MDPI

Article

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

Nabin Giri [1,†] and Jianlin Cheng [1,†]*

[1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia; {ngzvh, chengji} @missouri.edu.
* Correspondence: chengji@missouri.edu.
† Current address: Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA.

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.This results demonstrate that the deep learning bioinformatics approach is a promising direction to model protein-ligand interaction on cryo-EM data using prior structural information. The source code, data, and instruction to reproduce the results are available on GitHub repository : https://github.com/jianlin-cheng/DeepProLigand

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

Citation: Giri, N.; Cheng, J. DeepProLigand. *Biomolecules* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Submitted to *Biomolecules* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Proteins are a building block of life and carry out many vital biological functions. Whether it is 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 organism. 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.

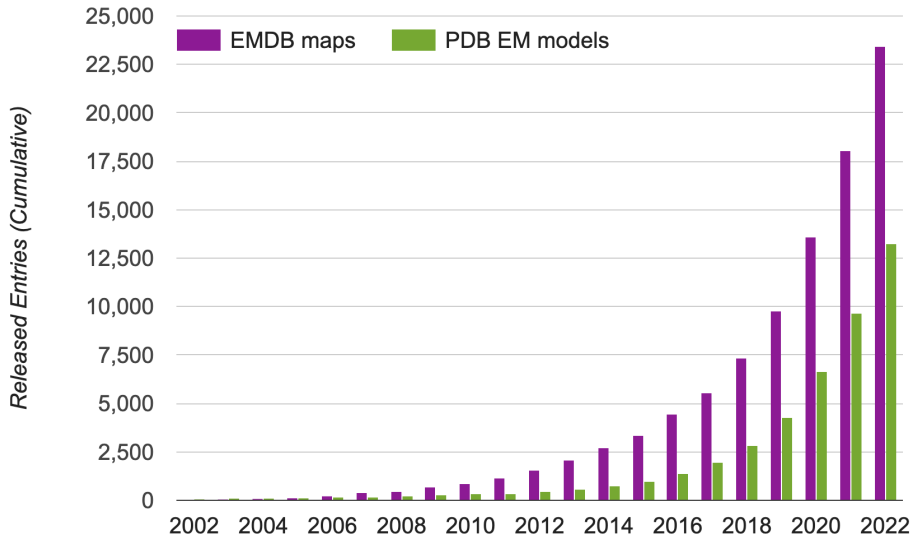


Figure 1. The growth of cryo-EM density maps and cryo-EM-derived protein structures. The statistics was obtained from EMDataResource [16], an unified data resource for 3-Dimension electron microscopy (3DEM) on November 22, 2022.

Cryo-EM, an acronym for cryogenic electron microscopy technique, [1] is a revolutionary technology that enables the determination of 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 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 interaction, 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. 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. Inspired by the recent success demonstrated by AlphaFold, that uses novel deep learning architecture to predict the protein structures using amino acid sequence data as an input, new deep learning architectures such as Equibind [31], an SE(3)-equivariant geometric deep learning model for direct-shot prediction of receptor binding location and ligand's bound pose, and DIFFDOCK [30], a diffusion generative model tailored to the task of molecular docking, have been developed. We in this work combined the deep learning-based protein structure prediction tool - DeepTracer, with the 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 February 1 to April 1, 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.

