make interactions with these residues directly or possibly through bridging water molecules. While it is likely that an antibiotic would bind non-specifically to this site, EMERALD ranks our presented orientation the highest across all three trajectories, and there is a large predicted energy gap (about 10 kcal/mol) between the converged conformation and the best-scoring conformation with the phenyl group outside this hydrophobic pocket, suggesting that this pose is strongly favored by EMERALD.

Another instance of improving density fit and receptor interactions is a lipid phosphatidylinositol 4,5-bisphosphate (PIP_2) bound to transient receptor potential melastatin member 8 (TRPM8) [47]. The EMERALD docked model correlates with the map 10% better than the deposited model, placing all the phosphate groups into density and placing the likely-disordered glycerol backbone and beginning of the lipid tails in weaker density (Fig. 2.7c, d). Moreover, the 4,5 phosphate groups of our docked model make more interactions with basic residues that bind PIP_2 in other structures of TRPM8 [47]. While the starts of the acyl chains are oriented away from the transmembrane region of the protein, this is likely occurring because the chains are truncated. Considering the phosphate placements in the deposited model do not appear in the top 20 lowest-energy models for any trajectory and the reasons above, EMERALD predicts a more accurate model of PIP_2 binding.

Additional cases with confident alternative models are shown in Fig. 2.7e-h. For the ATP analog in a structure of the ATP11C flippase [48] the gamma phosphate sticks out of density in the deposited model (Fig. 2.7e) but is modeled into the density and interacting with a nearby

lysine residue in the docked model (Fig. 2.7f). Finally, our EMERALD model of a small molecule GO52 bound to the CD4-HIV-1 Env SOSIP complex [49] confidently fits the amide and piperidine groups into the map better than the deposited model, while keeping the same hydrophobic interactions as the deposited model (Fig. 2.7g, h).

We next identified cases where: a) the EMERALD model and deposited structure were different, and b) half maps were available in EMDB. For these cases, models were refined into one half map and validated against the other using real-space density correlation. Half maps for the instances with the "non-match, similar or better quality" designation and EMERALD convergence were obtained from the EMDB when available, giving 62 cases. Maps were sharpened with phenix.auto_sharpen [50], and the deposited and docked models were refined into the first half map using a dualspace refinement in Rosetta. The density cross correlations for both ligand refined models were calculated with the first and second half map. If the difference in density correlation with the second map was greater than 0.05 between the two models, the model with the higher correlation was considered better. When comparing the deposited and EMERALD models (Fig. 2.8a), we found 2 instances where EMERALD's model fits the validation map worse (Fig. 2.8b-e), 7 cases where it fits the validation map better (one of which is shown in Fig. 2.7d) and saw equivalent quality for the remaining 53 cases.

2.3.4 Ambiguous and failed cases

While our analysis confidently discovers alternate ligand models, 58% of docked molecules with similar quality to the deposited model have medium or low confidence. We found that small molecules that have pseudo-symmetry or have flexible moieties represent these low-confidence cases because of the challenges they provide from their often noisy and inconclusive density. In some instances, two or more replicates of EMERALD agree on a

substructure of the molecule (dark blue, Fig. 2.9a, b) [51], but differ in a rotamer of a functional group or a flexible group (light blue, Fig. 2.9a, b). For other ligands, ambiguous density leads to little agreement among the reference model and low-energy Rosetta models (Fig. 2.9c, d). The authors for the allosteric modulator of a dopamine receptor note the lack of confidence in the deposited structure [52] but have mutagenesis studies to confirm the conformation modeled (Fig. 2.9c) [53]. However, one model found with EMERALD aligns with their opposing model and fulfills an unexplained region of density in the deposited model (Fig. 2.9d). Altogether, these entries show the difficulty in interpreting cryoEM data at medium to low resolution leading to ambiguous density explanations for a single map, and the limits to automated ligand docking using our protocol.

To learn what improvements could be made to EMERALD in the future, we looked at instances where EMERALD predicts a ligand with worse metrics than the reference model. We found that these cases often had density that is noncontiguous or noisy, leading to incorrect skeletonization. For a ubiquinone-binding electron transport protein [54], the density skeleton only finds density near the head group (Fig. 2.10c). Without a complete skeleton, the initial population struggles to find the deposited conformation, placing the head group exposed to solvent (Fig. 2.10b). In this case, if the 2.63 Å data is instead truncated at 4.0 Å resolution, the density becomes more contiguous, and the skeleton generated by EMERALD matches the ligand conformation much more closely (Fig. 2.10d). With a complete skeleton, the docked model is no longer worse than the deposited model. The head group of the lowest-energy model makes the same hydrogen bond interactions as the deposited model, and the docked model improves density correlation by 0.03 (Fig. 2.10e). This underscores the importance of the initial sampling

step, especially when evaluating ligands with many rotatable bonds and identifies areas for future upgrades in EMERALD.

2.3.5 Ligand Modeling of linoleic acid

To demonstrate our protocol's utility in structure determination, we used EMERALD to create a model for linoleic acid bound in a previously undetermined protein structure. Determining this model manually would be an arduous task considering high flexibility of the ligand (Fig. 2.11a). Despite the difficulty of modeling the suspected ligand, EMERALD predicts a small molecule conformation that fits the density, makes an anchoring electrostatic interaction with a neighboring arginine residue, and introduces little torsional strain throughout the hydrophobic tail (Fig. 2.11b). This placement is supported by the structure of linoleic acid bound to a related protein [55]. Creating the model required no user input once ligand restraint files were made, and the ease and accuracy when modeling linoleic acid prove the value of EMERALD for structure determination.

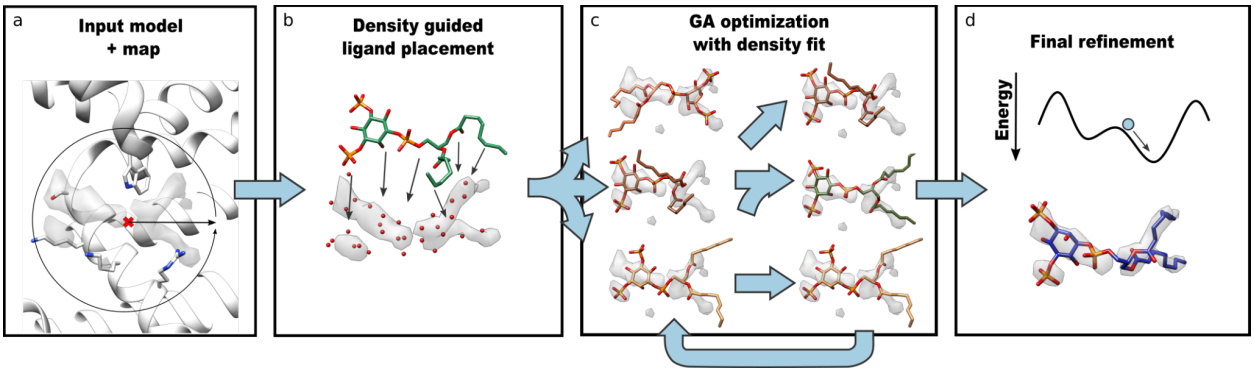

Figure 2.1. Overview of EMERALD docking protocol. (a) The cryoEM map, coordinates of the receptor, and the location of the binding site (red cross) are provided as inputs. The binding pocket is calculated depending on the radius of the ligand (circle) to determine boundaries and side chains to consider when modeling. (b) All unmodeled density in the pocket is converted to a pseudo-atomic skeleton (independent of ligand identity), which is used to generate an initial set of ligand conformers. (c) Using a genetic algorithm, the pool of ligand conformers is optimized against Rosetta energy and density fit. The population of ligand conformers evolve over 10 generations with low energy conformations surviving and combining attributes with each other. (d) The 20 poses with lowest energy are refined in Rosetta.

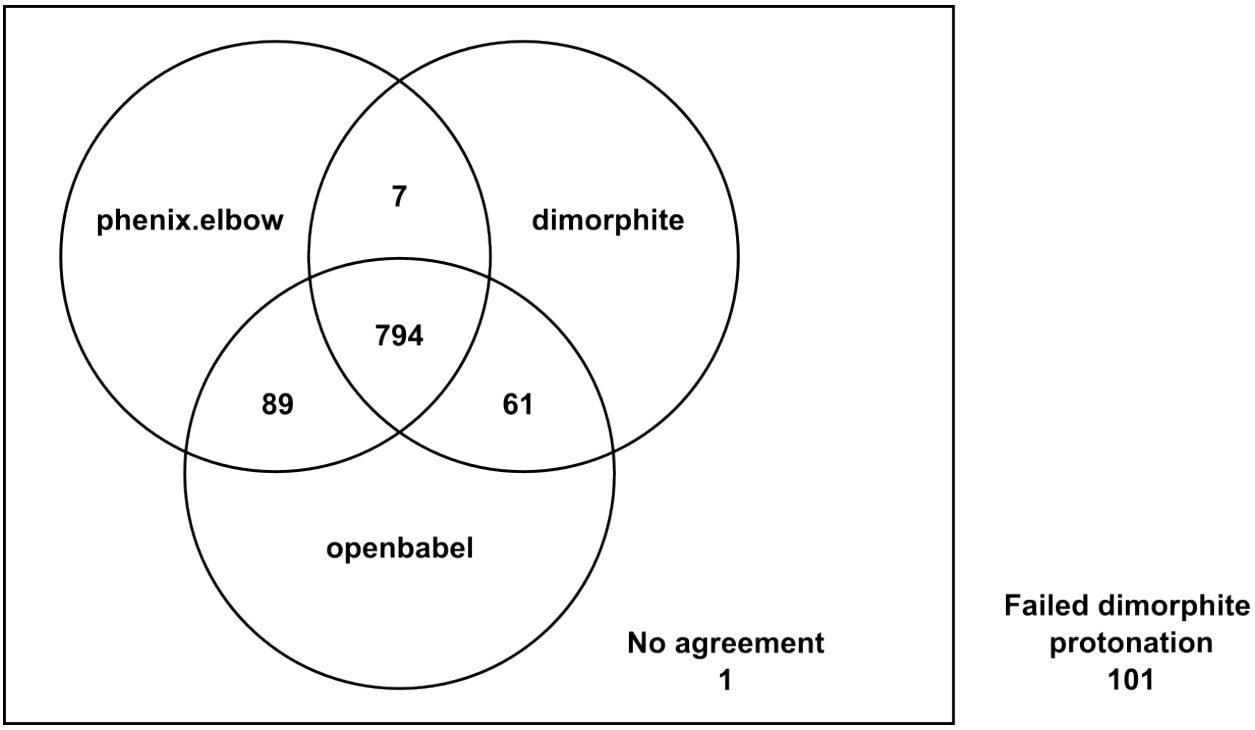

Figure 2.2. Agreement of protonation state assignments. Three protonation assignment methods were used to determine the protonation state for a small molecule in EMERALD. The state assigned by 2 or more methods was used for docking.

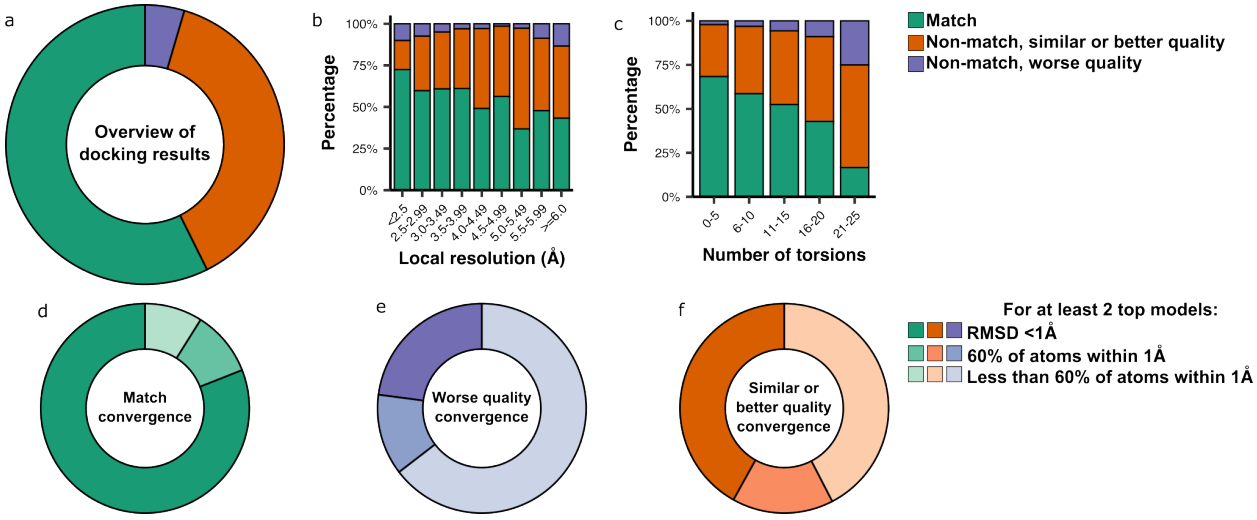

Figure 2.3. Benchmarking EMERALD against the EMDB. (a) A comparison of EMERALD models to the deposited structures over 1053 EMDB-deposited complexes. In total, 58% of EMERALD-docked models were placed within 1 Å RMSD of the deposited ligand ("match", green); 38% were more than 1 Å RMSD of the deposited ligand but had similar or better density correlations and numbers of hydrogen bonds ("similar or better quality", orange), and 4% were more than1 Å RMSD from the deposited ligand and had worse density correlations or number of hydrogen bonds ("worse quality", blue). (b, c) Bins of binding pocket local resolution (b) and number of torsion angles in the small molecule (c) shown as percentage by docking result. (d-f) The convergence of the best ranking models across multiple runs for matches (d), worse quality (e), and similar or better quality (f) cases. The darkest shade had multiple runs converge with all atoms within 1 Å of each other, the middle shade had multiple runs converge with at least 60% of atoms within 1 Å, and the lightest shade had divergent top-scoring models.

