biomolecules

MDPI

Article

Improving Protein–Ligand Interaction Modeling with cryo-EM Data, Templates, and Deep Learning in 2021 Ligand Model Challenge

Nabin Giri and Jianlin Cheng *

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA
* Correspondence: chengji@missouri.edu

Abstract: Elucidating protein–ligand interaction is crucial for studying the function of proteins and compounds in an organism and critical for drug discovery and design. The problem of protein–ligand interaction is traditionally tackled by molecular docking and simulation, which is based on physical forces and statistical potentials and cannot effectively leverage cryo-EM data and existing protein structural information in the protein–ligand modeling process. In this work, we developed a deep learning bioinformatics pipeline (DeepProLigand) to predict protein–ligand interactions from cryo-EM density maps of proteins and ligands. DeepProLigand first uses a deep learning method to predict the structure of proteins from cryo-EM maps, which is averaged with a reference (template) structure of the proteins to produce a combined structure to add ligands. The ligands are then identified and added into the structure to generate a protein–ligand complex structure, which is further refined. The method based on the deep learning prediction and template-based modeling was blindly tested in the 2021 EMDataResource Ligand Challenge and was ranked first in fitting ligands to cryo-EM density maps. These results demonstrate that the deep learning bioinformatics approach is a promising direction for modeling protein–ligand interactions on cryo-EM data using prior structural information.

Keywords: ligand challenge; cryo-EM; protein–ligand interaction; bioinformatics; machine learning; deep learning

Citation: Giri, N.; Cheng, J. Improving Protein–Ligand Interaction Modeling with cryo-EM Data, Templates, and Deep Learning in 2021 Ligand Model Challenge. *Biomolecules* **2023**, *13*, 132. https://doi.org/10.3390/biom13010132

Academic Editor: Ugo Bastolla

Received: 22 November 2022
Revised: 4 January 2023
Accepted: 6 January 2023
Published: 9 January 2023

1. Introduction

Proteins are a building block of life and carry out many vital biological functions. Whether acting as an enzyme to accelerate the chemical reactions, or as regulatory molecules binding to other molecules to activate their functions, the detailed characterization of proteins and their interaction with their binding partners (e.g., the natural substrates or drugs as ligands) is of great importance. Protein–ligand interactions are necessary requirements for signal transduction, immune responses, and gene regulation in living organisms. The study of protein–ligand interactions is important in understanding the mechanisms of biological regulation and provides a theoretical basis for the design and discovery of new drugs. A fundamental objective of computational structural biology is to understand and model such molecular interactions of living systems in sufficient detail so that the behavior of the system can be predicted or modified as desired. In order to characterize the thermodynamic and kinetic behavior of components and their interactions of living organisms, an image of interacting molecules, such as protein–ligand complexes, at near atomic resolution is required to analyze and understand the physical and geometrical constrains of the molecules.

Cryo-EM, an acronym for the cryogenic electron microscopy technique [1], is a revolutionary technology that enables the determination of a 3D structure of macro-molecular complexes at atomic resolution. With the development of various techniques in the cryo-EM realm to generate high resolution maps, as seen in Figure 1, EMDataResource [2] has seen

a surge in the deposition of cryo-EM derived protein density maps which elucidate the protein and ligand interactions in the molecules. The EMDataResource 2021 Ligand Model Challenge [3] was hosted to rigorously benchmark the current methods for generating models using cryo-EM density maps to improve the prediction and validation of protein and ligand interactions, and to identify the metrics which are most suitable for comparing the fit of atomic coordinate models into the cryo-EM maps.

One of the most popular approaches to modeling the protein–ligand complexes is the molecular docking [4–9], which uses physics- or statistical potential-based molecular simulations to generate protein–ligand complex models and a scoring function for estimation of their binding affinities to rank them. With the recent advancement in the field of deep learning, another most prominent approach to modeling protein–ligand complexes is deep learning-based methods. Deep learning-based methods predict protein–ligand binding sites [10–13] using various neural network architectures such as convolution neural networks (CNN), long short-term memory networks (LSTM), and residual networks (ResNet). These methods primarily use three databases: BioLiP [14], ATPBind [15] and Sc-PDB [16] to train and validate their deep learning models before making binding-site predictions. Similarly, deep learning architectures, such as CNNs, graph neural networks, and attention mechanisms, are used for the prediction of protein–ligand binding affinity [17–29]. These methods mainly make use of two databases: PDBbind [30] and the CASF databases [31] for binding affinity predictions. More advanced methods such as Equibind [32]—an SE(3)-equivariant geometric deep learning model for direct-shot prediction of receptor binding location and ligand's bound pose—and DIFFDOCK [33]—a diffusion generative model tailored to the task of molecular docking—have been developed recently. However, even with significant research efforts, despite some success, the protein–ligand interaction prediction problem still remains unsolved because existing methods cannot leverage vast structural data effectively.

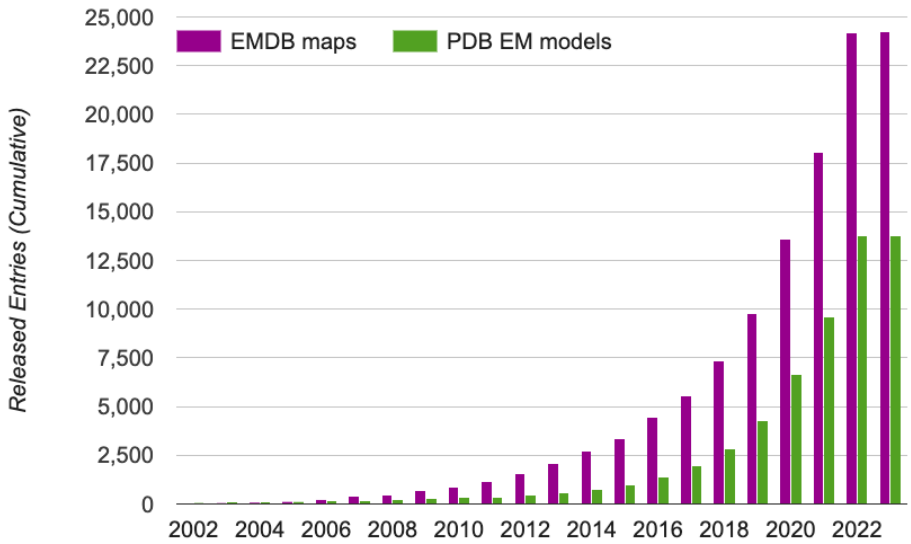


Figure 1. The growth of cryo-EM density maps and cryo-EM-derived protein structures. The statistics were obtained from EMDataResource [34], a unified data resource for 3-Dimension electron microscopy (3DEM) on 20 November 2022.

Inspired by the success of AlphaFold [35], which uses a novel deep learning architecture to predict the protein structures using amino acid sequence data as an input, as well as the various deep learning-based protein–ligand modeling techniques, we adapted the deep learning-based approach for our work. In this work, we combined the deep learning-based protein structure prediction tool DeepTracer with template-based protein–ligand interaction prediction in order to determine the structures of protein–ligand complexes for the 2021 Ligand Model Challenge that was held from 1 February to 1 April 2021. Based on the official

results provided by the assessors of the challenge, our method performed best in fitting ligands to cryo-EM maps (measured across all targets), demonstrating the unique value of the novel deep learning bioinformatics approach for modeling protein–ligand interaction.