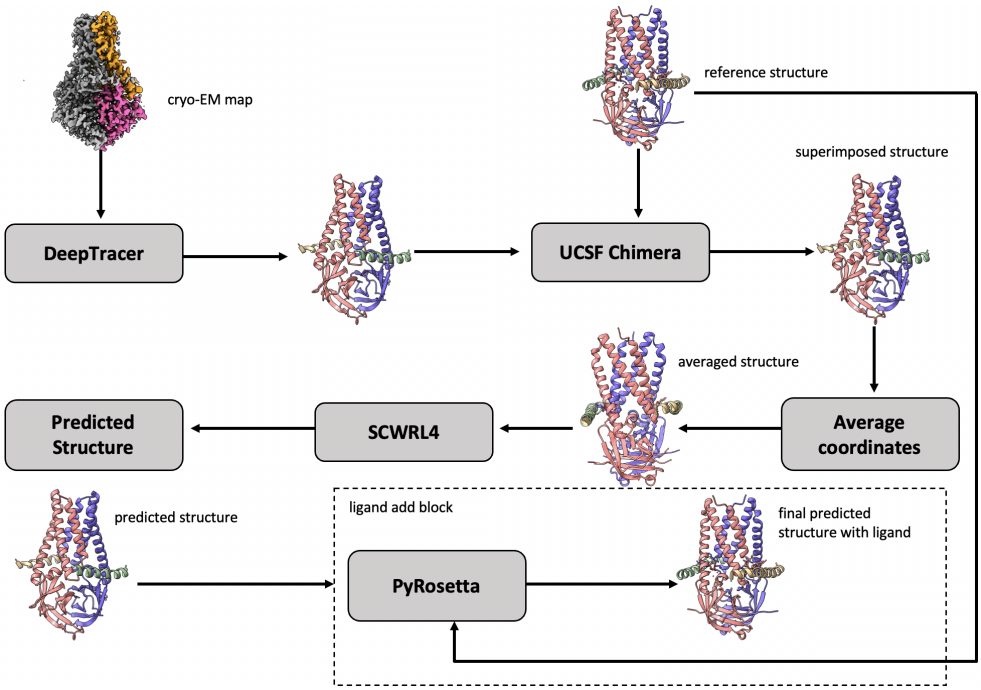


Figure 2. The workflow of `DeepProLigand` generating protein complex structure from cryo-EM map and reference structure. The cryo-EM map (EMD-22898) illustrated in the workflow is of SARS-CoV-2 ORF3a ion channel in lipid nanodiscs [12].

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 structure. 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 on the article [11]. To predict the bound conformation (3D atomic structure) of 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 structure using only 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 generate the structure of 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:

2.1. Protein complex reconstruction from cryo-EM density maps and reference structures

Using DeepTracer [13], we first predicted the 3D backbone coordinates of the protein complex directly from cryo-EM density map. DeepTracer uses a 3D U-Net architecture which is modified from the original 2D U-Net [14] architecture developed for biomedical image segmentation. The output from 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 [15]. 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 structure 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 [17] matchmaker function to superimpose both structure together. Once the structures are superimposed,

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

we savedthe superimposed structure relative to reference structure into a new PDB file. The new PDB file now contains the atoms of both reference and predicted structure into 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 by SCWRL4 [18] tool. Finally, a full-atom combined structure, consisting of multiple chains, is 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 [15]. 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.

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

$$threshold = 1\ Angstrom \tag{2}$$

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 follow the Algorithm 1 and generate new coordinates x_{avg}, y_{avg}, and x_{avg} which are the average coordinates of both reference and predicted structure. After computing the distance using Equation 1, we used 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 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 standard PDB file format. Table 1, 2, and 3 shows 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 [18] to add the side-chain conformation into the protein structure. The deep learning-based method utilized to predict the backbone atoms had 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 [10].

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

Target T201: EMD 7770		
Chain ID	Averaged Residues	% of residues averaged
Chain A	845	82.8
Chain B	845	82.8
Chain C	461	45.2
Chain D	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	Averaged Residues	% of residues averaged
Chain A	762	91.4
Chain B	105	92.1
Chain C	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	Averaged Residues	% of residues averaged
Chain A	184	95.3
Chain B	0	0
Chain C	20	64.5
Chain D	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 [19] to identify ligands and add them into the predicted structure by using the reference structure as 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 chemical component dictionary and returns a bool value (i.e, either True for ligand or False for non ligand) for each residue.

2.3. Refinement of Protein-Ligand Complex Model

After the prediction of 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 increased from low value to normal value. The scoring function we used is *ref2015_cst.wts*, which is a default score function, repeated 5 times. Finally, after the refinement of the protein complex, we used UCSF Chimera's *Fit in Map* function to do a rigid body optimization of the refined model. The 3D structure was rotated and aligned so that it fit to the density map. This

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

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

β-Galactosidase [20] 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 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 0.3 Angstrom (Å) distance threshold for chain C and 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 threshold of 0.5 Å. The ligand PTQ was appended using the Algorithm 2. Figure 3 shows the map-model overlay of cryo-EM density map EMD-7770 and our reconstructed protein structure model.