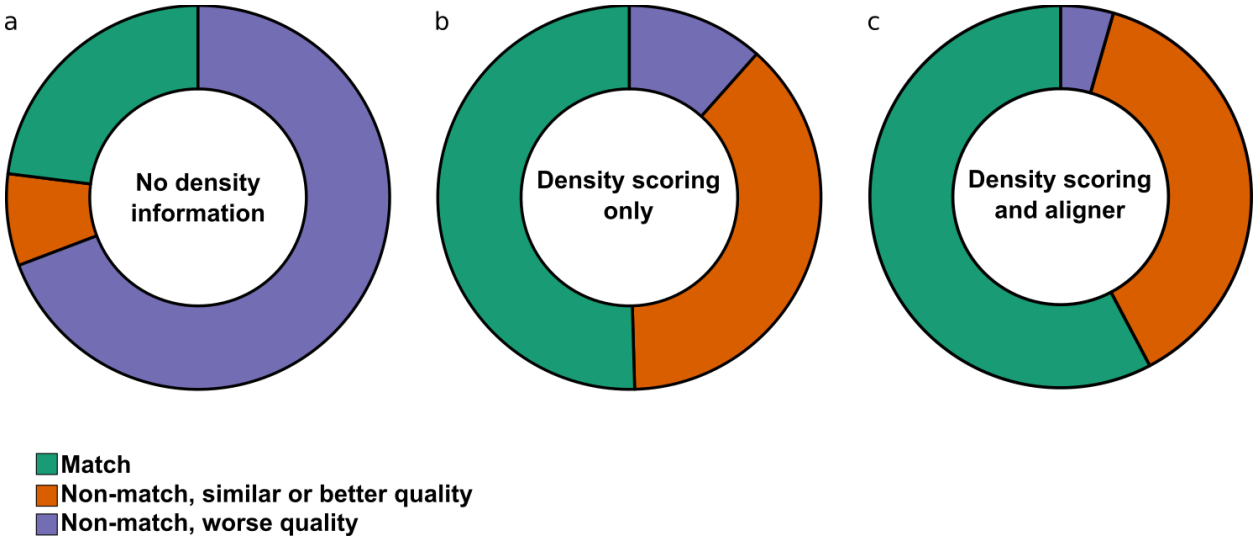

Figure 2.4. Docking results with varying density information. Docking results when docked with no density information in the genetic algorithm (a) and with density scoring evaluation during the genetic algorithm, but not during sampling (b). (a) Without density information, ligand docking can produce a model within 1 Å RMSD to the deposited model for only 23% of cases, and models often do not fit the density, having a worse density correlation or fewer hydrogen bonds for 69% of cases. (b) Adding a density fit score during GA evaluation greatly improves modeling success, getting 50% of cases within 1 Å RMSD to the deposited structure, but is still lower than when density information is included in sampling (c).

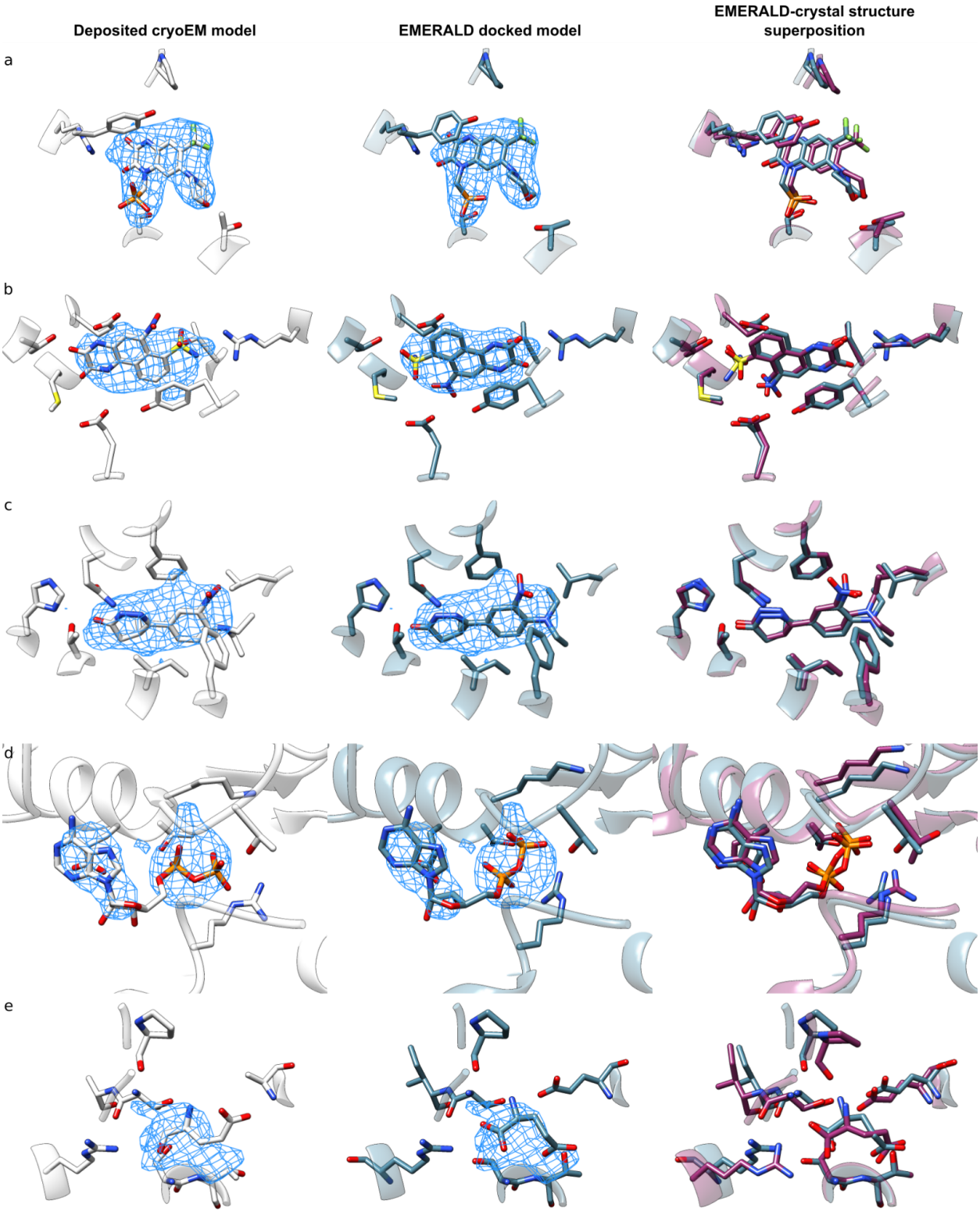

Figure 2.5. EMERALD predicted conformations match high-resolution crystal structures. (a-e) Comparison of the deposited model (left, *white*), EMERALD model (center and right, *blue*), and higher resolution crystal model (right, *purple*). (a) Antagonist ZK200775 in AMPA

receptor (EMDB: 23292, PDB: 7LEP, local resolution: 3.45 Å) and its associated crystal model (PDB: 5ZG2). (b) Molecule NBQX bound to the AMPA receptor (EMDB: 12805, PDB: 7OCE, local resolution: 2.75 Å) and its associated crystal model (PDB: 6FQH). (c) DNMDP bound to the SLFN12-PDE3A complex (EMDB: 23495, PDB: 7LRD, local resolution: 2.95 Å) and its associated crystal model (PDB: 7KWE). (d) ADP bound to ClpB (EMDB: 21553, PDB: 6W6E, local resolution: 4.42 Å) and its associated crystal model (PDB: 5LJ8). (e) Glutamate ligand in an AMPA receptor (EMDB: 12806, PDB: 7OCF, local resolution: 4.26 Å) and its associated crystal model (PDB: 3TKD).

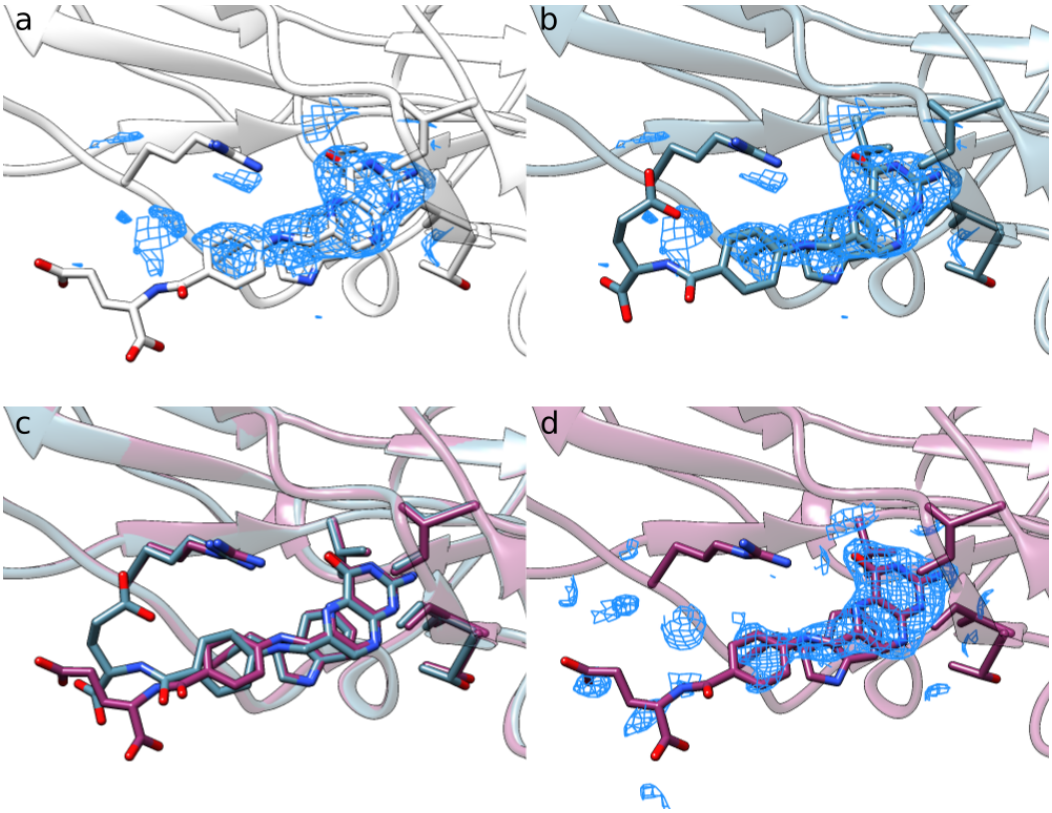

Figure 2.6. Ligand models of folate in MERS-CoV. (a,b) The deposited model (EMDB: 23674, PDB: 7M5E, local resolution: 3.81 Å) (a) and docked model (b) are similar in ordered regions of the ligand with strong cryoEM density and vary in the solvent exposed unordered region. (c) A superposition of the docked model (*blue*) and crystal model (*purple*) (PDB: 5VYH) reveals alignment where the ligand is interacting with the receptor but disagreement in the unordered region. (d) The crystal model shown with its 2mFo-Fc map (contour = 1σ) shows weak density in the solvent exposed region.

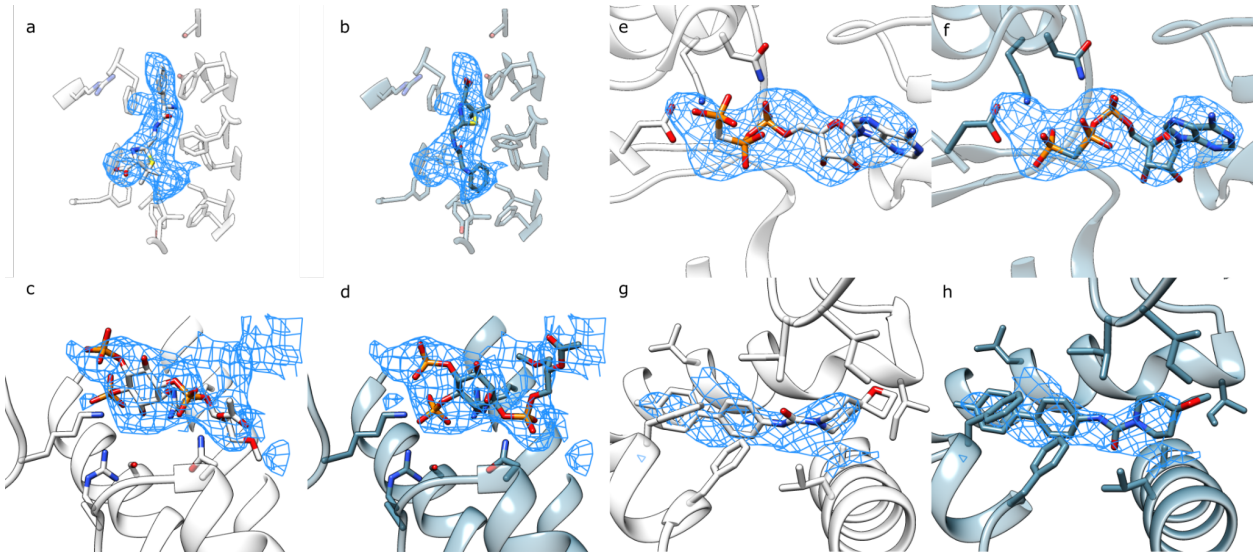

Figure 2.7. Alternate conformations found by EMERALD in cases without crystal structures. (a,b) The deposited structure (a) and EMERALD docked model (b) of Ampicillin bound to Mtr pump (EMDB: 21228, PDB: 6VKS, local resolution: 3.84 Å). (c,d) The deposited structure (c) and EMERALD docked model (d) of PIP_2 bound to the TRMP8 ion channel (EMDB: 0487, PDB: 6NR2, local resolution: 4.07 Å). (e,f) The deposited model (e) and EMERALD model (f) of an ATP analog in flippase ATP11C (EMDB: 30163, PDB: 7BSP, local resolution: 4.06 Å). (g,h) The deposited model (g) and EMERALD model (h) of GO52 bound to the CD4-HIV-1 Env SOSIP complex (EMDB: 22049, PDB: 6X5C, local resolution: 4.11 Å).