2. Materials and Methods

We attempted to solve the problem of protein–ligand interaction by using a set of bioinformatics methods, incorporating cryo-EM data and known structural information such as reference protein structures. In particular, we leveraged the recent advance of applying deep learning to directly predict the structure of proteins from high-resolution cryo-EM density maps; a succinct review of the methods can be found in Ref. [36]. To predict the bound conformation (3D atomic structure) of a protein–ligand complex, we utilized an existing deep learning-based tool as a key component of our model building pipeline (DeepProLigand). DeepProLigand predicts the 3D coordinates of protein structures using only a cryo-EM density map as an input. This protein structure model is a starting point for the downstream ligand positioning and model refinement tasks. The workflow illustrated in Figure 2 demonstrates our approach to generating the structure of a protein complex by incorporating a fully automatic deep learning-based method as its primary building block. The modeling pipeline of Figure 2 has three key steps described as follows:

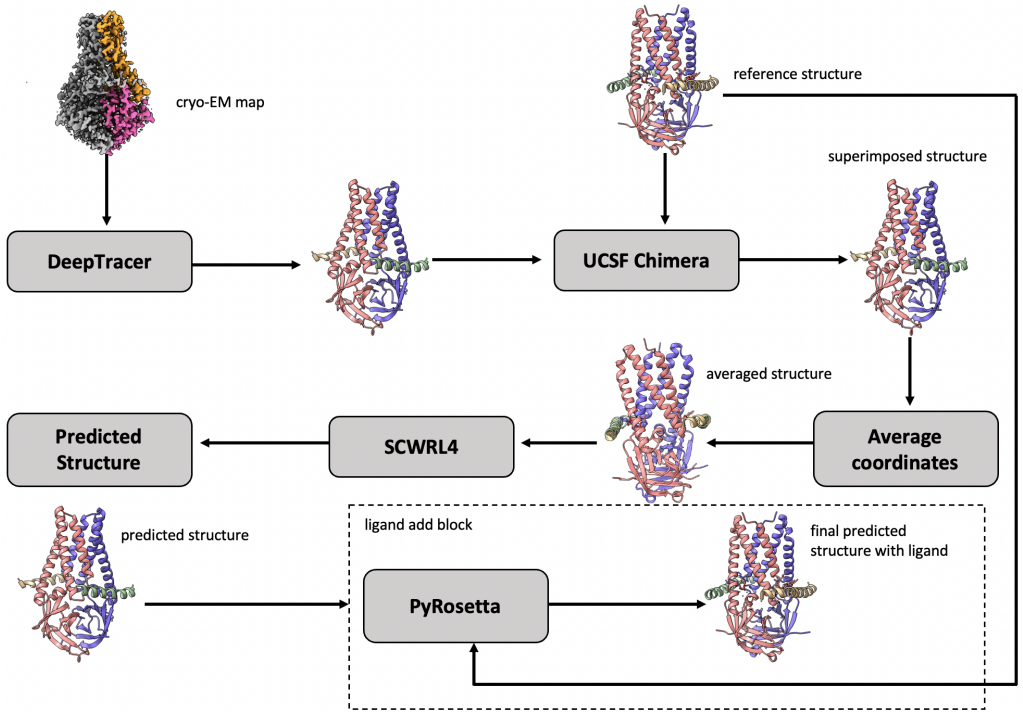


Figure 2. The workflow of DeepProLigand generating protein complex structure from cryo-EM map and reference structure. The cryo-EM map (EMD-22898) illustrated in the workflow is of a SARS-CoV-2 ORF3a ion channel in lipid nanodiscs [37].

2.1. Protein Complex Reconstruction from cryo-EM Density Maps and Reference Structures

Using DeepTracer [38], we first predicted the 3D backbone coordinates of the protein complex directly from a cryo-EM density map. DeepTracer uses a 3D U-Net architecture which is modified from the original 2D U-Net [39] architecture developed for biomedical image segmentation. The output from the DeepTracer block is a predicted 3D backbone coordinate structure that has the carbon, carbon alpha, nitrogen and oxygen atoms in the Protein Data Bank Format (PDB), which is standardized by wwPDB [40]. The predicted structure reflects the conformation of the protein in the ligand binding mode. Because the reference structure of the protein (prior structure without cryo-EM information) is also provided by 2021 Ligand Model Challenge organizers, we used the structural alignment to

combine them to generate a posterior structure, conceptually similar to combining the prior probability and likelihood to generate a posterior probability in the Bayesian reasoning. Specifically, in order to combine the reference structure and the predicted structure together in terms of geometrical alignment, we utilized the UCSF Chimera's [41] matchmaker function to superimpose both structures together. Once the structures were superimposed, we saved the superimposed structure relative to reference structure into a new PDB file. The new PDB file then contained the atoms of both reference and predicted structures in the same geometrical space, allowing us to average the coordinates of the corresponding backbone atoms and utilize the reference structure's residue, and chain labeling for all the shared components between the two structures. The side chains were added on top of the combined backbone structures using the SCWRL4 [42] tool. Finally, a full-atom combined structure, consisting of multiple chains, was produced for the downstream processing. It is worth noting that our approach of generating protein complex structures was different from the traditional approach of fitting a reference structure into a cryo-EM density map.

Algorithm 1 depicts the pseudo code of averaging the backbone atoms' coordinates of reference and predicted structures. We started by initializing an empty PDB file named "average structure" that followed the guidelines of wwPDB [40]. For each residue of the reference structure, if the distance between the reference structure's carbon alpha atom and any carbon alpha atom of predicted backbone structure in the 3D geometrical space was less than a threshold value, then all the backbone atoms' coordinates of the residue in the particular reference structure were averaged with the predicted structure's corresponding residue and saved in the average structure PDB file; otherwise, the coordinates of the residue in reference structure were simply saved in the average structure PDB file. We refer to the arithmetic mean, the sum of collection of numbers divided by the count of numbers, as "average" throughout the paper. The default distance threshold value we used is 1 Angstrom (Å); however, the threshold value for each chain can be modified as desired.

$$distance = \sqrt{(x_r - x_p)^2 + (y_r - y_p)^2 + (z_r - z_p)^2}, \quad (1)$$

$$threshold = 1\ Angstrom. \quad (2)$$

Here, let x_r, y_r, and z_r be the coordinates for each carbon alpha residue of the reference structure and x_p, y_p, and z_p be the coordinates for each carbon alpha residue of the predicted structure. With this notation, we followed Algorithm 1 and generated new coordinates x_{avg}, y_{avg}, and x_{avg} which are the average coordinates of both reference and predicted structures. After computing the distance using Equation (1), we used a threshold value as shown in Equation (2) for Target 202 and Target 203 of the 2021 EMDataResource Ligand Challenge. For Target 201 of the challenge, since most of the chains were turned into coils, we used a threshold value of 0.3 Angstrom (Å) for chain C and 0.5 Angstrom (Å) for all other chains. The averaged coordinate structure was saved into a standard PDB file format. Tables 1–3 show the number of residues averaged per chain for each target. Target T201 has 73.55% residues averaged, target T202 has 86.60% residues averaged, and finally target T203 has 57.70% residues averaged.

After the backbone atoms were computed using Algorithm 1, we utilized SCWRL4 [42] to add the side-chain conformation into the protein structure. The deep learning-based method utilized to predict the backbone atoms had a high impact on determining the side-chains conformation as well, because high side chain accuracy is often achieved when the backbone prediction is accurate, as also demonstrated by AlphaFold [35].

Algorithm 1 Average predicted structure and reference structure.

Require: threshold ▷ threshold can be modified per chain

compute *distance* using Equation (1)
initialize: $x_{avg} = 0$, $y_{avg} = 0, and\ z_{avg} = 0$
if $distance < threshold$ **then**
    $x_{avg} \Leftarrow (x_r + x_p)/2$
    $y_{avg} \Leftarrow (y_r + y_p)/2$
    $z_{avg} \Leftarrow (z_r + z_p)/2$
else
    $x_{avg} \Leftarrow x_r$
    $y_{avg} \Leftarrow y_r$
    $z_{avg} \Leftarrow z_r$
end if

Table 1. Number of residues averaged for target T201: EMD 7770.