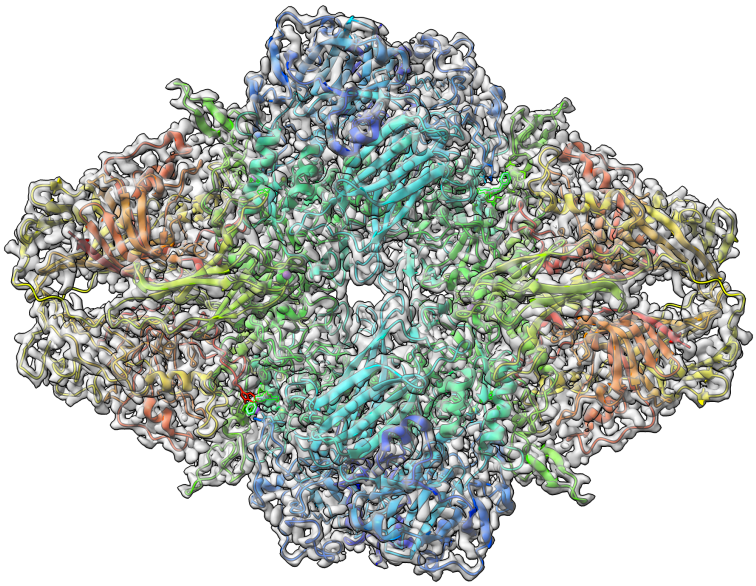


Figure 3. Target 201 (EMD-7770) map-model overlay at the recommended contour 0.52 (3.3 σ) with **T0201EM0004_1** (ours)

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) [21] target with 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 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 on Figure 2, and during averaging of the structure, we used 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.

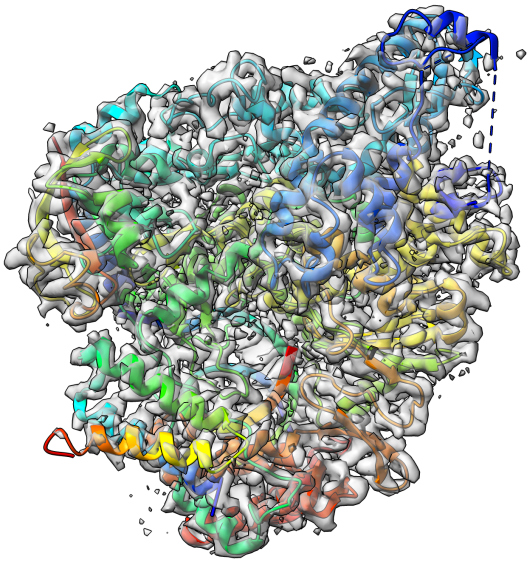


Figure 4. Target 202 (EMD-30210) map-model overlay at the recommended contour 0.058 (4.3 σ) with **T0202EM004_1** (ours)

2.4.3. Target 203: SARS-CoV-2 protein 3a in lipid nanodiscs

SARS-CoV-2 3a ion channel in lipid nanodiscs [22] 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 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 1 Angstrom (Å) distance threshold for all chains of the complex. The ligand PEE was appended using the Algorithm 2. Figure 5 shows the map-model overlay of cryo-EM density map EMD-22898 and our reconstructed protein structure model.

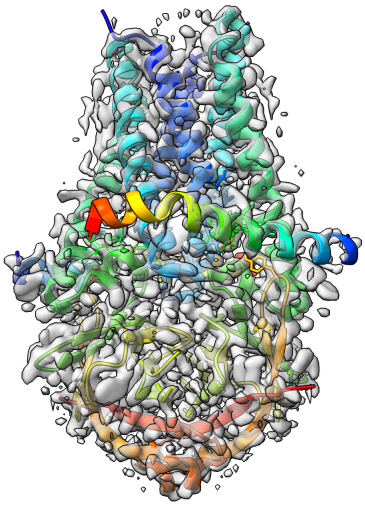


Figure 5. Target 203 (EMD-22898) map-model overlay at the recommended contour 0.7 (10.3 σ) with **T0203EM004_1** (ours)

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 map for ligand was assessed by the Q-score [23] and the Z-scores. The Q-score measures how similar map values around an atom are to a Gaussian-like function we would see if the atom were well resolved. Q-score wascalculated 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 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 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 ligand for Target 202, 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 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 in 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 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