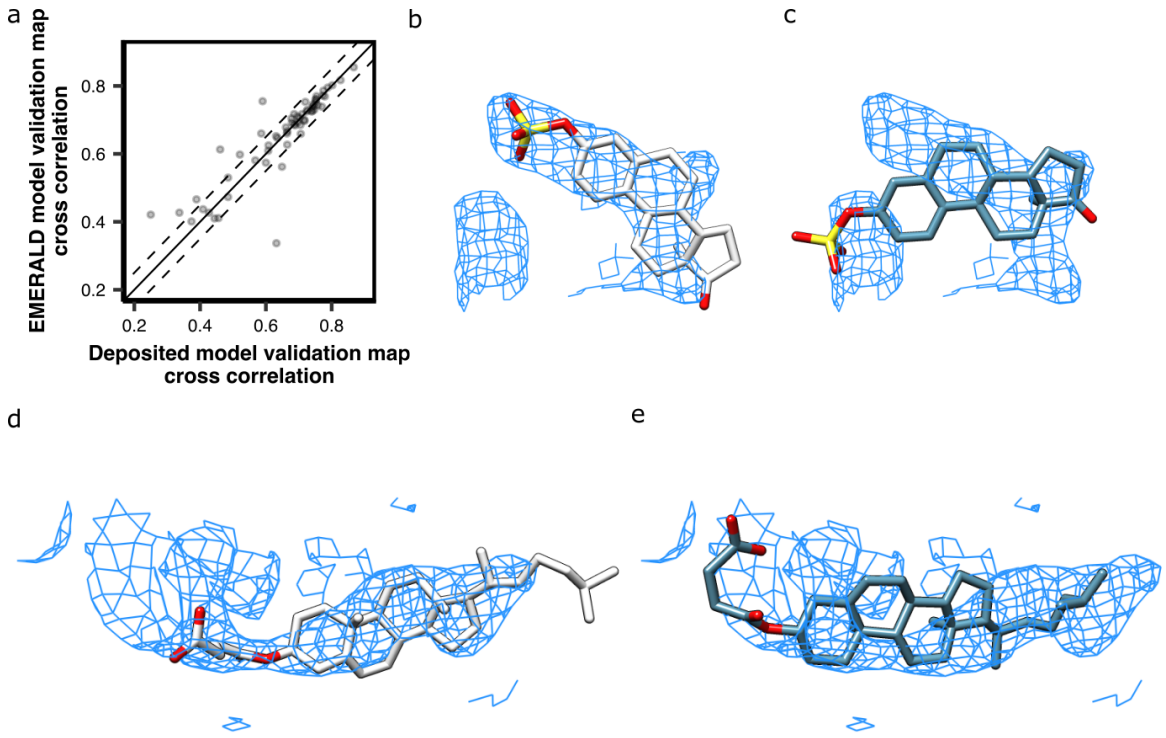

Figure 2.8. Validation of models by calculating map correlation to half maps. (a) Comparison of the map correlation of the deposited model and EMERALD model to a half map. Dashed lines are at +/- 0.05 correlation values above or below equal values. Points above the dashed line are likely cases where the EMERALD model fits better than the deposited (shown in Fig. 2.7c, d). Points below the bottom dashed line are likely cases where the EMERALD model is worse than the deposited model. (b-e) Two cases where the EMERALD model was counted as similar or better than the deposited model but is not by half map analysis. Deposited (b) and EMERALD (c) model of estrone 3-sulfate in ABCG2 (EMDB: 12939, PDB: 7OJ8, local resolution: 3.59 Å). Deposited (d) and EMERALD (e) model of cholesterol hemisuccinate in PfCRT (EMDB: 20806, PDB: 6UKJ, local resolution: 3.21 Å).

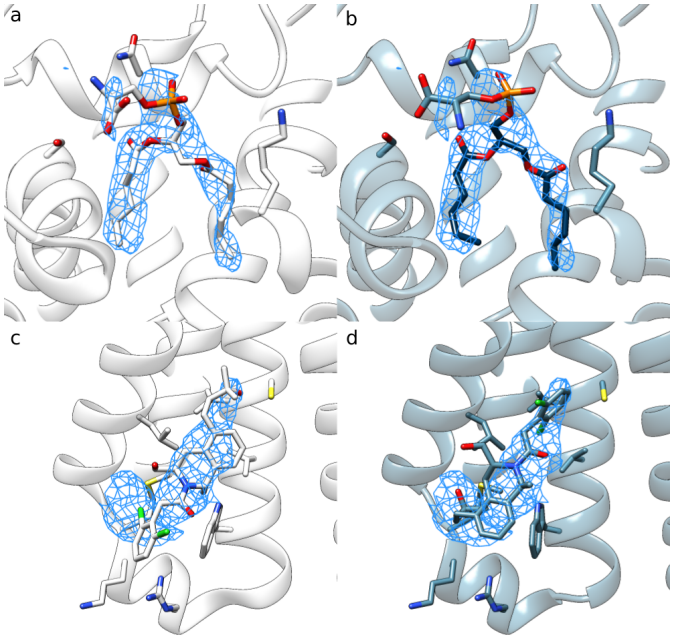

Figure 2.9. Examples of low-confidence docked models. (a) The deposited model of a phosphoserine lipid in ATPase (EMDB: 21844, PDB: 6WLW, local resolution: 2.75 Å) and the EMERALD-docked model (b) place the fatty acid tails in strong density but have differences in the head group. The lowest-energy models for all 3 triplicates find the same lipid tail orientations (*dark blue*). (c) The deposited model of LY3154207 bound to DRD1 (EMDB: 30395, PDB: 7CKZ, local resolution: 3.09 Å) and EMERALD-docked model (d) adopt different conformations. With low-resolution density and few residue binding partners, both models are equally plausible.

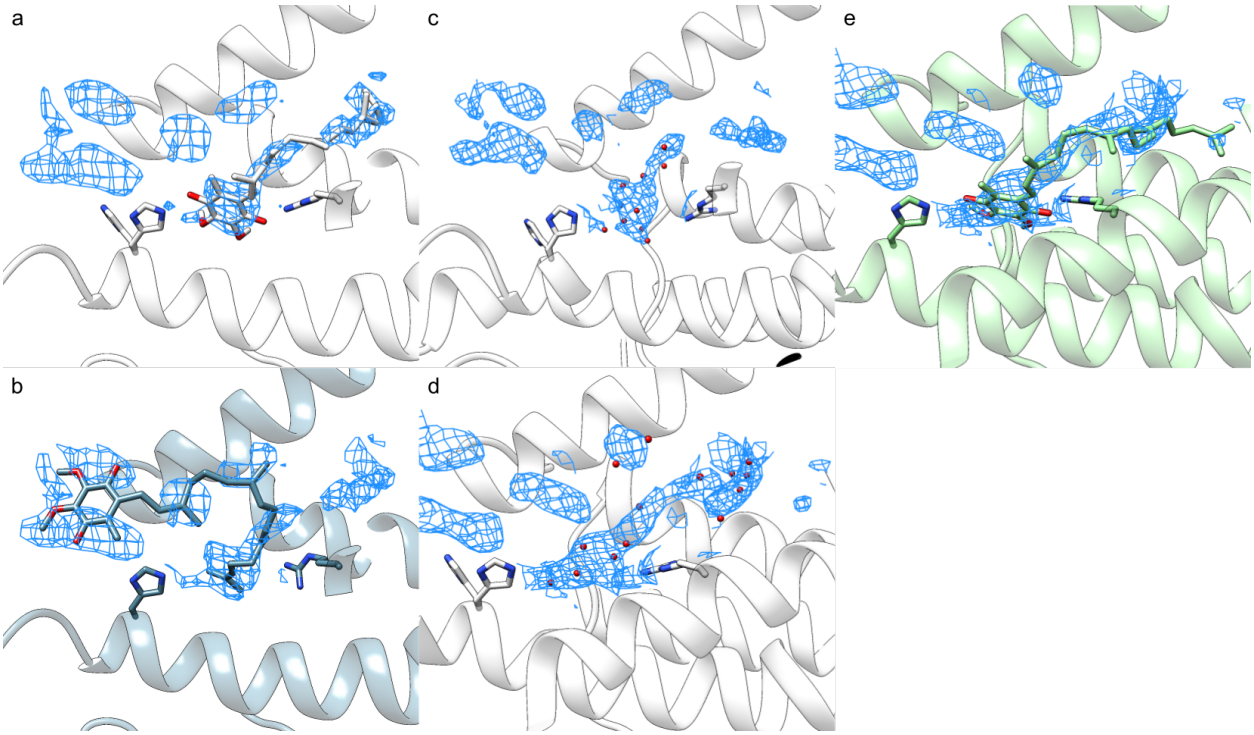

Figure 2.10. Correction of failed docked model with low-pass filter. (a,b) Ubiquinone binding site in cytochrome bo3 in the deposited model (EMDB: 30475, PDB: 7CUW, local resolution: 2.74 Å) (a) and docked model (b). EMERALD cannot find the known binding conformation and places the ligand in noisy density. (c) The density skeleton (red spheres) determined by our erosion protocol only includes density for a small portion of the ligand. (d) When the EM map is low-pass filtered at a 4Å cutoff, the density is more continuous, and the skeleton covers all of the ligand density. (e) The lowest-energy model from EMERALD after the map processing has a similar density correlation and similar hydrogen bonds as the deposited model.

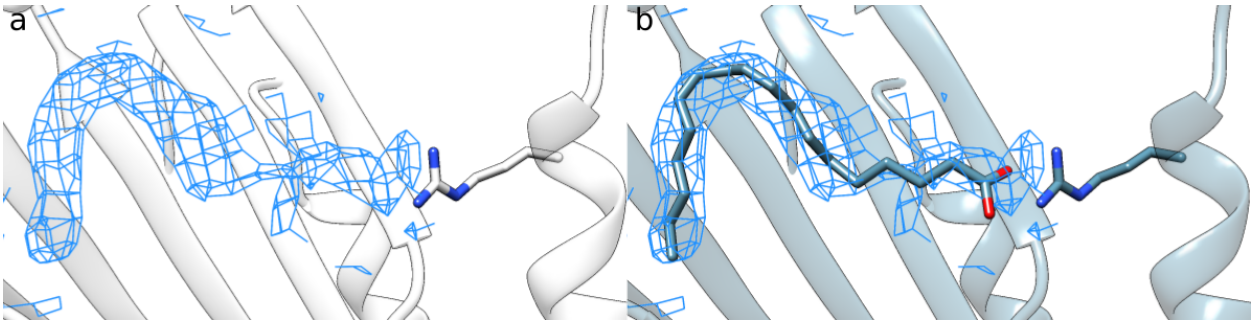

Figure 2.11. Blind modeling of linoleic acid. (a) Unmodeled density for linoleic acid (local resolution: 2.9 Å). (b) Output model from density-guided docking. The model makes an anchoring electrostatic interaction with a nearby arginine residue and models the tail strain-free into the density.

Chapter 3. Modifications of docking protocol for special ligand classes

3.1 Docking ligands with many torsion angles

EMERALD proved to be an effective tool to model small molecules, but there are common instances of small molecules that require special methodology for docking—most notably, lipids. CryoEM has provided an avenue to study protein targets previously troublesome to X-ray crystallography. One class of proteins benefitting from advancements in cryoEM are integral membrane proteins. Integral membrane proteins are embedded within the cell membrane and make interactions with lipid molecules as substrates or as annular lipids at interfaces between the protein and membrane. With increasing ability to study these proteins [56], it is imperative to model the large, flexible lipids bound to membrane proteins. While an initial benchmark of EMERALD accurately modeled ligands with 20 or fewer torsion angles, the failure rate increases to 25% once the number of torsion angles exceeds 20 (Fig. 2.3c), a common number to reach for phospholipids. A modified protocol of EMERALD was developed to accurately handle docking small molecules with 20 or more torsions. To improve EMERALD's performance for docking high-torsion ligands, two aspects of the protocol were modified: skeleton generation and repopulating the pool of ligand conformers during GA optimization.

3.1.1 Dynamic skeleton generation for ligand alignment

As mentioned in Chapter 2.1, alignment of ligand conformations in the initial pool to unmodeled density blobs can drastically reduce the conformation space needed to be searched during the GA. This necessity is only amplified as the number of torsion angles in a small molecule, and thus the possible conformation space, increases. The previously mentioned erosion protocol was limited in three ways. First, voxels initially considered for erosion were limited

within 10 Å of the center of the binding site, and ligands of a certain size exceed this boundary. Second, for high-torsion ligands, it is common for density to be noncontiguous because of low signal from disordered regions. Noncontiguous density blobs were unlikely to be grouped together out of caution of adding noisy regions of density to the skeleton, leading to sections of the ligand unrepresented in the skeleton (Fig. 2.10). Finally, a single skeleton was used for all alignments in the initial population, allowing only one shot at generating a quality skeleton.

All these limitations have been addressed in an updated erosion algorithm for large ligands (Fig. 3.1). All voxels considered for erosion are within the maximum radius of the docked ligand. For every ligand conformation, a skeleton is independently created. Unmodeled regions of the EM map are detected by being above a threshold previously mentioned in Chapter 2.1. Now, instead of using two rounds of erosion to separate noncontiguous regions of density, all blobs of density are separated if more than 1.5 Å apart.

The separated blobs are then categorized into “base” blobs and “satellite” blobs depending on their volume and proximity to the receptor. To approximate blob volume based on ligand size, density maps of ligands from an empirically determined set of lipids were generated using molmap in Chimera. Blob volumes at a threshold of 0.4 were calculated and plotted against the number of heavy atoms in the ligand. The resulting best fit line was used to calculate an expected blob volume shown by Eq. 2.

(Eq.2)

$$Expected\ blob\ volume\ =\ 6.67\ \times\ n_{heavy\ atoms}\ -\ 45.114$$

All blobs of density greater than 60% of the expected size are considered base blobs. If no blob reaches that mark, then the largest blob of density is considered the base. Base blobs are assigned a score determined by the size of the blob

(Eq. 3)

$$score_{size} = \begin{cases} \dfrac{blob\ size}{expected\ size}, & \dfrac{blob\ size}{expected\ size} \leq 1 \\ 2 - \dfrac{blob\ size}{expected\ size}, & \dfrac{blob\ size}{expected\ size} > 1 \end{cases}$$

And the proximity to the receptor

(Eq. 4)

$$score_{proximity} = \begin{cases} 1 - e^{-(blob\ distance-5)^2}, & d_{blob} \leq 5\ \text{Å} \\ 0, & d_{blob} > 5\ \text{Å} \end{cases}$$

where d_{blob} is the blob's average distance to any receptor atom. The size score and proximity score are averaged together to provide a probability that a base blob should be considered. A single base blob is chosen determined by the probabilities.