Target T201 : EMD 7770			
Chain ID	**Total Residues**	**Averaged Residues**	**% of Residues Averaged**
Chain A	1021	845	82.8
Chain B	1021	845	82.8
Chain C	1021	461	45.2
Chain D	1021	852	83.4
Average % across chains			73.55

Table 2. Number of residues averaged for target T202: EMD 30210.

Target T202 : EMD 30210			
Chain ID	**Total Residues**	**Averaged Residues**	**% of Residues Averaged**
Chain A	834	762	91.4
Chain B	114	105	92.1
Chain C	63	48	76.2
Average % across chains			86.6

Table 3. Number of residues averaged for target T203: EMD 22898.

Target T203 : EMD 22898			
Chain ID	**Total Residues**	**Averaged Residues**	**% of Residues Averaged**
Chain A	193	184	95.3
Chain B	193	0	0
Chain C	31	20	64.5
Chain D	31	22	71.0
Average % across chains			57.7

2.2. Template-Based Prediction of Protein-Ligand Interaction

After the protein structure that can accommodate ligands was generated using Algorithm 1, we utilized PyRosetta [43] to identify ligands and add them into the predicted structure by using the reference structure as a template, as depicted in Algorithm 2. The reference structure contains the ligands' atomic coordinates. Since PyRosetta is a residue based tool, when a pose is created, all the atoms in a structure including ligand atoms are indexed by residue indices. Following Algorithm 2, we let *res* be each residue in

the reference structure that we checked for whether it was a ligand. PyRosetta's *is_ligand* function works by comparing the ligand to a chemical component dictionary and returns a bool value (i.e., either True for ligand or False for non ligand) for each residue.

Algorithm 2 Identify ligands and include them into average structure.

Require: pyrosetta
initialize *pyrosetta*
pose_ref = *pose_from_pdb*(*reference structure*)
if *pose_ref.residue*(*res_id*).*is_ligand*() == *True* **then**
    with *open*("*average_structure.pdb*","*a*") *as file* :
        file.write(*residue*)
else
    do nothing
end if

2.3. Refinement of Protein-Ligand Complex Model

After the prediction of the protein–ligand complex structure using the approach outlined above, we further refined the predicted complex structure using Rosetta FastRelax. Relax does not perform extensive refinement and only searches the local low-energy backbone and side-chain conformations near the starting conformations by implementing rounds of packing and minimizing, with repulsive weight in the scoring function gradually increasing from a low value to a normal value. The scoring function we used was *ref2015_cst.wts*, which is a default score function, repeated five times. Finally, after the refinement of the protein complex, we used UCSF Chimera's *Fit in Map* function to perform a rigid body optimization of the refined model. The 3D structure was rotated and aligned so that it fit to the density map. This refinement step was optional. During the blind experiment of the 2021 EMDataResource Ligand Challenge, we submitted both an unrefined model and a refined model for each target.

2.4. Target cryo-EM Density Maps of 2021 Ligand Challenge

We blindly tested the protein–ligand modeling pipeline DeepProLigand on three targets that were released as 2021 EMDataResource Ligand Challenge targets from February to April 2021. The next section elaborates the three targets and the experimental setting used for each target.

2.4.1. Target 201: *Escherichia coli* Beta-Galactosidase

The β-Galactosidase [44] target with atomic resolution of 1.9 Angstrom (Å) contains protein Beta-galactosidase, magnesium ion, sodium ion, water and 2-phenylethyl 1-thio-beta-D-galactopyranoside (PTQ) as a ligand. The EMDB ID of the target in EMDataResource is EMD-7770. We predicted the 3D structure of the complex using the workflow of DeepProLigand, as highlighted in Figure 2 and, during averaging of the structure, we initialized a 0.3 Angstrom (Å) distance threshold for chain C and a 0.5 Angstrom (Å) distance threshold for all other chains of the complex by re-initializing the threshold value of Equation (2). The reason for threshold of 0.3 Å in chain C was because most of the chains were turned into coils/turns with a threshold of 0.5 Å. The ligand PTQ was appended using Algorithm 2. Figure 3 shows the map–model overlay of cryo-EM density map EMD-7770 and our reconstructed protein structure model.

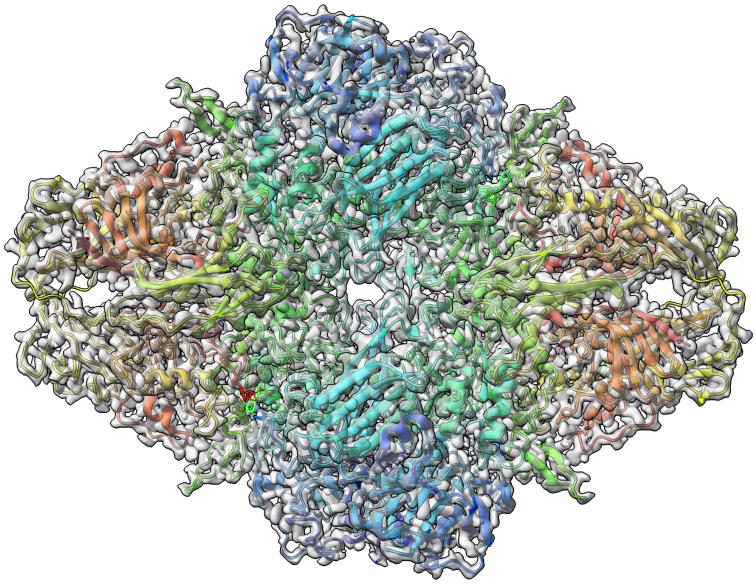


Figure 3. Target 201 (EMD-7770) map-model overlay at the recommended contour 0.52 (3.3 σ) with T0201EM0004_1 (ours).

2.4.2. Target 202: SARS-CoV-2 RNA-Dependent RNA Polymerase

The nsp12-nsp7-nsp8 complex bound to the template-primer RNA and triphosphate form of Remdesivir(RTP) [45] target with an atomic resolution of 2.5 Angstrom (Å) contains RNA-directed RNA polymerase, Non-Structural Protein 8, Non-Structural Protein 7, Primer, Templete, ZINC ION, PYROPHOSPHATE 2-, MAGNESIUM ION, water, and [(2 R,3 S,4 R,5 R)-5-(4-azanylpyrrolo[2,1-f][1,2,4]triazin-7-y)-5-cyano-3,4-bis(oxidanyl)oxolan-2-yl]methyl dihydrogen phosphate as a ligand. The EMDB ID of the target in EMDataResource is EMD-30210. We predicted the 3D structure of the complex using the workflow of DeepProLigand as highlighted in Figure 2 and, during averaging of the structure, we used a 1 Angstrom (Å) distance threshold for all chains of the complex. The ligand F86 (remdesivir, covalent inhibitor) was appended using Algorithm 2. Figure 4 shows the map–model overlay of cryo-EM density map EM-30210 and our reconstructed protein structure model.

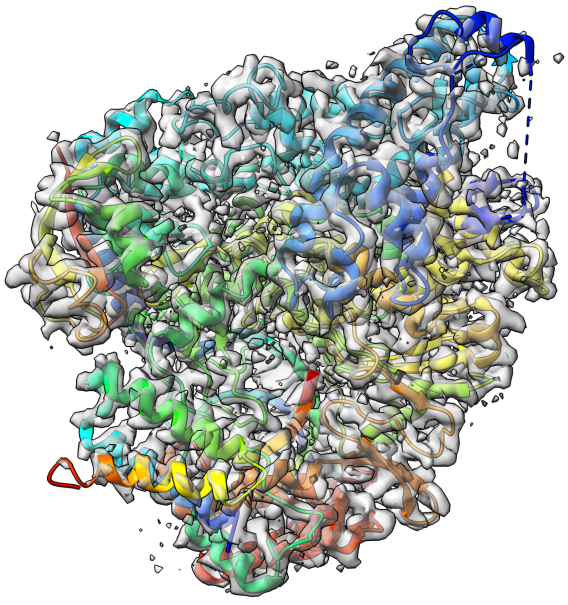


Figure 4. Target 202 (EMD-30210) map-model overlay at the recommended contour 0.058 (4.3 σ) with T0202EM004_1 (ours).