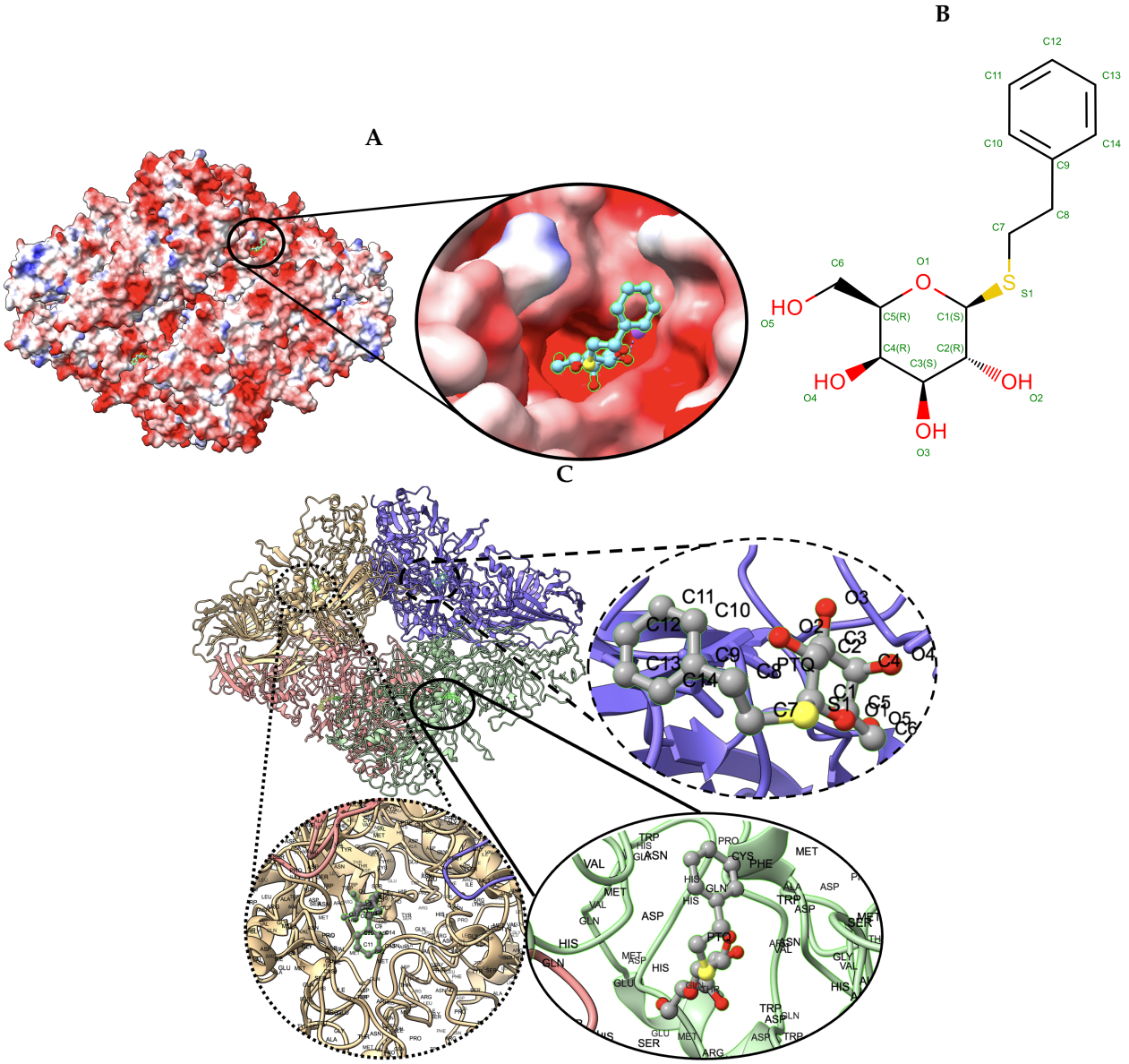


Figure 6. Target 201. **(A) T0201EM004_1** (ours) docked by Target 201 (EMD-7770) and visualized with electrostatic potential surface generated in UCSF Chimera. **(B)** Ligand PTQ, image extracted from Protein Data Bank (PDB). **(C)** Protein-ligand interactions in **T0201EM004_1** (ours) model. Chains 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.

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 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 deep learning approach in modeling enables us to predict the protein structure directly from cryo-EM map within minutes, making the approach highly useful on both prediction accuracy and time.