Satellite blobs are added to base blobs to form a starting point for density erosion. Satellite blobs are scored by the proximity to the receptor like base blobs in Eq. 4. The proximity of the satellite to the base is also considered using Eq. 5

(Eq. 5)

$$satellite\ score_{proximity\ to\ base} = \begin{cases} 1 - e^{\frac{-(d_{base}-7.5)^2}{16}}, & d_{base} \leq 7.5\ \text{Å} \\ 0, & d_{base} > 7.5\ \text{Å} \end{cases}$$

where d_{base} is the satellite blob's average distance to the base blob. Additionally, the current size of the base blob is considered by Eq. 6.

(Eq. 6)

$$satellite\ score_{size} = \begin{cases} 1.2 - \dfrac{blob\ size}{expected\ size}, & 1.2 - \dfrac{blob\ size}{expected\ size} \geq 0 \\ 0, & 1.2 - \dfrac{blob\ size}{expected\ size} < 0 \end{cases}$$

Weights in Eqs. 3-6 were determined empirically, as to create the best skeletons. All three values are multiplied together to generate a probability that the satellite should be added to the base, and

satellite blobs are added to the base accordingly. Once satellite blobs have been added, a skeleton is generated by erosion considering voxels that share a face or edge with one another.

3.1.2 GA optimization with directionality

Along with changes to skeleton generation, repopulation of the pool of ligand conformers was retooled for better performance on large ligands. With the default protocol, conformers can undergo either crossover or mutation when repopulating the pool. During crossover, two parent conformers combine characteristics to generate children, randomly selecting a torsion angle value from either parent for every torsion angle (Fig. 3.1c, bottom left). For mutations, a single parent either has its position in the pocket perturbed or torsion angles are perturbed at random. For ligands with many torsions, the randomness in torsion assignment undermines the previous alignment to density step since consecutive torsion angles should already be assigned so that the ligand fits into the density. An alternate form of crossover and mutation was implemented that considers directionality of torsions when generating children.

With the alternate crossover and mutation method, torsions are represented as a tree data structure, with one torsion acting as a root and all successive torsions being downstream of it (Fig. 3.1c). When performing crossover or mutation, a random torsion is chosen within the tree. For mutations, all downstream torsions of the selected torsion position are randomly mutated. During crossover events, a new conformer is created consisting of the downstream torsions from one parent and the upstream torsions of the other parent (Fig. 3.1c, bottom right). Together, newly generated conformers are likely to keep portions of the ligand already fit into density, while still allowing opportunities to search for lower energy models.

3.1.3 Docking results for ligands with 20 or more torsion angles

The effects of changes to EMERALD on modeling ligands with many torsions were tested on 172 ligand-bound structures. The dataset was collected using the same criteria as described in Chapter 2.2.1, except only cases with 20 or more torsions were included. Analysis of docked models compared to deposited models is as described in Chapter 2.3.1. With the modified EMERALD protocol, 17/172 ligands were within 1 Å RMSD to the deposited ligand (10%, Fig. 3.2b), a modest improvement from 12/172 from default EMERALD. Considering how difficult it is to achieve low-RMSD values with large ligands, a better measurement for improvement is the proportion of cases with docked ligands worse than the deposited model. The updates to the protocol lower the number of failures from 47 to 17.

The strength of the large ligand docking protocol can be seen when docking a phosphatidylserine molecule to ATP11C flippase (Fig. 3.2c-e). Using the default EMERALD protocol, the docked phosphatidylserine molecule was aligned only to the central portion of the ligand density, and sampling could not effectively explore the remaining conformational space. This led to the head group and a tail fitting into noise in the map rather than interacting with the protein (Fig. 3.2d). When modeled using the large ligands update, the initial voxel search for skeletonization allowed density for the head group to be detected, and the noncontiguous density for a tail is discovered. The final model matches the deposited structure, fitting the serine head in a buried pocket and the tails are packed against the surface of the protein (Fig. 3.2e).

The challenge of modeling flexible ligands still leads to 17 cases unable to be modeled as well as the deposited model. Fig. 3.2f shows the binding of a phosphatidylethanolamine molecule to the prokaryotic potassium transport complex. Despite the clear density for all portions of the ligand, the updated EMERALD protocol can only match the tails of the lipid and

places the head group in noise (Fig. 3.2g). Another replicate with the updated protocol overfits the molecule into density and places a lipid tail into detected density likely corresponding to another lipid (Fig. 3.2h). While it is possible that EMERALD could successfully dock the molecule with more replicates, the size of the ligand and possibility of other binding locations for the lipid tail prevent consistent success for this case and others. Nonetheless, a failure rate of 10% is similar to docking smaller ligands and provides confidence that EMERALD can model ligands of any size.

3.2 Docking ligands coordinating metal ions

Many biochemical processes require the transportation or binding of metal ions, and this extends to protein function where metals can play a structural or catalytic role. Around 40% of enzymes with known structures require metal ions for their function [57], necessitating metal prediction methods for any protein-ligand modeling software.

Like modeling small molecules, placement of metal ions becomes increasingly difficult at lower resolutions. Not only will density corresponding to metals be weak or non-distinct, but placements of metal-binding residues are unclear and coordinated water molecules are not represented in the density map at low resolutions. While metal site determination algorithms exist [58], they do not consider density data or energetics, instead focusing solely on metal geometry. Moreover, they are not capable of modeling small molecules and metal ions simultaneously which is required if both locations are unknown.

EMERALD was modified with the ability to simultaneously model small molecules and the metal ions they coordinate (Fig. 3.3). Two additions were required for simultaneous small molecule-ion docking. First, a metal site scorer was implemented to accurately place and score metal ions when new small molecule conformations were created (Fig. 3.3a). Site scores at every

grid point were calculated based on ideal metal coordination geometry and normalized density value at the location. The metal geometry score factors ideal coordination bond length (Eq. 7), coordination bond angles (Eq. 8), and the angle between the metal site, coordinating atom, and base atom (Eq. 9).

(Eq. 7)

$$s_{bond} = \sum_{i=1}^{n} \begin{cases} 0, & \Delta_i \geq 1 \\ 1 - \Delta_i, & \Delta_i < 1 \end{cases}$$

where $\Delta_i = d_{ideal} - d_i$, d_{ideal} is the ideal coordination bond length and d_i is the measured coordination bond distance for each coordinating atom.

(Eq. 8)

$$s_{coordination} = \sum_{i=1}^{n} \sin(\theta_i + (\frac{\pi}{2} - \theta_{ideal}))$$

where θ_{ideal} is the ideal angle for the coordination geometry and θ_i is the measured angle for each pair of coordination atoms.

(Eq. 9)

$$s_{base} = \sum_{i=1}^{n} \cos(|b_{ideal} - b_i|)$$

where b_{ideal} is the base angle for ideal electron orientation and b_i is the measured base angle for each coordinating atom. The total score for the metal site is simply the geometric mean of all 3 scores and a normalized density value, multiplied by a factor determined by the number of potential coordinating atoms at the point (Eq. 10, Eq. 11).

(Eq. 10)

$$site\ score = -1 * \frac{s_{bond}}{n_{atoms}} * \frac{s_{base}}{n_{atoms}} * \frac{s_{coordination}}{n_{angles}} * density_{norm} * x$$

(Eq. 11)

$$x = \begin{cases} 0.25n, & n \leq 6 \\ 7.5 - n, & n > 6 \end{cases}$$

For every ligand pose, the ion is placed at the point on the energy grid with the best metal site score, and the site score is utilized when evaluating the energy of docked poses in the GA. The site scorer is very accurate, as evidenced by its ability to filter iron-binding designed proteins described in [59].

Further optimization of metal coordination sites is provided during the final refinement step of EMERALD. The default small molecule energy function in Rosetta was never optimized to evaluate ion-mitigated interactions and often places metal-coordinating atoms too close to the ion. To ensure proper coordination geometry after minimization, four types of constraints are added during refinement (Fig. 3.3b). Each ion-coordinating atom pair has distance constraints added between them, modeled in Eq. 12. A base angle constraint and dihedral constraint, modeled by Eq. 13 and Eq. 14 respectively, are also added between appropriate atoms in the ion-coordinating atom pair. Lastly, coordination geometry constraints are added among all coordinating atoms-ion-coordinating atom groups (Eq. 15). Small molecules often shift during the refinement stage, so two rounds of refinement are used: the first without constraints, and the second with the metal ion constraints.

(Eq. 12)

$$f(d) = \left(\frac{d - d_{ideal}}{0.5}\right)^2$$

where d_{ideal} is the distance for an ideal coordination bond length.

(Eq. 13)

$$f(b) = \left(\frac{b - b_{ideal}}{0.5}\right)^2$$

where b_{ideal} is the base angle for ideal electron orientation.

(Eq. 14)

$$f(\chi) = \left(\frac{\chi_{nearest} - \chi_{ideal}}{0.5}\right)^2$$

where χ_{ideal} is the dihedral angle for ideal electron orientation and $\chi_{nearest}$ is the nearest periodic value of the angle χ to χ_{ideal}.

(Eq. 15)

$$f(\theta) = \left(\frac{\theta_{nearest} - \theta_{ideal}}{0.1}\right)^2$$

where θ_{ideal} is the angle between coordinating atoms for ideal coordination geometry and $\theta_{nearest}$ is the nearest periodic value of the angle θ to θ_{ideal}.

To benchmark the capabilities of modeling metal-mediated small molecule binding, 45 cases of nucleotide-Mg bound structures were analyzed. Unfortunately, at the time of searching the EMDB for entries with ion-coordinating small molecules on September 03, 2021, the only instances were nucleotide-Mg pairs. Of the 45 cases, 16 produced models with small molecule models within 1 Å RMSD and Mg ions within 1 Å of the deposited model. Cases with poor ion placement often lack side chain residues necessary to produce a strong signal for ion binding.

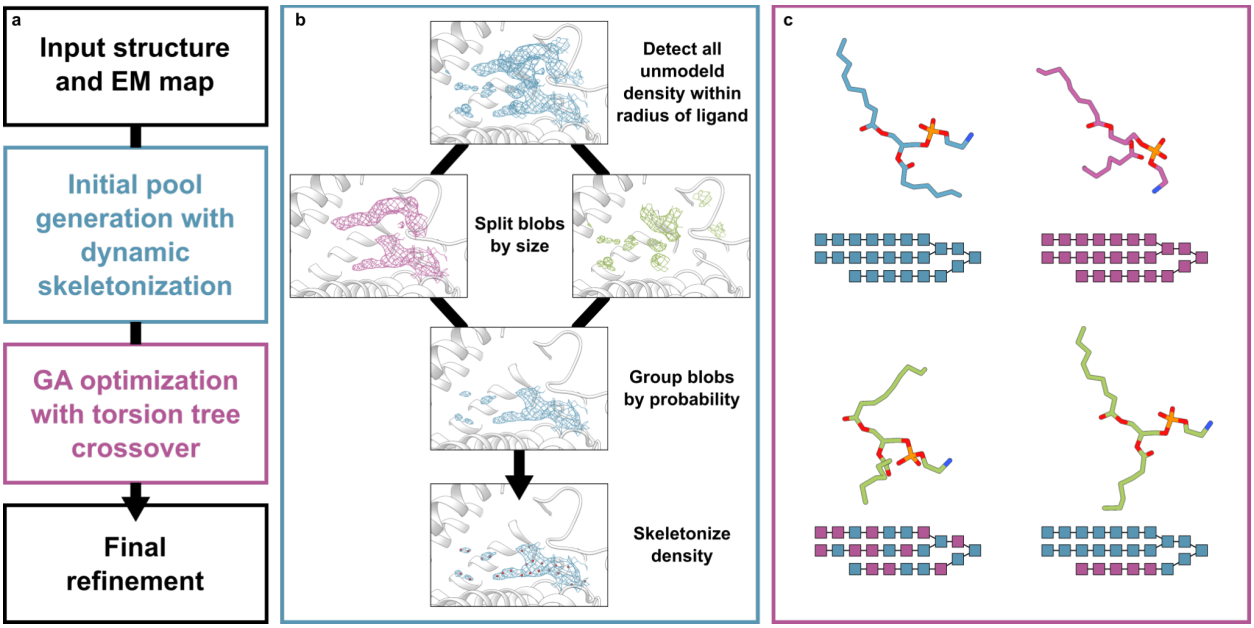

Figure 3.1. Updates to EMERALD for docking high-torsion small molecules. (a) Overview of protocol. Changes were made to skeletonization during initial conformer generation and crossover events during the genetic algorithm. (b) Dynamic skeletonization. A skeleton is created for each initial conformer generated. All unmodeled density within the radius of the ligand is considered for skeletonization. Blobs are split into larger "base" blobs (purple) and smaller "satellite" blobs (green). All blobs receive scores as a proxy for probability which determine the blobs for erosion. (c) Torsion tree crossover. Torsions in the small molecule are represented in a tree structure (squares) with directionality. With default behavior, the torsions from the parents (blue, purple, top) are randomly assigned to the child (green, bottom left). The updated crossover swaps sections of the tree between parents, leading to children similar to one of the parents (green, bottom right).