2.4.3. Target 203: SARS-CoV-2 Protein 3a in Lipid Nanodiscs

The SARS-CoV-2 3a ion channel in lipid nanodiscs [37] target with atomic resolution of 2.08 Angstrom (Å) contains ORF3a protein, Apolipoprotein A-I, water and 1,2-Dioleoyl-sn-glycero-3-phosphoethanolamine as a ligand. The EMDB ID of the target in EMDataResource is EMD-22898. We predicted the 3D structure of the complex using the workflow of DeepProLigand as highlighted in Figure 2 and, during averaging of the structure, we used a 1 Angstrom (Å) distance threshold for all chains of the complex. The ligand PEE was appended using Algorithm 2. Figure 5 shows the map–model overlay of cryo-EM density map EMD-22898 and our reconstructed protein structure model. The source code, data, and instructions to reproduce the results are available in the GitHub repository: https://github.com/jianlin-cheng/DeepProLigand, accessed on 8 January 2023.

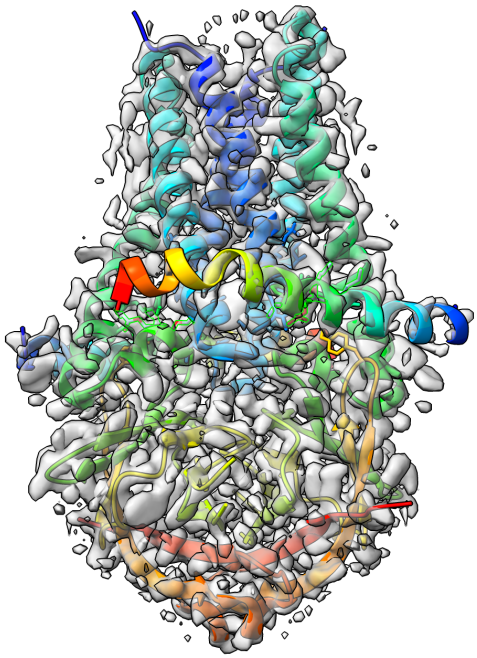


Figure 5. Target 203 (EMD-22898) map-model overlay at the recommended contour 0.7 (10.3 σ) with T0203EM004_1 (ours).

3. Results

The analysis of the models in this section is based on the official results provided by the organizers of the 2021 Ligand Model Challenge. The fit to a map for a ligand was assessed by the Q-score [46] and the Z-scores. The Q-score measures how similar map values around an atom are to a Gaussian-like function which we would see if the atom were well resolved. The Q-score was calculated as a correlation between two vectors: u, which contained map values at points around the atom, and v, which contained values obtained from the reference Gaussian.

We used the Q-score to compare the map-to-model fit for all the models that were submitted to the challenge. Table 4 shows the Q-score of the ligand for all the models submitted for Target 201; our model is highlighted in bold for scrutiny. Figure 6 shows the ligand (PTQ)'s binding pose and orientation in our best predicted model, T0201EM004_1. Ligand PTQ bound to all four chains of Target 201, resulting in four binding sites for the ligand. We visualized three binding locations for the ligand with its binding pose and orientations in Figure 6. Table 5 shows the Q-score of the ligand for Target 202 and, similar to Target 201, our model is highlighted in bold for scrutiny. Figure 7 shows ligand (F86)'s binding pose and orientation in our best predicted model, T0201EM004_1. Ligand F86 bound to only one location in Target 202. We visualized the binding location for the ligand with its binding pose and orientations in Figure 7.

Table 4. Evaluation of Target 201: *Escherichia coli* beta-galactosidase on Q-score for all the models submitted in the 2021 Ligand Model Challenge

Team Name	PTQ (Ligand)
T0201EM014_1	0.82
T0201EM005_2	0.81
T0201EM012_1	0.81
T0201EM006_1	0.79
T0201EM009_1	0.78
T0201EM015_1	0.77
T0201EM005_1	0.73
T0201EM002_2	0.72
T0201EM004_1	**0.71**
T0201EM003_1	0.71
T0201EM003_2	0.71
T0201EM010_1	0.69
T0201EM011_1	0.64
T0201EM001_2	0.64
T0201EM013_1	0.63
T0201EM001_3	0.62
T0201EM002_3	0.60
T0201EM003_3	0.58
T0201EM002_1	0.55
T0201EM001_1	0.33
T0201EM007_1	0.31
T0201EM008_1	-

Note: Table is sorted in descending order using ligand: PTQ score. "-" means, we were unable to calculate the score of the model. Our best model is highlighted in bold.

Table 6 shows the Q-score of the ligand for Target 203. Similar to Target 201 and 202, our model is highlighted in bold for scrutiny in Table 6. Figure 8 shows the ligand (PEE)'s binding pose and orientation in our best predicted model, T0201EM004_1. Ligand PEE bound to two locations in Target 203. We have visualized the binding locations for the ligand with its binding pose and orientation in Figure 8.

Table 5. Evaluation of Target 202: SARS-CoV-2 RNA-dependent RNA polymerase on Q-score for all the models submitted to the 2021 Ligand Model Challenge.

Team Name	F86 (Ligand)
T0202EM004_1	**0.74**
T0202EM009_1	0.71
T0202EM006_1	0.69
T0202EM005_1	0.68
T0202EM012_1	0.68
T0202EM002_2	0.68
T0202EM003_2	0.68
T0202EM010_1	0.67
T0202EM003_1	0.67
T0202EM008_1	0.63
T0202EM001_1	0.60
T0202EM001_2	0.59
T0202EM013_1	0.59
T0202EM007_1	0.57
T0202EM011_1	0.56
T0202EM002_1	0.52

Note: Table is sorted in descending order using ligand: F86 score. Our best model is highlighted in bold.