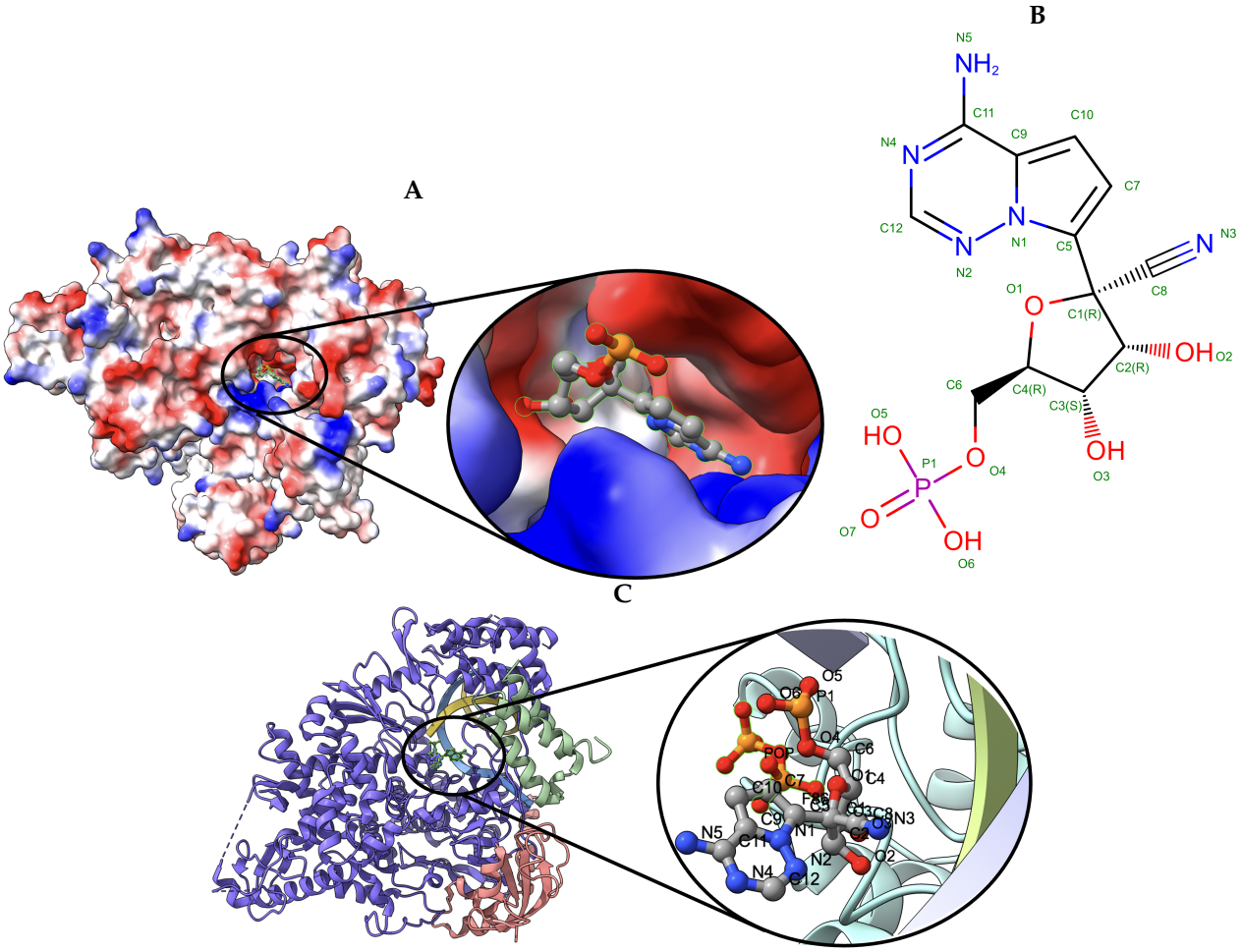


Figure 7. Target 202. **(A) T0202EM0004_1** (ours) docked by Target 202 (EMD-30210) and visualized with electrostatic potential surface generated in UCSF Chimera. **(B)** Ligand F86, image extracted from Protein Data Bank (PDB). **(C)** Protein-ligand interactions in **T0202EM004_1** (ours) model. Chains 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)

Table 6. Evaluation of Target 203: SARS-CoV-2 ORF3a putative ion channel in nanodisc on Q-score 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.