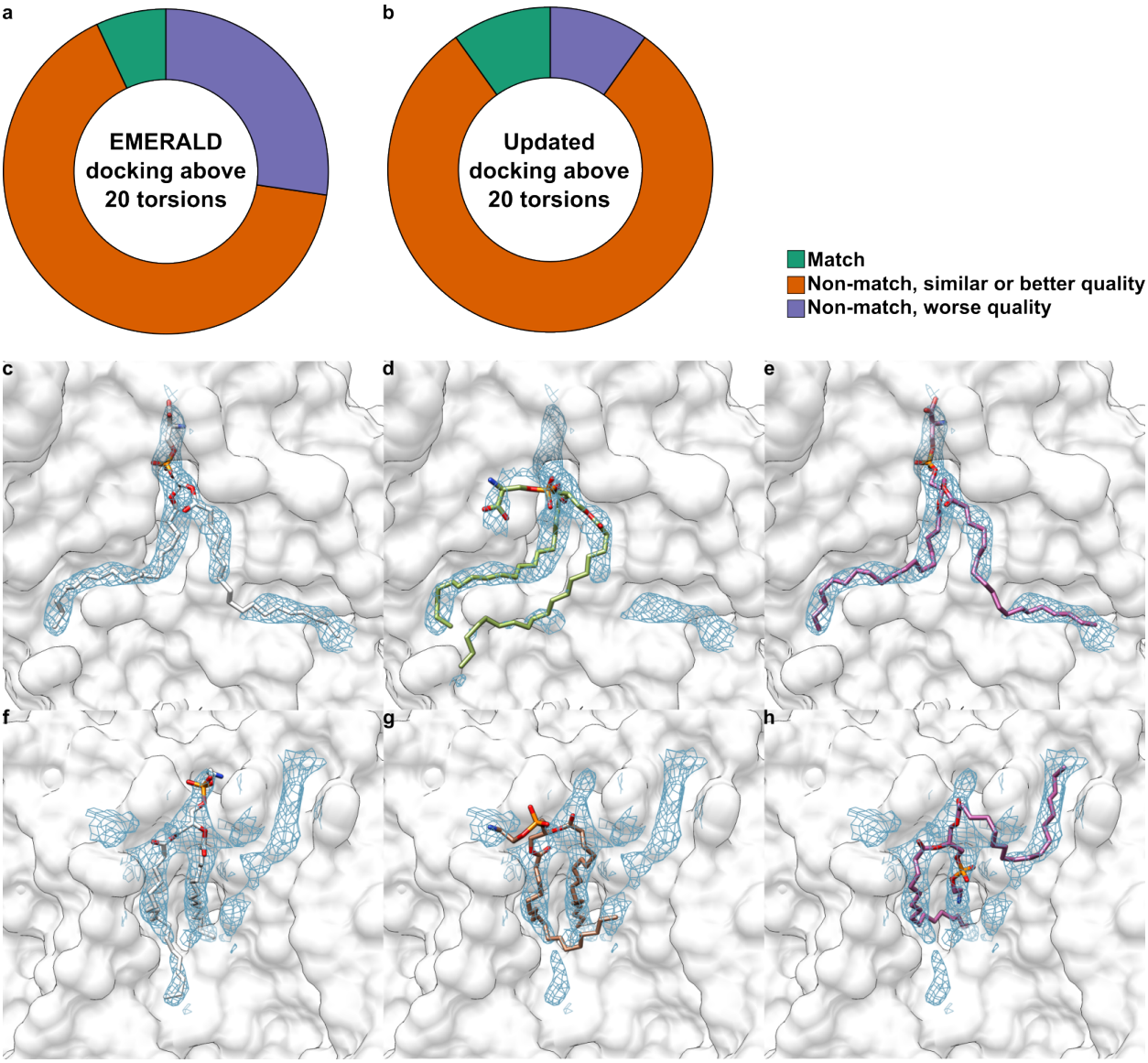

Figure 3.2. Docking results of high-torsion small molecules. (a) A comparison of EMERALD models to the deposited structures of 172 EMDB-deposited complexes with over 20 torsion angles. In total, 7% of EMERALD-docked models were placed within 1 Å RMSD of the deposited ligand ("match", green); 66% were more than 1 Å RMSD of the deposited ligand but had similar or better density correlations and numbers of hydrogen bonds ("similar or better quality", orange), and 27% were more than1 Å RMSD from the deposited ligand and had worse density correlations or number of hydrogen bonds ("worse quality", blue). (b) The same analysis as (a) but using the docking protocol for high-torsion small molecules. 10% "match", 80% are "similar or better quality", and 10% are "worse quality". (c-e) Example of an improvement in docking phosphatidylserine in ATP11C flippase (EMDB: 30168, PDB: 7BSV, local resolution: 3.13 Å). (c) Deposited model; (d) default EMERALD docked model; (e) high-torsion updates docked model. (f-h) Failure of high-torsion docking of phosphatidylethanolamine in KdpFABC

(EMDB: 12184, PDB: 7BGY, local resolution: 3.24 Å). (f) Deposited model; (g,h) high-torsion updates docked models.

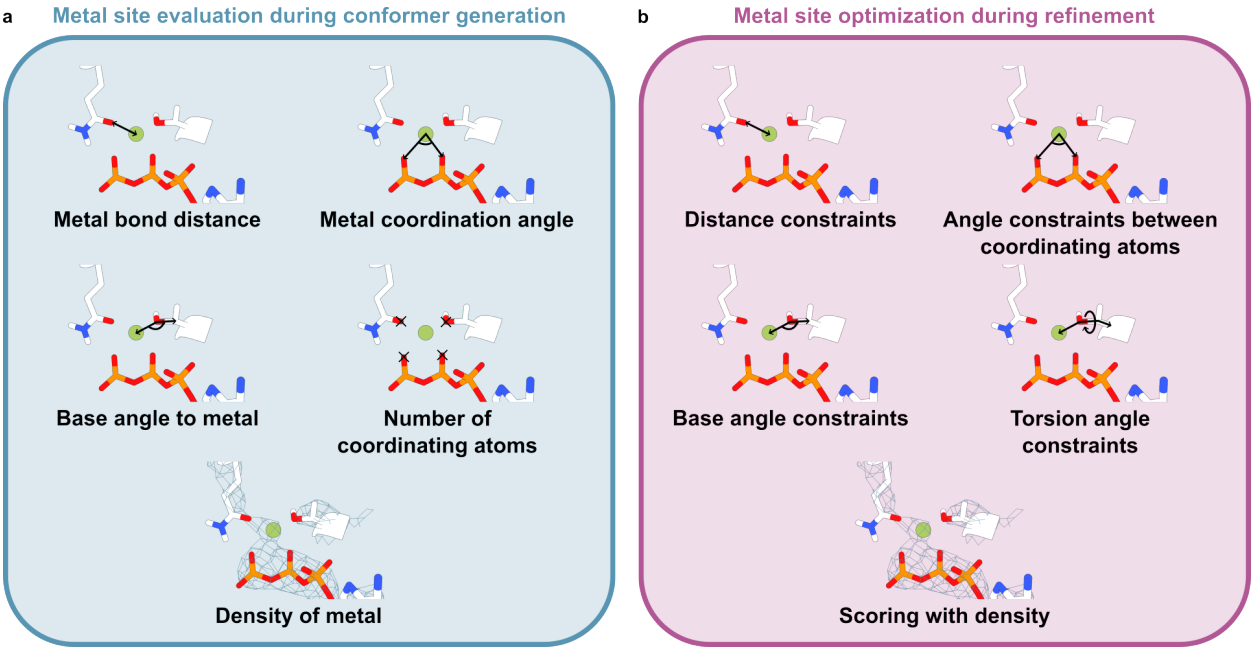

Figure 3.3. Considerations when modeling metal-coordinating small molecules. (a) Factors for metal site scoring. Ideal metal coordination distance and angles, the position of coordinating electrons, the amount of coordinating atoms, and signal in the EM map all contribute to a metal site score. (b) Energy terms added to refinement to ensure proper metal coordination geometry. Rosetta constraints are added between the metal and detected coordinating atoms for proper distance, coordination geometry, and electron position. The density fit is also used during scoring.

Chapter 4. Identifying Ligands in cryoEM data

4.1 Additions to EMERALD for ligand identification

The ability to accurately model small molecules provides a framework to determine ligand identity. A list of candidate ligands can be screened by docking the small molecules in EMERALD and evaluating the best ranking molecule for each identity. Naively, the total Rosetta energy of the model could be used to rank the possible ligand identities. However, the Rosetta energy function is simply a sum of total energy terms [60], so between ligands of varying size the energy function will be biased towards the largest ligand. Thus, it was necessary to establish metrics that appropriately factor ligand size to discern ligand identities.

Once again, a combination of an energy-based score and density-fit score were used to assess the quality of the docked models. In GALigandDock, there exists a binding affinity estimator that calculates the change in free energy upon binding (dG), providing an energetic term. A ligand density correlation can be used as a density-based score, but large ligands may be disproportionately favored if they can fill in the density despite having some atoms sticking out of density. To counteract this, a penalty is applied to the ligand density correlation determined by the number of atoms placed near low-density voxels and their centrality to the ligand. Each atom is assigned a centrality score, c, calculated by 1 plus the second longest path to a terminal atom. Using the mean, μ, and standard deviation, σ, of the density values of all atoms in the small molecule, the penalty applied to the ligand correlation is defined in Eq. 16.

(Eq. 16)

$$penalty = \sum_{i=1}^{n} \begin{cases} 0, & density_i \geq \mu - \sigma \\ \log(2c_i + 1), & density_i < \mu - \sigma \end{cases}$$

With both scoring metrics established, it was necessary to develop a single score to evaluate docked ligand models across different identities. In order to combine the metrics into a single unit, dG and penalized density scores of docked models were converted to Z-scores using expected values from linear regression models. Entries from the EMERALD benchmarking dataset where the EMERALD model and deposited model were within 1 Å RMSD (see Chapter 2.3.1) were used for establishing linear regression models. Since the original benchmarking dataset only included structures with ligands with 25 or fewer torsion angles, entries with 25 or more ligand torsions and a docked model within 1.5 Å RMSD to the deposited model (as described in Chapter 3.1) were also added.

The dG and penalized density values for the deposited ligand model after relaxation were calculated and plotted against features of the ligand-bound entries. Comparisons of dG and penalized density to ligand and map features revealed a small dependency for dG on ligand size (represented as the number of heavy atoms) (Fig. 4.1a) and dependencies on ligand size and local resolution for the penalized density correlation (Figs 4.1b,c). These features were used to fit a linear regression model to predict the binding affinity and penalized density correlation, yielding Eq. 17 and Eq. 18, respectively.

(Eq. 17)

$$dG_{expected} = 5.74 - 0.565n;\ \sigma = 9.296$$

(Eq. 18)

$$density\ correlation_{expected} = 0.846 - 0.0232r\ -\ 0.00249n;\ \sigma = 0.07732$$

where n is the number of heavy atoms and r is the local resolution. The expected value and standard deviation are utilized to calculate a Z-score for a docked model given a particular identity. Once individual Z-scores were determined, they were combined into a single Z-score by averaging them and dividing by $\sqrt{0.5}$. Z-scores are converted into a probability distribution by a softmax

function (Eq. 19) and analyzed by calculating a cross-entropy value to a one-hot encoded probability distribution with the deposited identity as 1 (Eq. 20)

(Eq. 19)

$$S(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_j^n e^{z_j}}$$

where $\boldsymbol{z}$ is the vector of Z-scores and n is the length of the vector.

(Eq. 20)

$$(P^* \mid P) = \sum_i P^*(i) \log P(i)$$

where P^* is the true one-hot encoded probability distribution and P is the predicted probability distribution.

4.2 Identifying common ligands

With the ability to compare docked ligand models across identities, the identification capabilities of EMERALD were benchmarked on entries with common ligand identities, defined as the 180 most common ligands in the PDB as reported in the default library for ligand identification with phenix [20]. All entries from the EMERALD dataset that matched the deposited structure or converged on the same model and had a ligand identity found in the 180 most common ligands were considered. This dataset consisted of 268 cases with 36 unique ligands from the list. Ligands were screened in EMERALD with the deposited ligand identity and 9 random decoys from the other 35 ligands found in the dataset, followed by analysis as described above and shown in Fig. 4.2a.

EMERALD identification ranked the deposited identity first in 138/268 cases (Fig. 4.2b). To analyze how favored the molecule is, cross-entropy values were calculated. Cross-entropy measures the difference between two probability distributions, and for our case, the closer a distribution is to a one-hot encoded distribution, the closer to zero the cross-entropy value will

be. A kernel density plot of cross-entropy values can be seen in Fig. 4.2c. Nearly all the cases (252/268) predict the probability of the deposited identity as higher than a uniform distribution.

Entries where EMERALD identification confidently assigns the correct identity have ligand identities drastically different than most ligands in the default library. For instance, the spermine molecule bound to P5B-ATPase fills a narrow, acidic cavity that is inaccessible and unfavorable to the bulkier, negatively charged nucleotides that make up most of the common identities (Fig. 4.2d), leading to a cross-entropy value close to zero.

Poor performing cases typically have very positive binding affinities, usually caused by the small molecule clashing in a tight binding pocket. For an ADP molecule bound to the p97 ATPase mutant [61], a nearby arginine residue is modeled towards the ADP binding site. This rotamer causes clashing and prevents the beta phosphate in the molecule from fitting into the map and interacting with the Walker loop motif (Fig. 4.2e, left), leading to an estimated binding affinity of 26 kcal/mol. Simply expanding the pocket size during docking detects the arginine as a moveable side chain and rotates the residue away from the binding site (Fig. 4.2e, center) and decreases the dG to 21 kcal/mol. The ADP molecule now adopts a conformation similar to higher resolution structures of the complex, but still experiences high binding affinity, likely from continued clashing with the Walker motif backbone (Fig. 4.2e, right). While EMERALD produced a more favorable pose than the deposited model, the poor binding affinity still leads to a lower probability prediction than smaller molecules that do not clash. Given that EMERALD is incapable of large movements in backbone orientation, cases like this ATPase demonstrate the importance of the starting protein model for ligand identification.

Since the ligand identification can discern identities for ligands with drastically different properties, we wanted to examine EMERALD's ability to distinguish similar identities. A

majority of deposited identities in the dataset are ADP, ATP, or GDP molecules, which vary in either a phosphorylation group or functional groups on the purine ring. Our ligand identification protocol correctly identifies ADP as the ligand by a significant amount over both ATP ($p = 3.059\times10^{-5}$, Wilcoxon signed-rank) and GDP ($p = 1.392\times10^{-12}$, Wilcoxon signed-rank) (Fig. 4.3a). In the case of ADP bound to yeast 26S proteasome [62], the protocol scores ADP better for both binding affinity and density fit (Fig. 4.3b), penalizing GDP for adopting the less favorable *syn* conformation and the guanine ring slightly sticking out of the EM map (Fig. 4.3c).

When comparing the probabilities of docked ATP molecules to decoy ADP molecules, there is no significant difference (Fig. 4.3d), with a distribution similar to the ATP analog phosphoaminophosphonic acid-adenylate ester (ANP). ADP molecules can satisfy ATP density blobs without sticking out of the map and can bind to the same motifs as ATP. Additionally, the sample may contain a mixture of ATP and ADP bound states. All these possibilities are seen at the ATP binding site of the ABC transporter ABCG2. The deposited ATP model places the gamma phosphate near two acidic residues and misses key interactions with lysine and serine residues in a Walker motif (Fig. 4.3e). The docked ATP model remedies these issues (Fig. 4.3f), and the decoy ADP model makes the same interactions with the Walker motif while fitting the EM map better (Fig. 4.3g). Along with a slightly better predicted binding affinity, our prediction model favors ADP bound over ATP. The ABCG2 sample was prepared with both ATP and ADP [63], albeit with ten times more ATP, but our model suggests that the bound structure is likely a mixture of both nucleotides, favoring ADP binding.