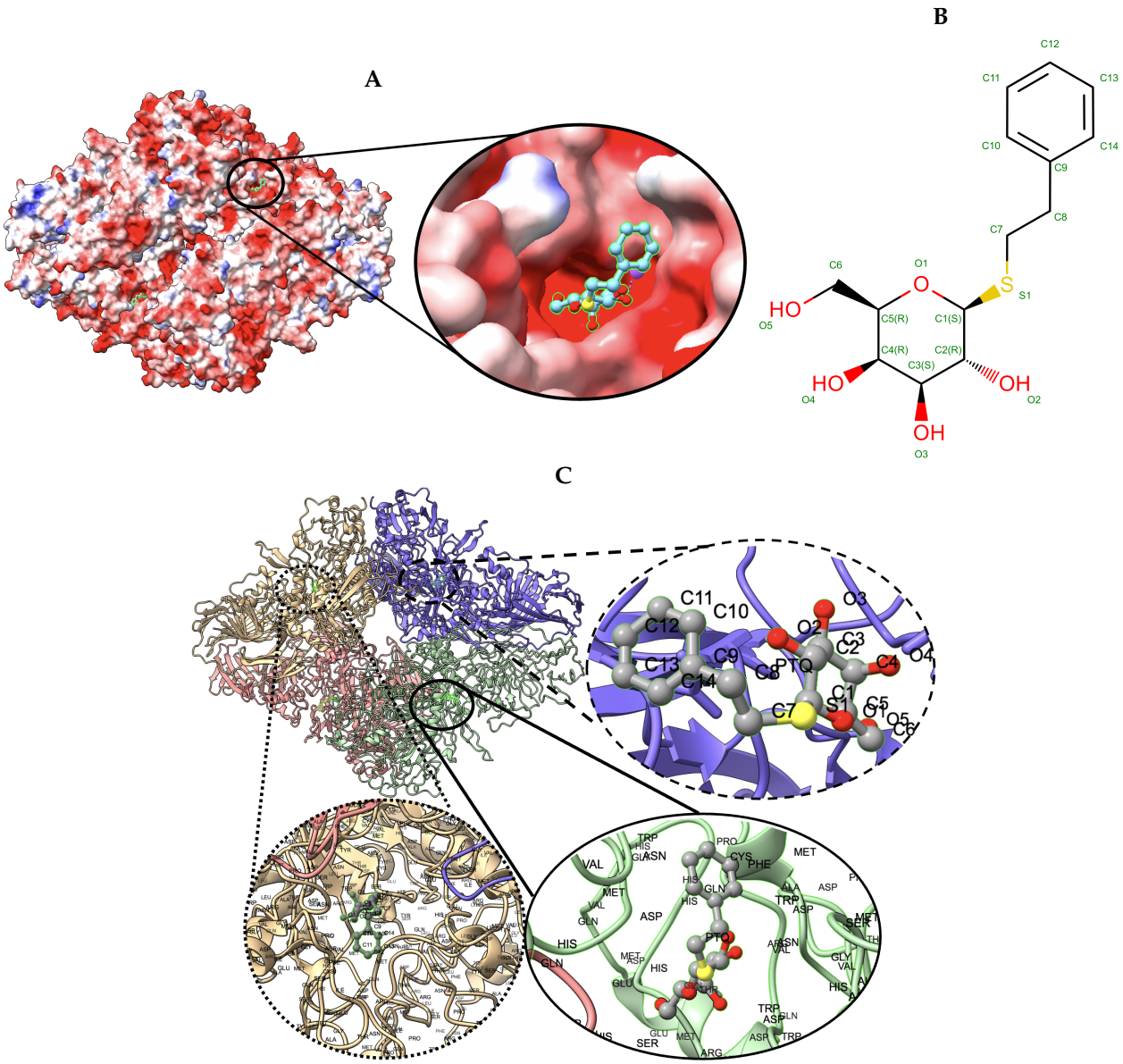


Figure 6. Target 201. (**A**) T0201EM004_1 (ours) docked by Target 201 (EMD-7770) and visualized with electrostatic potential surface generated in UCSF Chimera. (**B**) Ligand PTQ, image extracted from Protein Data Bank (PDB). (**C**) Protein–ligand interactions in T0201EM004_1 (ours) model. Chains are colored differently (chain A: blue, chain B: pink, chain C: green and chain D: golden). The ligand is labeled with its atom names as well as the ligand name (PTQ). For chain D: golden and chain C: green, we have labeled the chain residue names for understanding protein–ligand interaction better.

Figure 9 shows cumulative Z-scores on Q-scores of 17 groups participating in the 2021 Ligand Model Challenge. Our DeepProLigand predictor (EM004) performs best overall on all three targets. Specifically, our protein–ligand model was ranked first for Target 202, second for Target 203, and in the middle for Target 201 as shown in the Z-scores on Q-scores in Figure 9. The Q-scores of our best model for the three targets are shown in Table 7. Even though there are too few targets to draw a definite conclusion, the good results indicate that the deep learning structure prediction in conjunction with the template reference structure is able to build a good protein structure framework to accommodate ligands, and the template-based protein–ligand prediction can assemble the ligands with the protein structure well for some targets. Incorporating a deep learning approach in modeling enables us to predict the protein structure directly from the cryo-EM map within minutes, making the approach highly useful in terms of both prediction accuracy and time.

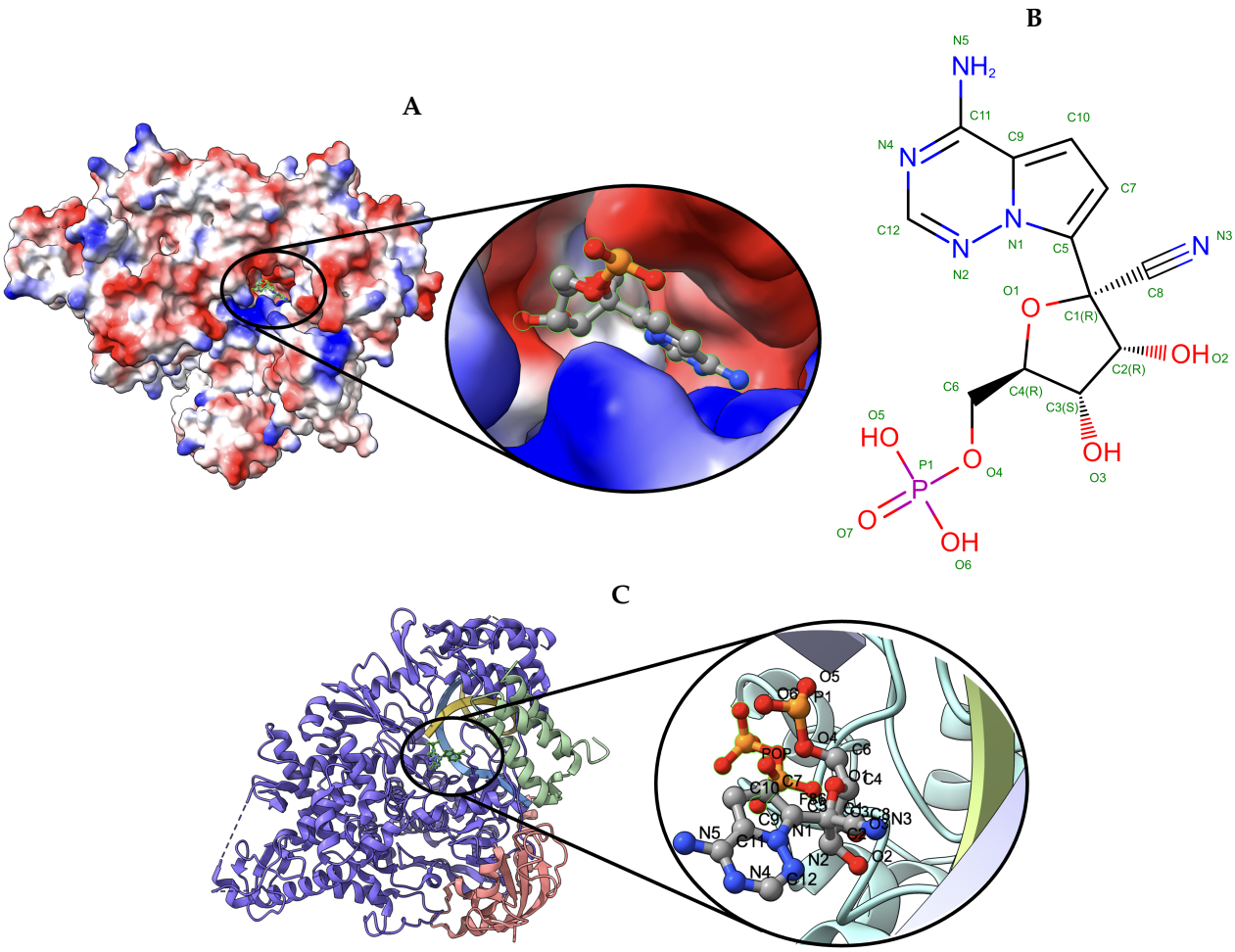


Figure 7. Target 202. (**A**) T0202EM0004_1 (ours) docked by Target 202 (EMD-30210) and visualized with electrostatic potential surface generated in UCSF Chimera. (**B**) Ligand F86, image extracted from Protein Data Bank (PDB). (**C**) Protein–ligand interactions in T0202EM004_1 (ours) model. Chains are colored differently (chain A: blue, chain B: orange, chain C: green, chain P: yellow, and chain T: teal). The ligand is labeled with its atom names as well as the ligand name (F86).