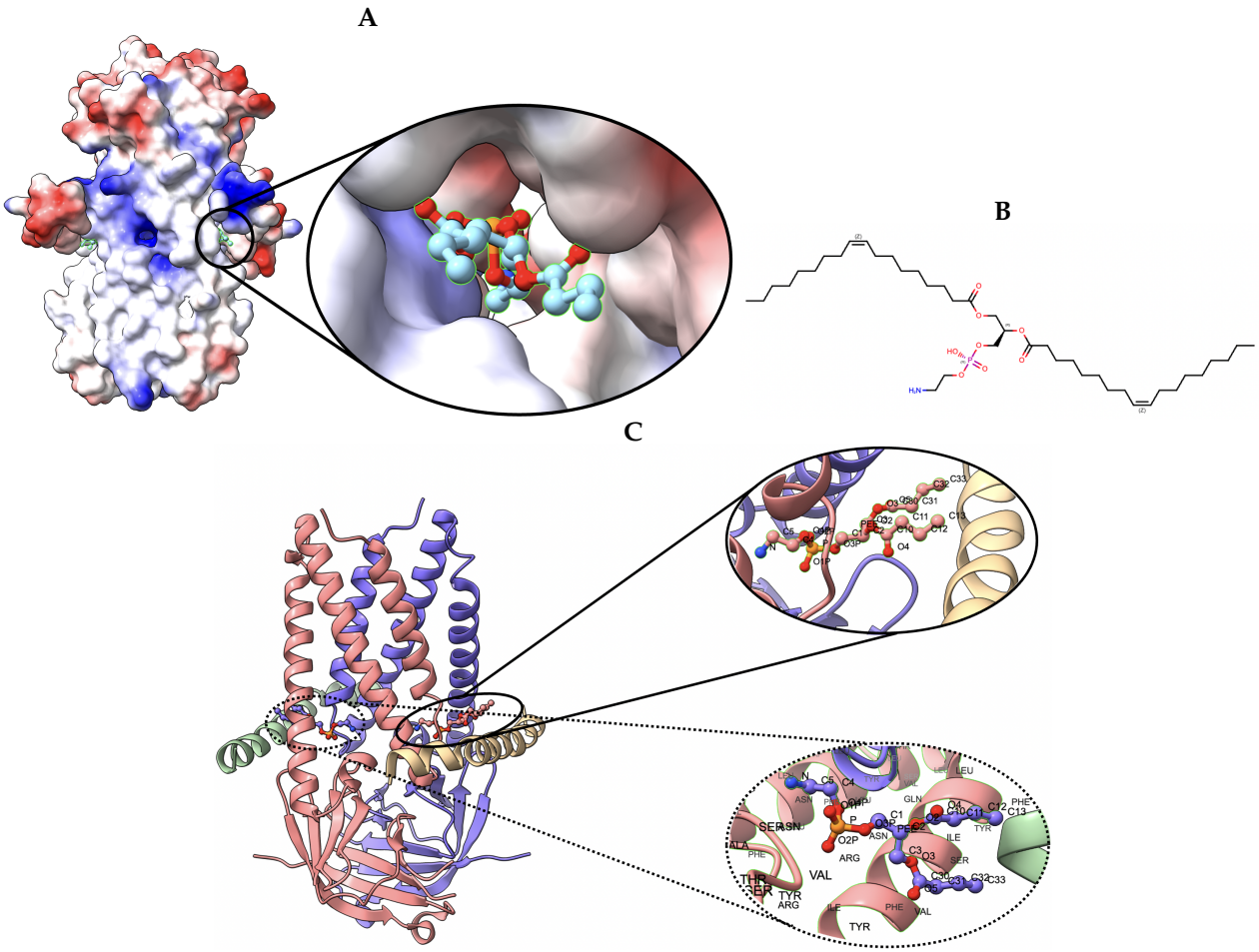


Figure 8. Target 203. **(A) T0203EM0004_1** (ours) docked by Target 203 (EMD-22898) and visualized with electrostatic potential surface generated in UCSF Chimera. **(B)** Ligand PEE, image extracted from Protein Data Bank (PDB). **(C)** Protein-ligand interactions in **T0203EM0004_1** (our) model. Chains 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 ligands name (PEE)

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.

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 structure 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

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.

compatible with geometric and stereochemical restraints of proteins from cryo-EM density maps and reference structures.

4. Conclusion and Future Work

In this work, we demonstrate that the deep learning prediction of protein structure 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 SE(3)-equivariant Transformer network [24–27] 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 deep learning approach to prediction of 3D structures of protein-ligand complexes leveraging cryo-EM and other related data is a promising avenue to accelerate the advancement of the protein-ligand interaction study [28,29]. With the proliferation of cryo-EM maps being deposited in EMDataResource database, the use of deep learning-based methods can help to determine the structure of the protein-ligand complexes rapidly and ultimately helps to expedite the drug discovery process.

Author Contributions: Conceptualization, JC; methodology, JC and NG; experiments and data collection, NG; data analysis, NG and JC; writing—original draft preparation, NG and JC; writing—review and editing, JC and NG; visualization, NG; supervision, JC; project administration, JC and NG; funding acquisition, JC. 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).