4.3 Identifying lipid molecules

With the success of modeling common ligands, we turned our attention to identification of lipid molecules. As previously mentioned, membrane proteins are enticing targets to study

with cryoEM and often have lipid molecules bound as substrates or annular lipids. However, disordered regions of the ligand can cause noncontiguous density that is difficult to assign. Thus, a tool that can identify blobs of density corresponding to bound lipids would be a boon. Entries from the high-torsion dataset from Chapter 2.4.2 that contained phospholipids as the ligand and had a ligand density cross correlation above 0.5 were considered for screening. To focus on the head group identity of the lipids, ligands were separated into 6 generic lipid classes, phosphatidic acid (PA), phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylglycerol (PG), phosphatidylinositol (PI), and phosphatidylserine (PS), and lipid tails were truncated to 9 saturated carbon molecules.

Identification and results analysis were performed as described for the common ligands. The protocol only ranks the correct lipid identity first for 19/70 cases (Fig. 4.4a). Additionally, the distribution of cross-entropy values is narrower for lipids than common ligand targets (Fig. 4.4b), and on average, the correct identity is only slightly favored over a uniform distribution. Like the common ligands, identities that are most unalike are the best identified—with the smallest (PA) and largest (PI) lipids achieving identifications with the lowest cross-entropy value (Fig. 4.4c). Models with PA bound lack density for a head group (Fig. 4.4d) while PI-bound maps have density that cannot be fulfilled by the smaller head groups (Fig. 4.4e).

The cluster of molecules with near-uniform probability distributions reveal ambiguity in maps for lipid molecules. The similarity in shape and chemical properties of PC, PE, PG, and PS molecules makes it difficult to discern their identities. For instance, our model favors structures for both the deposited identity of PS and the decoy PE model for the lipid bound to DGAT1. The amine group on the serine head makes an interaction with a backbone carbonyl (Fig. 4.4f), but the ethanolamine head group fits the density better (Fig. 4.4g), provided predicted probabilities

of 17% and 19% respectively. Given that the modelers may have been unsure of identities when initially depositing structures [64, 56] and how PS and PE molecules often share membrane leaflets [65], it is unlikely for our protocol to favor a single identity and both models offer plausible explanations of the EM data.

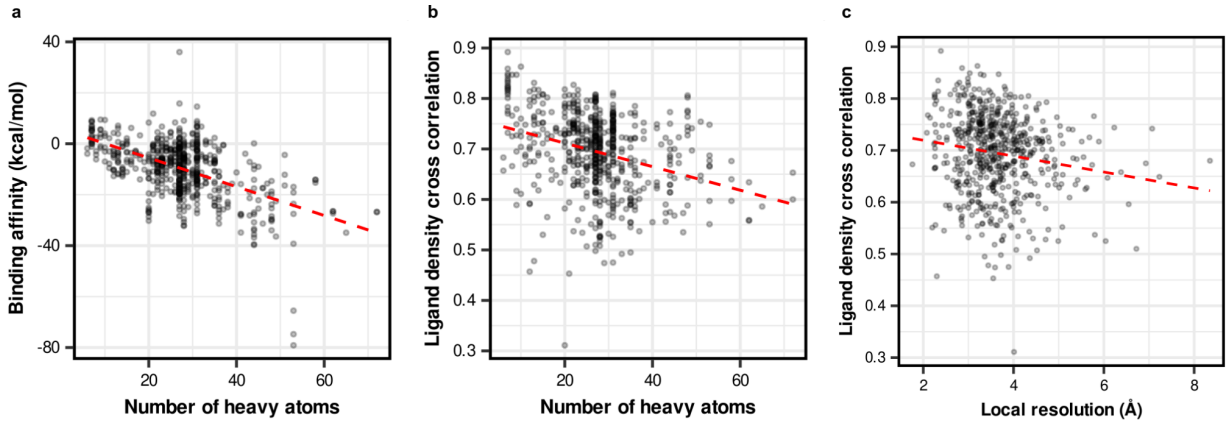

Figure 4.1. EMERALD outputs dependence on input features of deposited EMDB models. (a) Estimated binding affinity dependent on the number of heavy atoms in the small molecule. (b) Ligand cross correlation with the density map dependent on the number of heavy atoms in the small molecule. (c) Ligand cross correlation with the density map dependent on the local resolution of the map in the binding pocket. Red dashed lines represent the best fit line using a linear regression.

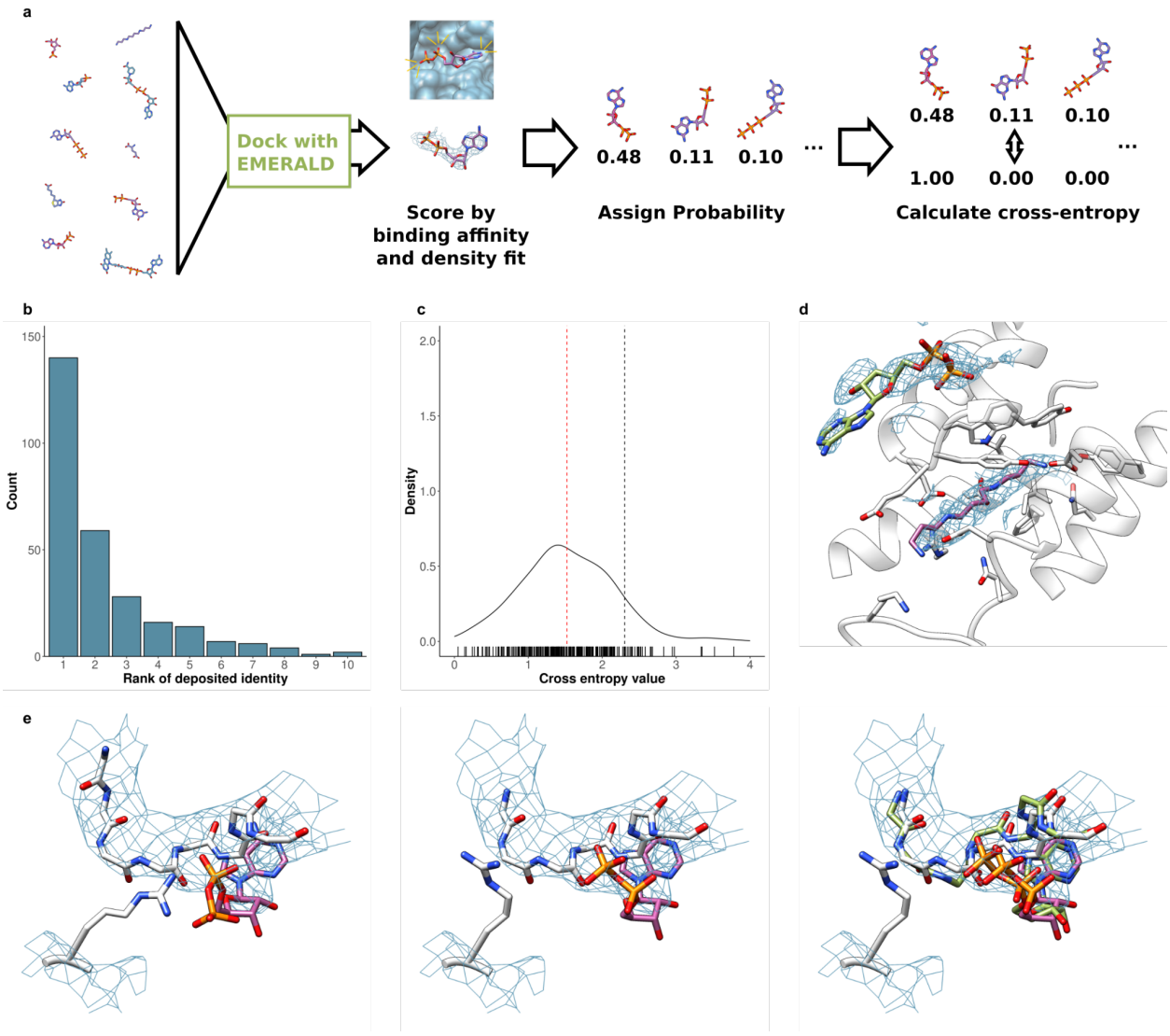

Figure 4.2. Screening common ligand identities with EMERALD. (a) Overview of identification protocol. The deposited ligand identity along with 9 decoys are docked with EMERALD. The estimated binding affinity and map correlation of docked models are used to predict the probability for each identity. The probability distribution is compared to a one-hot encoded identity distribution using cross-entropy. (b) Histogram of the rank of deposited identity after screening. (c) Kernel density distribution of cross-entropy values for each case. Red dashed line marks the mean cross-entropy value for the test set. Black dashed line represents the cross-entropy of a uniform distribution. (d) Successfully screened identity of spermine bound to P5B-ATPase (EMDB: 13012, PDB: 7OP3, local resolution 3.05 Å). Docked model of spermine in purple; Docked model of decoy identity in green. (e) Low cross-entropy case of ADP bound to p97 ATPase (EMDB: 23775, PDB: 7MDM, local resolution 4.93 Å). Left: docked ADP model with clashing arginine; center: docked ADP model after expanding pocket size for arginine to become moveable; right: docked ADP (purple) superimposed with high-resolution crystal structure (green) (PDB: 5DYG).

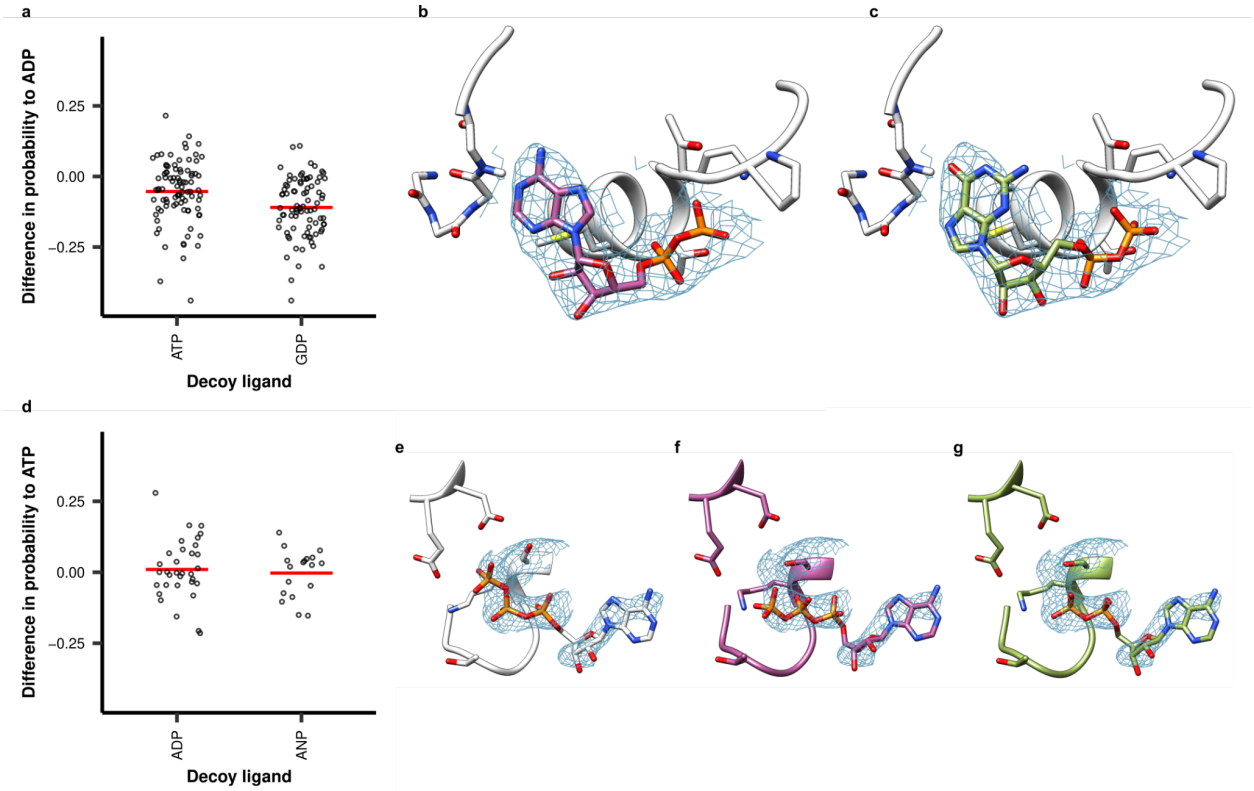

Figure 4.3. Ability of EMERALD identification to distinguish similar identities. (a) Probability of docked decoy ATP and GDP molecules relative to docked ADP molecules. (b, c) Docked models of native ADP molecule (b) and decoy GDP molecule (c) in yeast 26S proteasome (EMDB: 9045, PDB: 6EF3, local resolution: 4.17 Å). (d) Probability of docked decoy ADP and ANP molecules relative to docked ATP molecules. (e-g) Plausible corrected ligand conformation and identity in ABCG2 transporter (EMDB: 12951, PDB: 7OJH, local resolution: 3.27 Å). (e) Deposited conformation of ATP molecule. (f) EMERALD-docked ATP molecule. (g) EMERALD-docked decoy ADP molecule, which has a higher predicted probability than ATP.