We also noticed the limitation of our approach in terms of the geometric quality of the atoms in the predicted protein structure, however. Particularly, there were some atom–atom clashes in the models, which may be caused by the violations of some geometric constraints of atom–atom distances in the protein structure predicted by the deep learning, as well as in the averaging of the coordinates of the predicted structure and the reference structure. The violations of geometric and stereochemical restraints were not fixed by the current refinement protocol in the prediction pipeline. The refinement protocol even introduced some new clashes into the model. AlphaFold demonstrated that the well-trained sophisticated deep learning architecture can accurately capture the geometric restraints of atoms and bonds in protein structures by predicting high-quality protein structures of atomic resolution that are highly similar to natural protein structures; this means more advanced deep learning architectures can be developed to predict high-quality protein structures compatible with the geometric and stereochemical restraints of proteins from cryo-EM density maps and reference structures.

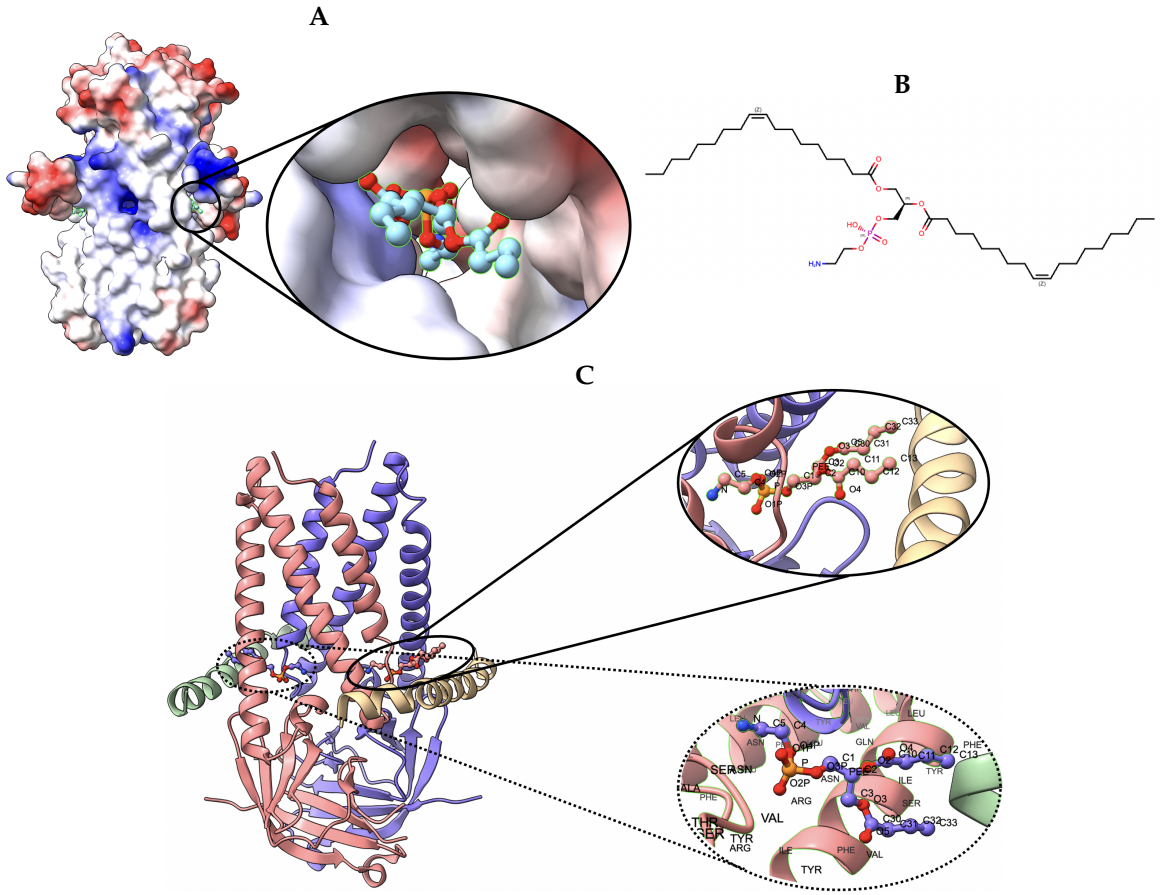


Figure 8. Target 203. (**A**) T0203EM0004_1 (ours) docked by Target 203 (EMD-22898) and visualized with electrostatic potential surface generated in UCSF Chimera. (**B**) Ligand PEE, image extracted from Protein Data Bank (PDB). (**C**) Protein–ligand interactions in T0203EM0004_1 (our) model. Chains are colored differently (chain A: blue, chain B: pink, chain C: green, and chain D: golden). The ligand are labeled with their atom names as well as the ligand's name (PEE).

Table 6. Evaluation of Target 203: SARS-CoV-2 ORF3a putative ion channel in nanodisc on Q-score for all the models submitted in the 2021 Ligand Model Challenge.

Team Name	PEE (Ligand)
T0203EM0016_1	0.77
T0203EM004_1	**0.76**
T0203EM0012_1	0.75
T0203EM005_1	0.74
T0203EM0010_1	0.73
T0203EM003_1	0.73
T0203EM003_2	0.70
T0203EM0011_1	0.72
T0203EM002_2	0.72
T0203EM009_1	0.71
T0203EM002_1	0.70
T0203EM006_1	0.70
T0203EM002_3	0.69
T0203EM0014_1	0.67
T0203EM001_2	0.66
T0203EM001_3	0.63
T0203EM008_1	0.63
T0203EM001_1	0.60
T0203EM007_1	0.51

Note: Table is sorted in descending order using ligand: PEE score. Our best model is highlighted in bold.

Table 7. Q-score of our best model (EM004_1) for three targets.

Target Name	Ligand
Target 201	0.71 (PTQ)
Target 202	0.74 (F86)
Target 203	0.76 (PEE)

Note: Among two models submitted for each target in the challenge, our best model's id is EM004_1 across all the targets. EM004_1 model is the non refined model.

Figure 9. Z-scores on Q-scores for ligand of all the models submitted to 2021 Ligand Model Challenge. The pointed arrow represents our model.

4. Conclusions and Future Work

In this work, we demonstrate that the deep learning prediction of protein structures from cryo-EM maps can generate good protein structures for constructing protein–ligand complexes and the template-based protein–ligand interaction prediction can fit ligands well into the predicted protein structures according to the outstanding performance of our protein–ligand modeling pipeline. It is also worth noting that our method was fully automatic and did not involve any manual tweaking of the models to improve the scores. As discussed before, the current protein–ligand prediction pipeline cannot resolve some violations of some geometric and stereochemical restraints of atoms in protein structures. We plan to soon develop advanced end-to-end deep learning architectures, similar to some components in AlphaFold, to better predict better protein structures from cryo-EM maps and reference structures. Moreover, we plan to design 3D-equivariant deep learning architectures like the SE(3)-equivariant Transformer network [47–50] to tackle the problem of geometric constraints which are not addressed by current methods. Finally, an end-to-end direct deep learning prediction of the structure of protein–ligand complexes from cryo-EM density maps, reference structures and ligand information to fully automate all the steps of the entire pipeline in this work will be pursued. We believe the application of a deep learning approach to the prediction of 3D structures of protein–ligand complexes leveraging cryo-EM and other related data is a promising avenue with which to accelerate the advancement of the study of protein–ligand interaction [51,52]. With the proliferation of cryo-EM maps being deposited in the EMDataResource database, the use of deep learning-based methods can help to determine the structure of the protein–ligand complexes rapidly and ultimately help to expedite the drug discovery process.