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

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

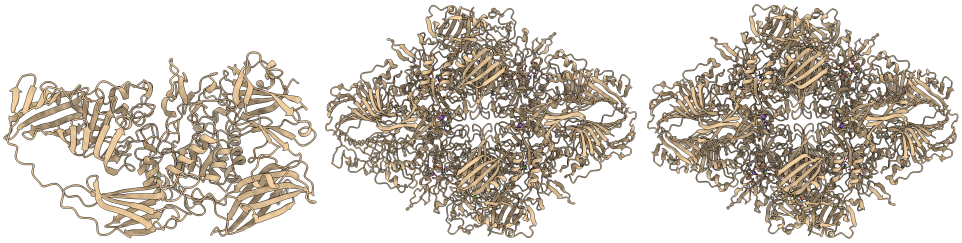
References

1. Cressey, D., Callaway, E. (2017). Cryo-electron microscopy wins chemistry Nobel. Nature, 550(7675)
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, Jerome, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. "AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings." Journal of Chemical Information and Modeling 61, no. 8 (2021): 3891-3898.
5. Morris, Connor J., and Dennis Della Corte. "Using molecular docking and molecular dynamics to investigate protein-ligand interactions." Modern Physics Letters B 35, no. 08 (2021): 2130002.
6. Fan, J., Fu, A., Zhang, L. (2019). Progress in molecular docking. Quantitative Biology, 7(2), 83-89
7. Stanzione, F., Giangreco, I., Cole, J. C. (2021). Use of molecular docking computational tools in drug discovery. Progress in Medicinal Chemistry, 60, 273-343
8. Sana Mansoor, Sammia Shahid, Kinza Ashiq, Norah Alwadai, Mohsin Javed, Shahid Iqbal, Urooj Fatima, Sabah Zaman, Muhammad Nazim Sarwar, Fwzah H. Alshammari, Eslam B. Elkaeed, Nasser S. Awwad, Hala A. Ibrahium, Controlled growth of nanocomposite thin layer based on Zn-Doped MgO nanoparticles through Sol-Gel technique for biosensor applications, Inorganic Chemistry Communications, Volume 142, 2022, 109702, ISSN 1387-7003, https://doi.org/10.1016/j.inoche.2022.109702.
9. Shahid Sammia, Ejaz Anam, Javed Mohsin, Mansoor Sana, Iqbal Shahid, Elkaeed Eslam B., Alzhrani Rami M., Alsaab Hashem O., Awwad Nasser S., Ibrahium Hala A., Fatima Urooj, Zaman Sabah, Nazim Sarwar Muhammad, The Anti-Inflammatory and Free Radical Scavenging Activities of Bio-Inspired Nano Magnesium Oxide, Frontiers in Materials, 9, 2022, 10.3389/fmats.2022.875163
10. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.
11. Giri, N., Roy, R. S., Cheng, J. (2022). Deep learning for reconstructing protein structures from cryo-EM density maps: recent advances and future directions. arXiv preprint arXiv:2209.08171.
12. Kern, D. M., Sorum, B., Mali, S. S., Hoel, C. M., Sridharan, S., Remis, J. P., ... Brohawn, S. G. (2021). Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. Nature structural molecular biology, 28(7), 573-582.
13. Pfab, J., Phan, N. M., Si, D. (2021). DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes. Proceedings of the National Academy of Sciences, 118(2), e2017525118.
14. Ronneberger, O., Fischer, P., Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.
15. "Protein Data Bank: the single global archive for 3D macromolecular structure data." Nucleic acids research 47, no. D1 (2019): D520-D528.
16. Catherine L. Lawson, Ardan Patwardhan, Matthew L. Baker, Corey Hryc, Eduardo Sanz Garcia, Brian P. Hudson, Ingvar Lagerstedt, Steven J. Ludtke, Grigore Pintilie, Raul Sala, John D. Westbrook, Helen M. Berman, Gerard J. Kleywegt, Wah Chiu, EMDataBank unified data resource for 3DEM, Nucleic Acids Research, Volume 44, Issue D1, 4 January 2016, Pages D396–D403, https://doi.org/10.1093/nar/gkv1126
17. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry, 25(13), 1605-1612
18. Krivov, G. G., Shapovalov, M. V., Dunbrack Jr, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins: Structure, Function, and Bioinformatics, 77(4), 778-795
19. Chaudhury, S., Lyskov, S., Gray, J. J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics, 26(5), 689-691.
20. Bartesaghi, A., Aguerrebere, C., Falconieri, V., Banerjee, S., Earl, L. A., Zhu, X., ... Subramaniam, S. (2018). Atomic resolution cryo-EM structure of -galactosidase. Structure, 26(6), 848-856.
21. Yin, W., Mao, C., Luan, X., Shen, D. D., Shen, Q., Su, H., ... Xu, H. E. (2020). Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. Science, 368(6498), 1499-1504.
22. Kern, D. M., Sorum, B., Mali, S. S., Hoel, C. M., Sridharan, S., Remis, J. P., ... Brohawn, S. G. (2021). Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. Nature structural molecular biology, 28(7), 573-582.