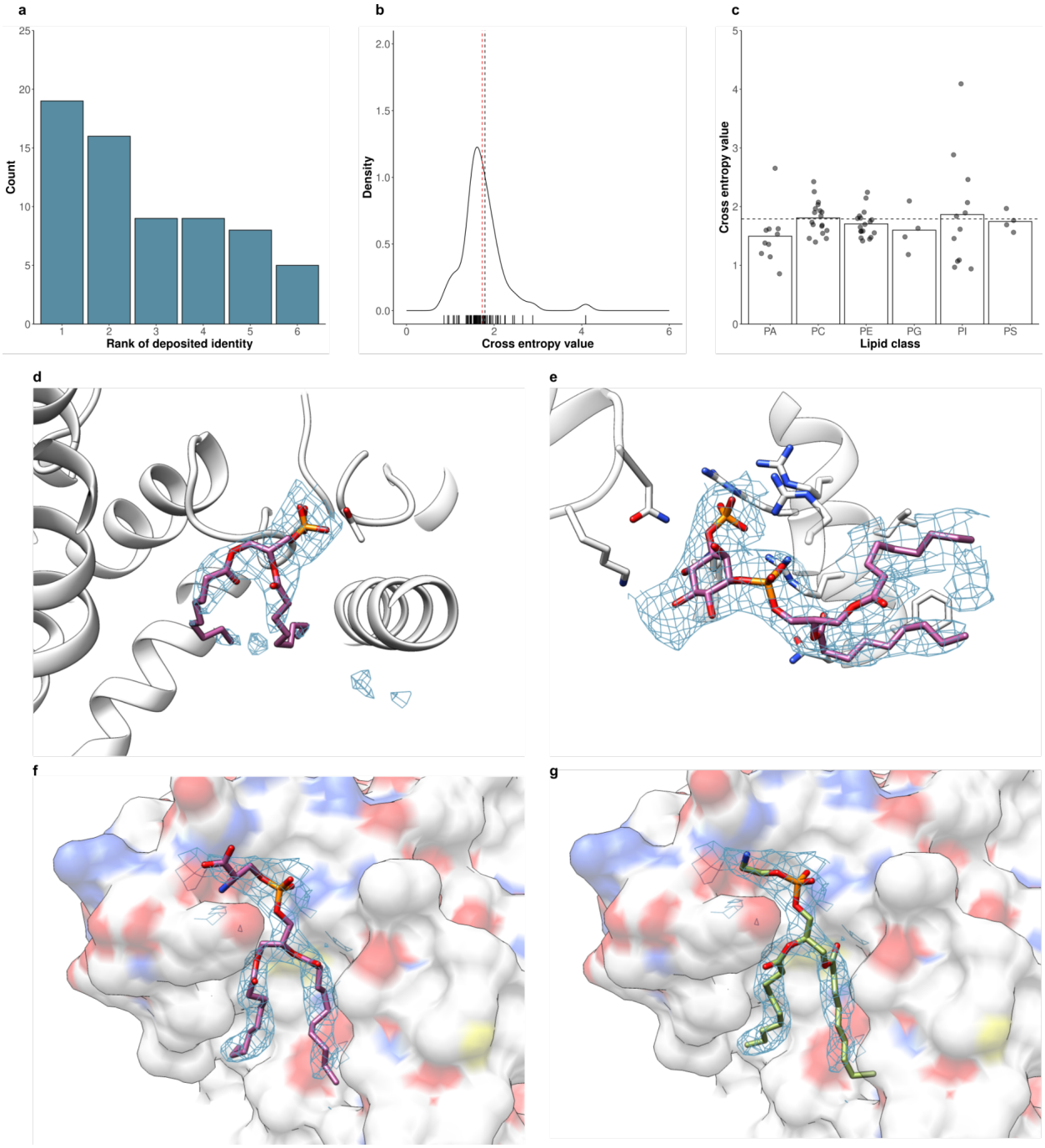

Figure 4.4. Screening phospholipid identity with EMERALD. (a) Histogram of the rank of the deposited identity after screening. (b) Kernel density distribution of cross-entropy values for each case. Red dashed line marks the mean cross-entropy value for the test set. Black dashed line represents the cross-entropy of a uniform distribution. (c) Cross-entropy values by phospholipid class. Black dashed line represents the cross-entropy value of a uniform distribution. PA = phosphatidic acid; PC = phosphatidylcholine; PE = phosphatidylethanolamine; PG = phosphatidylglycerol; PI = phosphatidylinositol; PS = phosphatidylserine. (d) Docked model of PA molecule bound to LAT1-4F2hc amino acid transporter (EMDB: 30837, PDB: 7DSL, local

resolution: 3.25 Å). The lipid density has no space for a head group. (e) Docked model of PI in TPC1 channel (EMDB: 7435, PDB: 6C9A, local resolution: 3.72 Å). The EM map has a clear section for a bulky head group. (f, g) Docked models of deposited identity PS (f) and decoy PE (g) in DGAT1 (EMDB: 21302, PDB: 6VP0, local resolution: 3.63 Å).

Chapter 5. Discussion and Future Directions

Here, multiple ligand modeling tools are presented that aid in structure determination by cryoEM. Ligands can be accurately fitted into EM maps with EMERALD, a density-augmented genetic algorithm docking protocol. EMERALD is capable of modeling ligands of varying types and sizes—routinely matching deposited models for small molecules with 20 or fewer torsion, predicting models with similar quality to deposited models for large ligands, and modeling ion-mediated small molecules with proper coordination geometry. Furthermore, EMERALD was expanded to assign probabilities to potential ligand identities if that is unknown.

While EMERALD is a powerful tool for small molecule determination, it is not without its limitations. The RosettaGenFF force field is not capable of handling all elements, so some common ligands like heme and iron-sulfur clusters cannot be parameterized. Additionally, outside of modeling ions and small molecules, two ligands cannot be modeled simultaneously. Two ligands can be docked successively in EMERALD but having two unmodeled density blobs nearby can lead to poorer performance when docking the first molecule. Docking protocols have been successful in modeling two ligands at once by considering the opposite ligand as a residue [25], providing a possibility to expand EMERALD's capabilities in the future. Several structures solved by cryoEM are electron transport proteins with nearby ligands or contain lipid molecules in proximity, so the lack of simultaneous modeling can be limiting.

The limitations of ligand fitting persist with ligand identification, as the identification method is predicated on EMERALD producing quality docked models. Since each ligand identity needs to be docked with EMERALD to evaluate its likelihood of binding, EMERALD identification can take multiple hours to complete. This makes EMERALD identification unreasonable if a user wanted to screen against hundreds or thousands of common metabolites.

Future updates can be included that can quickly eliminate candidate ligands based on density shape or pharmacophore, followed by docking of the most likely ligand candidates.

Deep learning has been a boon for structure determination and ligand docking with protocols for protein structure prediction [9, 10, 11], cryoEM map processing [66], binding site prediction [67], and even docked pose prediction [68, 69]. With all these uses of deep learning in structural biology, it is tempting to incorporate deep learning to solve ligand fitting and identification problems. For instance, a machine learning model may be used to predict dG or density correlations for ligand identification or produce a logistic model to assign probability to a ligand identity. However, the low amount and lack of diversity of ligand-bound cryoEM structures may not provide enough information to produce meaningful models. Moreover, several errors persist in deposited models, as shown in Chapter 2.3.3, and tools for validating cryoEM ligand models are lacking, meaning several deposited structures may not be trustworthy. Hopefully, with the tools developed here, microscopists can have more confidence in the identity and conformation of their model, causing an increase in the quality of ligand-bond structures solved by cryoEM available for training.

As is, EMERALD offers an automatic tool for ligand modeling that will prove helpful for the now common scenario of ligand-bound structure determination through cryoEM, and EMERALD identification can aid in situations where a modeler does not know what ligand is present. EMERALD-based tools will serve as a valuable addition to the toolkit of Rosetta EM modeling and evaluation methods [5, 70, 71, 72] for structure determination under one software package.

Bibliography

1. Muenks, A., Zepeda, S., Zhou, G., Veesler, D. & DiMaio, F.Automatic and accurate ligand structure determination guided by cryo-electron microscopy maps. *Nat Commun*. 14, 1164 (2023).
2. Yip, K.M., Fischer, N., Paknia, E., Chari, A. & Holger, S. Atomic-resolution protein structure determination by cryo-EM. *Nature* 587, 157–161 (2020).
3. Nakane, T. et al. Single-particle cryo-EM at atomic resolution. *Nature* 587, 152–156 (2020).
4. Merk, A. et al. 1.8 Å resolution structure of β-galactosidase with a 200 kV CRYO ARM electron microscope. *IUCrJ* 7, 639-643 (2020).
5. Wang, R.Y.R. et al. De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nat Methods* 12, 335–338 (2015).
6. Terwilliger, T.C., Adams, P.D., Afonine, P.V. & Sobolev, O.V. Cryo-EM map interpretation and protein model-building using iterative map segmentation. *Protein Science* 29, 87– 99 (2020).
7. Terashi, G. & Kihara, D. De novo main-chain modeling for EM maps using MAINMAST. *Nat Commun*. 9, 1618 (2018).
8. He, J. & Huang, S.Y. Full-length de novo protein structure determination from cryo-EM maps using deep learning. *Bioinformatics* 37, 3480-3490 (2021).
9. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).
10. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871-876 (2021).
11. Chowdhury, R. et al. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology* 40, 1617–1623 (2022).
12. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Cryst.* D60, 2126-2132 (2004).
13. Oldfield, T.J. X-LIGAND: an application for the automated addition of flexible ligands into electron density. *Acta Cryst.* D57, 696-705 (2001).
14. Zwart, P.H., Langer, G.G. & Lamzin, V.S. Modelling bound ligands in protein crystal structures. *Acta Cryst.* D60, 2230-2239 (2004).
15. Terwilliger, T.C., Klei, H., Adams, P.D., Moriarty, N.W. & Cohn, J.D. Automated ligand fitting by core-fragment fitting and extension into density. *Acta Cryst.* D62, 915-922 (2006).
16. Evrard, G.X., Langer, G.G., Perrakis, A. & Lamzin, V.S. Assessment of automatic ligand building in ARP/wARP. *Acta Cryst.* D63, 108-117 (2007).
17. Casañal, A., Lohkamp, B. & Emsley P. Current developments in Coot for macromolecular model building of electron cryo-microscopy and crystallographic data. *Protein Sci.* 29, 1069-1078 (2020).
18. Chojnowski, G., Sobolev, E., Heuser, P. & Lamzin, V.S. The accuracy of protein models automatically built into cryo-EM maps with ARP/wARP. *Acta Cryst.* D77, 142-150 (2021).

19. van Zundert, G.C.P., Moriarty, N.W., Sobolev, O.V., Adams, P.D. & Borrelli, K.W. Macromolecular refinement of X-ray and cryoelectron microscopy structures with Phenix/OPLS3e for improved structure and ligand quality. *Structure* 29, 913-921 (2021).
20. Terwilliger, T.C., Adams, P.D., Moriarty, N.W. & Cohn, J.D. Ligand identification using electron-density map correlations. *Acta Cryst.* D63. 101-107 (2006).
21. Carolan, C.G. & Lamzin, V.S. Automated identification of crystallographic ligands using sparse-density representations. *Acta Cryst.* D70, 1844-1853 (2014).
22. Kowiel, M. et al. Automatic recognition of ligands in electron density by machine learning. *Bioinformatics* 35, 452-461 (2019).
23. Robertson, M.J., van Zundert, G.C.P., Borrelli, K. & Skiniotis, G. GemSpot: A pipeline for robust modeling of ligands into cryo-EM maps. *Structure* 28, 707-716 (2020).
24. Vant, J.W. et al. Flexible fitting of small-molecules into electron microscopy maps using molecular dynamics simulations with neural network potentials. *J. Chem. Inf. Model* 60, 2591-2604 (2020).
25. Sweeney, A., Mulvaney, T. & Topf M. ChemEM: flexible docking of small molecules in Cryo-EM structures using difference maps. *bioRxiv* (2023). doi: https://doi.org/10.1101/2023.03.13.532279
26. Joseph, A.P. et al. Comparing cryo-EM reconstructions and validating atomic model fit using difference maps. *J. Chem. Inf. Model*. 60, 2552-2560 (2020).
27. Park, H., Zhou, G., Baek, M., Baker, D. & DiMaio, F. Force field optimization guided by small molecule crystal lattice data enables consistent sub-angstrom protein–ligand docking. *J. Chem. Theory Comput*. 17, 2000-2010 (2021).
28. Lawson, C.L., et al. EMDataBank unified data resource for 3DEM. *Nucleic Acids Res*. 44, D396-D403 (2016).
29. Greer, J. Three-dimensional pattern recognition: An approach to automated interpretation of electron density maps of proteins. *J. Mol. Bio.* 82, 279-301 (1974).
30. Moriarty, N.W., Grosse-Kunstleve, R.W. & Adams, P.D. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Cryst.* D65, 1074-1080 (2009).
31. O'Boyle, N.M. et al. Open Babel: An open chemical toolbox. *J. Cheminform.* 3, 33 (2011).
32. Ropp, P.J., Kaminsky, J.C., Yablonski, S. & Durrant, J.D. Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules. *J. Cheminform.* 11, 14 (2019).
33. Wang, J., Wang, W., Kollman P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* 25, 247260 (2006).
34. Jakalian, A., Bush, B.L., Jack, B.D. & Bayly, C.I. Fast, efficient generation of high-quality atomic charges. AM1-BCC Model: I. Method. *J. Comp. Chem*. 21, 132-146 (2000).
35. Vilas, J.L. et al. MonoRes: Automatic and accurate estimation of local resolution for electron microscopy maps. Structure 26, 337-344 (2018).