Author Contributions: Conceptualization, J.C.; methodology, J.C. and N.G.; experiments and data collection, N.G.; data analysis, N.G. and J.C.; writing—original draft preparation, N.G. and J.C.; writing—review and editing, J.C. and N.G.; visualization, N.G.; supervision, J.C.; project administration, J.C. and N.G.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by two NIH grants (R01GM146340 and R01GM093123), Department of Energy grants (DE-AR0001213, DE-SC0020400 and DE-SC0021303), and two NSF grants (DBI1759934 and IIS1763246).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The cryo-EM density maps are available in the EMDB website [2] with ID: EMD-7770, EMD-30210, and EMD-22898. The submitted models metadata and further detailed information on 2021 Ligand Model Challenge can be accessed from the official challenge web-page [3]. The protein complex generated from this method, DeepProLigand, and the software program are released, and can be accessed through the GitHub repository: https://github.com/jianlin-cheng/DeepProLigand, accessed on 20 November 2022.

Acknowledgments: We thank the organizers and assessors of the 2021 EMDataResource Ligand Challenge for providing the data for this research and anonymous reviewers for their valuable suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

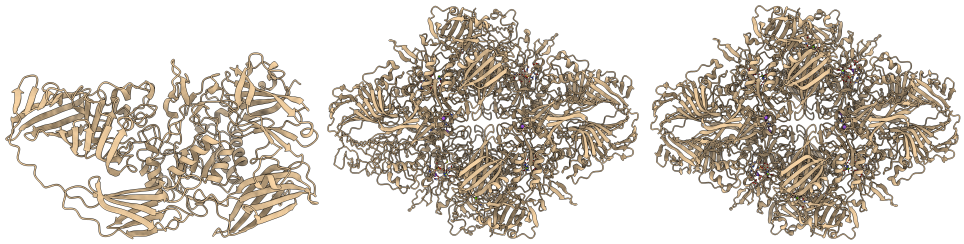
Abbreviations

The following abbreviations are used in this manuscript:

cryo-EM	Cryogenic electron microscopy
3D	Three-dimensional

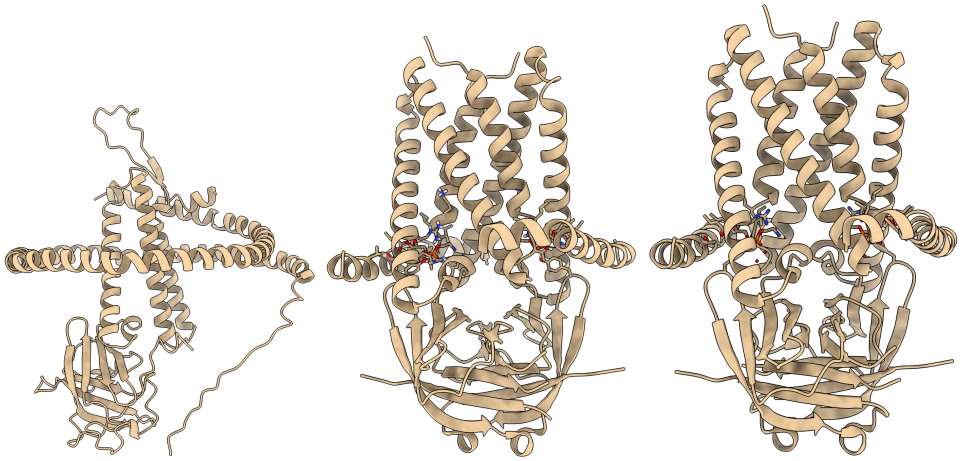
Appendix A. Sequence Based Modeling: An Approach for Predicting Protein Structure

DeepProLigand uses DeepTracer [38] to predict the protein structure as its main component because it uses both the cryoEM density map and the amino acid sequence. We conducted another study to predict the protein structure using AlphaFold [35]. While AlphaFold predicts the atomic coordinates of most proteins with remarkable accuracy, in this study, AlphaFold struggled to predict the atomic coordinates that fit locally into the density map per residue. Figures A1 and A2 shows the comparison of structures predicted by AlphaFold and DeepProLigand with the PDB deposited structure.



(**a**) (**b**) (**c**)

Figure A1. The target T201 is of EMD: 7770. (**a**) AlphaFold Predicted Structure. (**b**) DeepProLigand Predicted Structure. (**c**) PDB Deposited Structure with PDB ID: 6CVM.



(**a**) (**b**) (**c**)

Figure A2. The target T203 is of EMD: 22898. (**a**) AlphaFold Predicted Structure. (**b**) DeepProLigand Predicted Structure. (**c**) PDB Deposited Structure with PDB ID: 7KJR.

References

1. Cressey, D.; Callaway, E. Cryo-electron microscopy wins chemistry Nobel. *Nature* **2017**, *550*. [CrossRef] [PubMed]
2. EMDataResource. Available online: https://www.emdataresource.org/ (accessed on 22 November 2022).
3. 2021 Ligand Model Challenge. Available online: https://challenges.emdataresource.org/?q=2021-model-challenge (accessed on 22 November 2022).
4. Eberhardt, J.; Santos-Martins, D.; Tillack, A.F.; Forli, S. AutoDock Vina 1.2.0: New docking methods, expanded force field, and python bindings. *J. Chem. Inf. Model.* **2021**, *61*, 3891–3898. [CrossRef] [PubMed]
5. Morris, C.J.; Della Corte, D. Using molecular docking and molecular dynamics to investigate protein–ligand interactions. *Mod. Phys. Lett. B* **2021**, *35*, 2130002. [CrossRef]
6. Fan, J.; Fu, A.; Zhang, L. Progress in molecular docking. *Quant. Biol.* **2019**, *7*, 83–89 [CrossRef]
7. Stanzione, F.; Giangreco, I.; Cole, J.C. Use of molecular docking computational tools in drug discovery. *Prog. Med. Chem.* **2021**, *60*, 273–343. [PubMed]
8. Mansoor, S.; Shahid, S.; Ashiq, K.; Alwadai, N.; Javed, M.; Iqbal, S.; Fatima, U.; Zaman, S.; Sarwar, M.N.; Alshammari, F.H.; et al. Controlled growth of nanocomposite thin layer based on Zn-Doped MgO nanoparticles through Sol-Gel technique for biosensor applications. *Inorg. Chem. Commun.* **2022**, *142*, 109702. [CrossRef]
9. Shahid, S.; Anam, E.; Mohsin, J.; Sana, M.; Shahid, I.; Eslam, B.E.; Rami, M.A.; Alsaab Hashem, O.; Awwad Nasser, S.; Ibrahium Hala, A.; et al. The Anti-Inflammatory and Free Radical Scavenging Activities of Bio-Inspired Nano Magnesium Oxide. *Front. Mater.* **2022**, *9*, 875163. [CrossRef]
10. Cui, Y.; Dong, Q.; Hong, D.; Wang, X. Predicting protein–ligand binding residues with deep convolutional neural networks. *BMC Bioinform.* **2019**, *20*, 93. [CrossRef]
11. Xia, C.-Q.; Pan, X.; Shen, H.-B. Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. *Bioinformatics* **2020**, *36*, 3018–3027. [CrossRef]
12. Mylonas, S.K.; Axenopoulos, A.; Daras, P. DeepSurf: A surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics* **2021**, *37*, 1681–1690. [CrossRef]
13. Kandel, J.; Tayara, H.; Chong, K.T. PUResNet: Prediction of protein–ligand binding sites using deep residual neural network. *J. Cheminform.* **2021**, *13*, 65. [CrossRef] [PubMed]
14. Yang, J.; Roy, A.; Zhang, Y. BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **2013**, *41*, 1096–1103. [CrossRef] [PubMed]
15. Hu, J.; Li, Y.; Zhang, Y.; Yu, D.-J. ATPbind: Accurate protein–ATP binding site prediction by combining sequence-profiling and structure-based comparisons. *J. Chem. Inf. Model.* **2018**, *58*, 501–510. [CrossRef] [PubMed]
16. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. Sc-PDB: A 3D-database of ligandable binding sites-10 years on. *Nucleic Acids Res.* **2015**, *43*, D399–D404. [CrossRef] [PubMed]
17. Ashtawy, H.M.; Mahapatra, N.R. BgN-Score and BsN-Score: Bagging and boosting based ensemble neural networks scoring functions for accurate binding affinity prediction of protein–ligand complexes. *BMC Bioinform.* **2015**, *16*, S8. [CrossRef]
18. Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K deep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inf. Model.* **2018**, *58*, 287–296. [CrossRef]
19. Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep drug–target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821–i829. [CrossRef]