23. Krissinel, E., Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallographica Section D: Biological Crystallography, 60(12), 2256-2268.
24. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science, 373(6557), 871-876.
25. Fuchs, F., Worrall, D., Fischer, V., Welling, M. (2020). Se (3)-transformers: 3d roto-translation equivariant attention networks. Advances in Neural Information Processing Systems, 33, 1970-1981.
26. Morehead, A., Chen, X., Wu, T., Liu, J., Cheng, J. (2022). EGR: Equivariant Graph Refinement and Assessment of 3D Protein Complex Structures. arXiv preprint arXiv:2205.10390.
27. Wu, T., Cheng, J. (2022). Atomic protein structure refinement using all-atom graph representations and SE (3)-equivariant graph neural networks. bioRxiv
28. Dhakal, A., McKay, C., Tanner, J. J., Cheng, J. (2022). Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. Briefings in Bioinformatics, 23(1), bbab476.
29. Chang, Luca, Fengbin Wang, Kiernan Connolly, Hanze Meng, Zhangli Su, Virginija Cvirkaite-Krupovic, Mart Krupovic, Edward H. Egelman, and Dong Si. "DeepTracer ID: De Novo Protein Identification from Cryo-EM Maps." bioRxiv (2022).
30. Corso, G., Stärk, H., Jing, B., Barzilay, R., Jaakkola, T. (2022). DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. arXiv preprint arXiv:2210.01776.
31. Stärk, Hannes, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. "Equibind: Geometric deep learning for drug binding structure prediction." In International Conference on Machine Learning, pp. 20503-20521. PMLR, 2022.

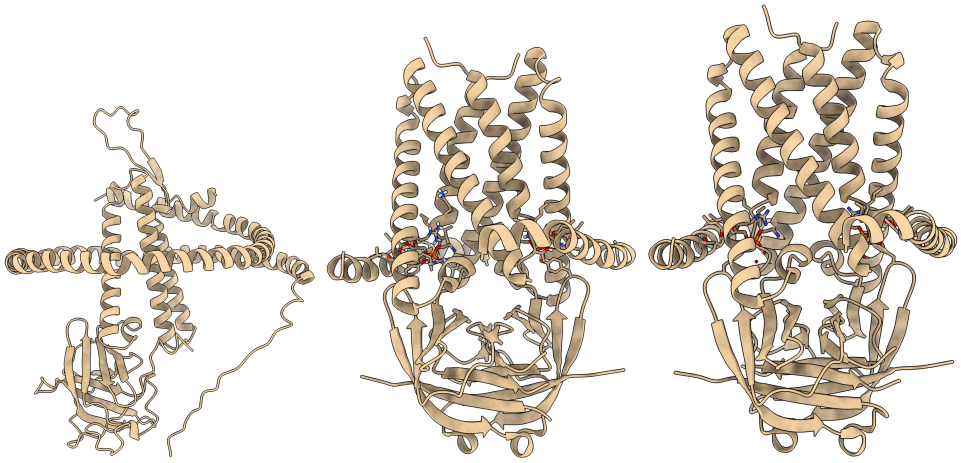
Appendix E Sequence based modeling: An approach for predicting protein structure

DeepProLigand uses DeepTracer 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. 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, which is the main focus of the 2021 Ligand Challenge. Figure A10, and A11 shows the comparison of structures predicted by AlphaFold and DeepProLigand with the PDB deposited structure.



(a) AlphaFold Predicted **(b)** DeepProLigand Predicted **(c)** PDB Deposited

Figure A10. The target T201 is of EMD: 7770, PDB ID: 6CVM.



(a) AlphaFold Predicted **(b)** DeepProLigand Predicted **(c)** PDB Deposited

Figure A11. The target T203 is of EMD: 22898, PDB ID: 7KJR.