36. Meng, E.C., Pettersen, E.F., Couch, G.S., Huang, C.C. & Ferrin, T.E. Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics* 7, 339 (2006).
37. Yu, J. et al. Hippocampal AMPA receptor assemblies and mechanism of allosteric inhibition. *Nature* 594, 448-453 (2021).
38. Sobolevsky, A.I., Rosconi, M.P. & Gouaux, E. X-ray structure of AMPA-subtype glutamate receptor: symmetry and mechanism. *Nature* 462, 745-756 (2009).
39. Zhang, D., Watson, J.F., Matthews, P.M., Cais, O. & Greger, I.H. Gating and modulation of a hetero-octameric AMPA glutamate receptor. *Nature* 594, 454-458 (2021).
40. Garvie, C.W. et al. Structure of PDE3A-SLFN12 complex reveals requirements for activation of SLFN12 RNase. *Nat. Commun.* 12, 4375 (2021).
41. Yin, Y. et al. Structural basis for aggregate dissolution and refolding by the Mycobacterium tuberculosis ClpB-DnaK bi-chaperone system. *Cell Rep.* 35, 109166 (2021).
42. Krintel, C. et al. Thermodynamics and structural analysis of positive allosteric modulation of the ionotropic glutamate receptor GluA2. *Biochem J.* 441, 173-178 (2012).
43. Park, Y.J. et al. Structures of MERS-CoV spike glycoprotein in complex with sialoside attachment receptors. *Nat. Struct. Mol. Biol.* 26, 1151–1157 (2019).
44. Sauer, M.M. et al. Structural basis for broad coronavirus neutralization. *Nat. Struct. Mol. Biol.* 28, 478–486 (2021).
45. Pallesen, J. et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc. Natl. Acad. Sci. USA* 114, E7348-E7357 (2017).
46. Lyu, M. et al. Cryo-EM structures of a gonococcal multidrug efflux pump illuminate a mechanism of drug recognition and resistance. *mBio* 11, e00996-20 (2020).
47. Yin, Y. et al. Structural basis of cooling agent and lipid sensing by the cold-activated TRPM8 channel. *Science* 363, eaav9334 (2019).
48. Nakanishi, H. et al. Transport cycle of plasma membrane flippase ATP11C by cryo-EM. *Cell Rep.* 32, 108208 (2020).
49. Ozorowski, G., Torres, J.L., Santos-Martins, D., Forli, S. & Ward, A.B. A strain-specific inhibitor of receptor-bound HIV-1 targets a pocket near the fusion peptide. *Cell Rep.* 33, 108428 (2020).
50. Terwilliger, T.C., Sobolev, O.V., Afonine, P.V. & Adams, P.D. Automated map sharpening by maximization of detail and connectivity. *Acta Cryst.* D74, 545-559. (2018).
51. Wang, L., Wu, D., Robinson, C.V., Wu, H. & Fu, T.M. Structures of a complete human V-ATPase reveal mechanisms of its assembly. *Mol. Cell* 80, 501-511 (2020).
52. Xiao, P. et al. Ligand recognition and allosteric regulation of DRD1-Gs signaling complexes. *Cell* 184, 943-956 (2021).
53. Hao, J. et al. Synthesis and pharmacological characterization of 2-(2,6-dichlorophenyl)-1-((1S,3R)-5-(3-hydroxy-3-methylbutyl)-3-(hydroxymethyl)-1-methyl-3,4-dihydroisoquinolin-2(1H)-yl)ethan-1-one (LY3154207), a potent, subtype selective, and orally available positive allosteric modulator of the human dopamine D1 receptor. *J. Med. Chem*. 62, 8711-8732 (2019).

54. Li, J. et al. Cryo-EM structures of *Escherichia coli* cytochrome *bo3* reveal bound phospholipids and ubiquinone-8 in a dynamic substrate binding site. *Proc Natl Acad Sci USA* 118, e2106750118 (2021).
55. Toelzer, C. et al. Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. *Science* 370, 725-730 (2020).
56. Cheng, Y. Membrane protein structural biology in the era of single particle cryo-EM. *Curr Op Struct Bio.* 52, 58-63 (2018).
57. Andreini, C., Bertini, I., Cavallaro, G., Holliday G.L. & Thornton, J.M. Metal ions in biological catalysis: from enzyme databases to general principles. *J Biol Inorg Chem* 13, 1205–1218 (2008).
58. Sciortino, G., Garribba, E., Pedregal, J.R., & Maréchal, J.D. Simple coordination geometry descriptors allow to accurately predict metal-binding sites in proteins. *ACS Omega*. 4, 3726-3731 (2019).
59. Wang, J. et al. Scaffolding protein functional sites using deep learning. *Science* 377, 387-394 (2022).
60. Alford, R.F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput*. 13, 3031-3048 (2017).
61. Zhang. X. et al. Conserved L464 in p97 D1-D2 linker is critical for p97 cofactor regulated ATPase activity. *Biochem. J*. 478, 3185-3204 (2021).
62. de la Peña, A.H., Goodall, E.A., Gates, S.N., Lander, G.C., Martin A. Substrate-engaged 26S proteasome structures reveal mechanisms for ATP-hydrolysis-driven translocation. *Science* 362, eaav0725 (2018).
63. Yu, Q. et al. Structures of ABCG2 under turnover conditions reveal a key step in the drug transport mechanism. *Nat Commun*. 12, 4376 (2021).
64. Wang, L. et al. Structure and mechanism of human diacylglycerol O-acyltransferase-1. *Nature* 581, 329-332 (2020).
65. van Meer, G., Voelker, D.R. & Feigenson, G.W. Membrane lipids: where they are and how they behave. *Nat Rev Mol Cell Biol*. 9, 112-124 (2008).
66. Sanchez-Garcia, R. et al. DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *Commun. Biol*. 4, 874 (2021).
67. Zhao, J., Cao, Y. & Zhang, L. Exploring the computational methods for protein-ligand binding site prediction. *Comp. Struct. Biotech. J*. 18, 417-426 (2020).
68. Wei, L. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems*. (2022).
69. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. DiffDock: Diffusion steps, twists, and turns for molecular docking. *arXiv* (2023). doi: https://arxiv.org/abs/2210.01776v2
70. Frenz, B., Walls, A., Egelman, E., Veesler, D. & DiMaio, F. RosettaES: a sampling strategy enabling automated interpretation of difficult cryo-EM maps. *Nat. Methods* 14, 797–800 (2017).
71. Frenz, B. et al. Automatically fixing errors in glycoprotein structures with Rosetta. *Structure* 27, 134-139 (2019).

72. Reggiano, G., Lugmayr, W., Farrell, D., Marlovits, T. C. & DiMaio, F. Residue-level error detection in cryoelectron microscopy models. *Structure* 31, 860-869 (2023).

Appendix

Examples to run EMERALD, requires version 2023.12 or later of Rosetta

Example xml for EMERALD docking:

```
<ROSETTASCRIPTS>
  <SCOREFXNS>
          <ScoreFunction name="relaxscore" weights="beta_genpot">
                  <Reweight scoretype="elec_dens_fast" weight="100"/>

                  <Reweight scoretype="gen_bonded" weight="1.0"/>
                  <Reweight scoretype="coordinate_constraint" weight="1.0"/>
        </ScoreFunction>
  </SCOREFXNS>

  <MOVERS>
          <SetupForDensityScoring name="setupdens" />
          <LoadDensityMap name="loaddens" mapfile="%%map%%" />
          <GALigandDock name="dock" scorefxn="relaxscore" ngen="10" npool="100"
rmsdthreshold="1.0" smoothing="0.0" ramp_schedule="0.1,1.0" grid_step="0.325"
padding="5.0" nativepdb="%%native%%" sidechains="auto" final_exact_minimize="bbsc"
random_oversample="100" use_pharmacophore="false" skeleton_threshold_const="5.0"
neighborhood_size="7" sample_ring_conformers="1" reference_pool="map"/>
  </MOVERS>
  <PROTOCOLS>
          <Add mover="setupdens"/>
          <Add mover="loaddens"/>
          <Add mover="dock"/>
  </PROTOCOLS>
  <OUTPUT scorefxn="relaxscore"/>
</ROSETTASCRIPTS>
```

Example command line for EMERALD docking

```
$ROSETTA/main/source/bin/rosetta_scripts.linuxgccrelease \
    -in:file:extra_res_fa $ligand_params_file \
    -in:file:overwrite_database_params \
    -gen_potential \
    -database $ROSETTA/main/database \
    -score::gen_bonded_params_file
scoring/score_functions/generic_potential/generic_bonded.round6p.txt \
    -s $input_model \
    -overwrite \
    -multi_cool_annealer 10 \
    -parser:protocol $xml_file \
    -parser:script_vars map=$em_map \
    -atom_mask 2 \
    -sliding_window 1 \
    -edensity::score_sliding_window_context \
    -edensity::mapreso $reso \
    -edensity::grid_spacing 1.0 \
    -no_autogen_cart_improper
```

Examples to run EMERALD for large ligands, requires Rosetta branch amuenks/lipid_dock

Example xml for large ligand docking:

```
<ROSETTASCRIPTS>
  <SCOREFXNS>
          <ScoreFunction name="relaxscore" weights="beta_genpot">
                  <Reweight scoretype="elec_dens_fast" weight="200"/>

                  <Reweight scoretype="gen_bonded" weight="1.0"/>
                  <Reweight scoretype="coordinate_constraint" weight="1.0"/>
        </ScoreFunction>
  </SCOREFXNS>

  <MOVERS>
          <SetupForDensityScoring name="setupdens" />
          <LoadDensityMap name="loaddens" mapfile="%%map%%" />
          <GALigandDock name="dock" scorefxn="relaxscore" ngen="10" npool="100"
rmsdthreshold="1.0" smoothing="0.0" ramp_schedule="0.1,1.0" grid_step="0.325"
padding="5.0" sidechains="auto" final_exact_minimize="bbsc" random_oversample="1"
use_pharmacophore="false" sample_ring_conformers="1" reference_pool="map"
reference_frac="1.0" skeleton_radius="10" method_for_radius="no_padding"
altcrossover="true" pmut="0.5" advanced_map_erosion="true”/>
  </MOVERS>
  <PROTOCOLS>
          <Add mover="setupdens"/>
          <Add mover="loaddens"/>
          <Add mover="dock"/>
  </PROTOCOLS>
  <OUTPUT scorefxn="relaxscore"/>
</ROSETTASCRIPTS>
```

Example command line for large ligand docking

```
$ROSETTA/main/source/bin/rosetta_scripts.linuxgccrelease \
    -in:file:extra_res_fa $ligand_params_file \
    -in:file:override_database_params \
    -gen_potential \
    -database $ROSETTA/main/database \
    -score::gen_bonded_params_file
scoring/score_functions/generic_potential/generic_bonded.round6p.txt \
    -s $input_model \
    -overwrite \
    -multi_cool_annealer 10 \
    -parser:protocol $xml_file \
    -no_autogen_cart_improper \
    -atom_mask 2 \
    -parser:script_vars map=$em_map \
    -sliding_window 1 \
    -edensity::score_sliding_window_context \
    -edensity::mapreso $reso \
    -edensity::grid_spacing 1.0
```

Examples to run EMERALD with metal cofactors, requires Rosetta branch amuenks/metal_dock

Example xml for docking with metal cofactor:

```
<ROSETTASCRIPTS>
  <SCOREFXNS>
          <ScoreFunction name="relaxscore" weights="beta_genpot">
                <Reweight scoretype="elec_dens_fast" weight="100"/>
                <Reweight scoretype="gen_bonded" weight="1.0"/>
                <Reweight scoretype="coordinate_constraint" weight="0.5" />
                <Reweight scoretype="atom_pair_constraint" weight="0.5" />
                <Reweight scoretype="angle_constraint" weight="0.5" />
        </ScoreFunction>
  </SCOREFXNS>

  <MOVERS>
          <SetupForDensityScoring name="setupdens" />
          <LoadDensityMap name="loaddens" mapfile="%%map%%" />
          <GALigandDock name="dock" scorefxn="relaxscore" ngen="10" npool="50"
smoothing="0.0" ramp_schedule="0.1,1.0" grid_step="0.325" padding="5.0"
sidechains="auto" final_exact_minimize="sc" random_oversample="100"
use_pharmacophore="false" skeleton_threshold_const="5.0" neighborhood_size="7"
pmut="0.2" reference_frac="0.5" reference_oversample="2" reference_pool="map"
metal_only="false" metal="MG" double_minimization="1"/>
  </MOVERS>
  <PROTOCOLS>
          <Add mover="setupdens"/>
          <Add mover="loaddens"/>
          <Add mover="dock"/>
  </PROTOCOLS>
  <OUTPUT scorefxn="relaxscore"/>
</ROSETTASCRIPTS>
```

Example command line for docking with metal cofactor

```
$ROSETTA/main/source/bin/rosetta_scripts.linuxgccrelease \
    -in:file:extra_res_fa $ligand_params_file \
    -in:file:override_database_params \
    -gen_potential \
    -database $ROSETTA/main/database \
    -score::gen_bonded_params_file
scoring/score_functions/generic_potential/generic_bonded.round6p.txt \
    -s $input_model \
    -overwrite \
    -multi_cool_annealer 10 \
    -parser:protocol $xml_file \
    -no_autogen_cart_improper \
    -atom_mask 2 \
    -parser:script_vars map=$em_map \
    -sliding_window 1 \
    -edensity::score_sliding_window_context \
    -edensity::mapreso $reso \
    -edensity::grid_spacing 1.0
```

Examples to run ligand identification, requires Rosetta branch amuenks/lipid_dock

Example xml for ligand identification:

```
<ROSETTASCRIPTS>
  <SCOREFXNS>
          <ScoreFunction name="relaxscore" weights="beta_genpot">
                  <Reweight scoretype="elec_dens_fast" weight="100"/>
                  <Reweight scoretype="gen_bonded" weight="1.0"/>
                  <Reweight scoretype="coordinate_constraint" weight="1.0"/>
        </ScoreFunction>
  </SCOREFXNS>

  <MOVERS>
          <SetupForDensityScoring name="setupdens" />
          <LoadDensityMap name="loaddens" mapfile="%%map%%" />
          <GALigandDock name="dock" scorefxn="relaxscore" ngen="10" npool="100"
rmsdthreshold="1.0" smoothing="0.0" ramp_schedule="0.1,1.0" grid_step="0.325"
padding="5.0" sidechains="auto" final_exact_minimize="bbsc" random_oversample="100"
use_pharmacophore="false" sample_ring_conformers="1" reference_pool="map"
reference_frac="0.5" method_for_radius="no_padding" estimate_dG="true"
multiple_ligands="%%ligands%%"/>
  </MOVERS>
  <PROTOCOLS>
          <Add mover="setupdens"/>
          <Add mover="loaddens"/>
          <Add mover="dock"/>
  </PROTOCOLS>
  <OUTPUT scorefxn="relaxscore"/>
</ROSETTASCRIPTS>
```

Example command line for large ligand docking

```
$ROSETTA/main/source/bin/rosetta_scripts.linuxgccrelease \
    -in:file:extra_res_fa $ligand_params_file1 \
    -in:file:extra_res_fa $ligand_params_file2 \
    -in:file:extra_res_fa $ligand_params_file3 \
    -in:file:override_database_params \
    -gen_potential \
    -database $ROSETTA/main/database \
    -score::gen_bonded_params_file
scoring/score_functions/generic_potential/generic_bonded.round6p.txt \
    -s $input_model \
    -overwrite \
    -multi_cool_annealer 10 \
    -parser:protocol $xml_file \
    -no_autogen_cart_improper \
    -atom_mask 2 \
    -parser:script_vars map=$em_map \
    -parser:script_vars ligands=”LG1,LG2,LG3” \
    -sliding_window 1 \
    -edensity::score_sliding_window_context \
    -edensity::mapreso $reso \
    -edensity::grid_spacing 1.0
```