20. Zheng, L.; Fan, J.; Mu, Y. Onionnet: A multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS Omega* **2019**, *4*, 15956–15965. [CrossRef]
21. Zhu, F.; Zhang, X.; Allen, J.E.; Jones, D.; Lightstone, F.C. Binding affinity prediction by pairwise function based on neural network. *J. Chem. Inf. Model.* **2020**, *60*, 2766–2772. [CrossRef]
22. Rezaei, M.A.; Li, Y.; Wu, D.; Li, X.; Li, C. Deep learning in drug design: Protein–ligand binding affinity prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *19*, 407–417. [CrossRef]
23. Jones, D.; Kim, H.; Zhang, X.; Zemla, A.; Stevenson, G.; Drew Bennett, W.F.; Kirshner, D.; Wong, S.E.; Lightstone, F.C.; Allen J.E. Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *J. Chem. Inf. Model.* **2021**, *61*, 1583–1592. [CrossRef] [PubMed]
24. Kwon, Y.; Shin, W.H.; Ko, J.; Lee, J. AK-score: Accurate proteinligand binding affinity prediction using an ensemble of 3D-convolutional neural networks. *Int. J. Mol. Sci.* **2020**, *21*, 8424. [CrossRef] [PubMed]
25. Karlov, D.S.; Sosnin, S.; Fedorov, M.V.; Popov, P. GraphDelta: MPNN scoring function for the affinity prediction of protein–ligand complexes. *ACS Omega* **2020**, *5*, 5150–5159. [CrossRef] [PubMed]
26. Wang, K.; Zhou, R.; Li, Y.; Li, M. DeepDTAF: A deep learning method to predict protein–ligand binding affinity. *Brief. Bioinform.* **2021**, *22*, bbab072. [CrossRef] [PubMed]
27. Azzopardi, J.; Ebejer, J.P. LigityScore: Convolutional neural network for binding-affinity predictions. *Bioinformatics* **2021**, *3*, 38–49.
28. Seo, S.; Choi, J.; Park, S.; Ahn, J. Binding affinity prediction for protein–ligand complex using deep attention mechanism based on intermolecular interactions. *BMC Bioinform.* **2021**, *22*, 542. [CrossRef]
29. Ahmed, A.; Mam, B.; Sowdhamini, R. DEELIG: A deep learning approach to predict protein–ligand binding affinity. *Bioinform. Biol. Insights* **2021**, *15*, 11779322211030364. [CrossRef]
30. Wang, R.; Fang, X.; Lu, Y.; Yang, C.Y.; Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119. [CrossRef]
31. Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative assessment of scoring functions: The CASF-2016 update. *J. Chem. Inf. Model.* **2018**, *59*, 895–913. [CrossRef]
32. Stärk, H.; Ganea, O.; Pattanaik, L.; Barzilay, R.; Jaakkola, T. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*; PMLR; MLResearch Press: Maastricht, The Netherlands, 2022; pp. 20503–20521.
33. Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *arXiv* **2022**, arXiv:2210.01776.
34. Lawson, C.L.; Patwardhan, A.; Baker, M.L.; Hryc, C.; Garcia, E.S.; Hudson, B.P.; Lagerstedt, I.; Ludtke, S.J.; Pintilie, G.; Sala, R.; et al. EMDataBank unified data resource for 3DEM. *Nucleic Acids Res.* **2016**, *44*, D396–D403. [CrossRef] [PubMed]
35. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
36. Giri, N.; Roy, R.S.; Cheng, J. Deep learning for reconstructing protein structures from cryo-EM density maps: Recent advances and future directions. *arXiv* **2022**, arXiv:2209.08171.
37. Kern, D.M.; Sorum, B.; Mali, S.S.; Hoel, C.M.; Sridharan, S.; Remis, J.P.; Toso, D.B.; Kotecha, A.; Bautista, D.M.; Brohawn, S.G. Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. *Nat. Struct. Mol. Biol.* **2021**, *28*, 573–582. [CrossRef] [PubMed]
38. Pfab, J.; Phan, N.M.; Si, D. DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2017525118. [CrossRef] [PubMed]
39. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.
40. Protein Data Bank: The single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **2019**, *47*, D520–D528. [CrossRef]
41. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [CrossRef]
42. Krivov, G.G.; Shapovalov, M.V.; Dunbrack, R.L., Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins Struct. Funct. Bioinform.* **2009**, *77*, 778–795. [CrossRef]
43. Chaudhury, S.; Lyskov, S.; Gray, J.J. PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **2010**, *26*, 689–691. [CrossRef] [PubMed]
44. Bartesaghi, A.; Aguerrebere, C.; Falconieri, V.; Banerjee, S.; Earl, L.A.; Zhu, X.; Grigorieff, N.; Milne, J.L.S.; Sapiro, G.; Wu, X.; et al. Atomic resolution cryo-EM structure of β-galactosidase. *Structure* **2018**, *26*, 848–856. [CrossRef]
45. Yin, W.; Mao, C.; Luan, X.; Shen, D.D.; Shen, Q.; Su, H.; Wang, X.; Zhou, F.; Zhao, W.; Gao, M.; et al. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science* **2020**, *368*, 1499–1504. [CrossRef] [PubMed]
46. Krissinel, E.; Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2004**, *60*, 2256–2268. [CrossRef] [PubMed]
47. Fuchs, F.; Worrall, D.; Fischer, V.; Welling, M. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1970–1981.

48. Morehead, A.; Chen, X.; Wu, T.; Liu, J.; Cheng, J. EGR: Equivariant Graph Refinement and Assessment of 3D Protein Complex Structures. *arXiv* **2022**, arXiv:2205.10390.
49. Wu, T.; Cheng, J. Atomic protein structure refinement using all-atom graph representations and SE (3)-equivariant graph neural networks. *bioRxiv* **2022**. [CrossRef]
50. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876. [CrossRef]
51. Chang, L.; Wang, F.; Connolly, K.; Meng, H.; Su, Z.; Cvirkaite-Krupovic, V.; Krupovic, M.; Egelman, E.H.; Si, D. DeepTracer ID: De Novo Protein Identification from Cryo-EM Maps. *bioRxiv* **2022**. [CrossRef]
52. Dhakal, A.; McKay, C.; Tanner, J.J.; Cheng, J. Artificial intelligence in the prediction of protein–ligand interactions: Recent advances and future directions. *Brief. Bioinform.* **2022**, *23*, bbab476. [CrossRef]