Assessment of AI-based Protein Structure Prediction for the NLRP3 Target

Jian Yin*, Jialin Yu, Weiren Cui, Junkun Lei, Wenhua Chen, Alexander L. Satz, Yifan Zhou, Hua Feng, Jason Deng, Wenji Su, and Letian Kuai

*Corresponding author: yin_jian@wuxiapptec.com

Address: WuXi AppTec (Shanghai) Co., Ltd. 240 Hedan Road, Shanghai 200131, China

Abstract

The recent success of AlphaFold and RoseTTAFold has demonstrated the values of AI methods in predicting highly accurate protein structures. Despite the advances, their roles in the context of small-molecule drug discovery need to be thoroughly explored. In this study, we evaluated the possibility whether the AI-based models can lead to reliable 3D structures of protein-ligand complexes. With the AI-generated protein structure, we were able to confidently predict the binding modes of small-molecule inhibitors for NLRP3, a challenging protein target in terms of obtaining the 3D model both experimentally and computationally. We therefore concluded that through careful evaluation, AI-predicted protein structures can be trusted and useful for small-molecule drug discovery projects.

Introduction

High-resolution X-ray crystallographic or NMR structure of proteins and protein-ligand complexes are critical for understanding the interactions between small-molecule drug candidates and their targets, and therefore are pivotal in the drug discovery process.[1] However, due to the limitations of experimental techniques, there are still a large fraction of protein structures that cannot be solved. Studies on novel drug targets are often hampered by a lack of protein structures with sufficient resolution and completeness.

The AI methods for protein structure predictions, including the AlphaFold (AF) [2] and RoseTTAFold (RF) [3] , have attracted tremendous interest for their ability to accurately predict protein structures, independent of structural templates with high sequence similarity. Albeit the success in recent evaluations of single protein structure predictions, their applications in small-molecule drug discovery have not been systematically investigated yet. For instance, protein targets may undergo induced-fit conformational changes due to the binding of small molecules[4]. Predictions might become even more challenging when the protein comprises multiple domains or subdomains. [5]

In the current study, we explored whether the AI predicted protein structures can be employed to predict the binding modes of known NLRP3 inhibitors. The target NLRP3 (NOD-like receptor family, pyrin domain-containing protein 3) [6] is a protein linked to many chronic inflammatory human diseases, such as atherosclerosis, Alzheimer's disease, and

nonalcoholic steatohepatitis. It is part of the NLRP3 inflammasome that, when activated, triggers the release of proinflammatory cytokines IL-1β and IL-18, and thus lead to an inflammatory form of cell death[7] . Despite the considerable interest in this target, a lack of structural information hinders the development of small-molecule inhibitors. We therefore carefully evaluated the quality of the AI predicted models, and assessed the plausibility of binding poses generated based on those structures. We demonstrated that the AI models combined with molecular modeling are valuable for exploring the mechanisms of the NLRP3 compounds and further optimization steps.

Results

Evaluation of the experimental and predicted structures of NLRP3

The human NLRP3 contains PYD and FISNA subdomains (Residues 3-218), a curved leucine-rich repeat domain (LRR; 652-1036) and a NACHT domain comprising the nucleotide-binding domain (NBD; 219-372), helical domain 1 (HD1; 373-434), winged helix domain (WHD; 435-541) and helical domain 2 (HD2; 542-651) (Figure 1). The only globally complete structure of this target available on PDB to date is a cryo-electron microscopy (cryo-EM) structure of inactive human NLRP3 bound to NEK7 (PDB code: 6NPY), a mitotic kinase which functions as a mediator of NLRP3 activation (Figure 1a). There is another cryo-EM structure of NLRP3 bound to drug MCC950 (Figure 1d) reported very recently, but it has not been deposited to PDB until this present work. [8]

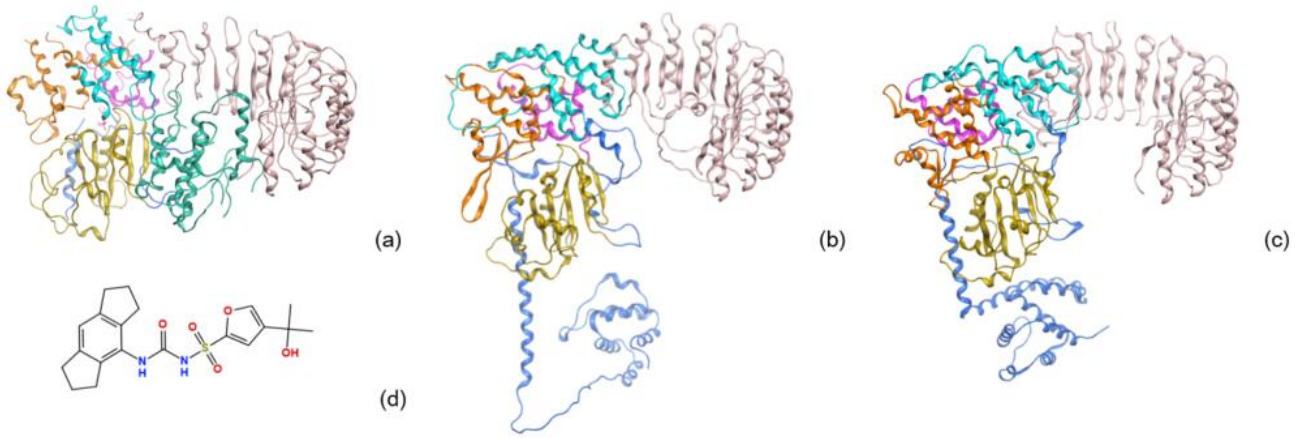

Figure 1. (a) Previously reported cryo-EM structure of NLRP3 complexed with NEK7 kinase (PDB code: 6NPY), and the percentage of missing residues in subdomains; structures of NLRP3 predicted by (b) AlphaFold and (c) RoseTTAFold. Gold: NBD; Purple: HD1; Green: NEK7; Orange: WHD; Cyan: HD2; Pink: LRR; Blue: others. (d) The 2D structure of MCC950, which was reported as a selective NLRP3 inhibitor.

We evaluated the suitability of NLRP3–NEK7 complex structure for small-molecule drug design based on three criteria: resolution, ratio of missing residues especially at the regions near the reported binding site of inhibitors, and the likelihood of conformational change upon the binding of drugs. Particularly, the resolution of this structure is 3.8 Å, which is considered as common for cryo-EM structures, but is relative low compared with x-ray crystal structures. It is also noteworthy that a significant portion of the residues (31.8%) are missing in the HD2 region, which is near of the ADP binding site, and also part of the reported binding pocket of MCC950. Additionally, the complexation of the NEK7 substrate may result in a conformation of NLRP3 that is distinct from that bound to small-molecule inhibitors.

Given these potential concerns on the cryo-EM structure of the NLRP3–NEK7 complex, we

obtained predicted structures of the single NLRP3 protein using both AF and RF. Five models were generated by each program locally installed, and the pLDDT (predicted Local Distance Difference Test) scores[2] were calculated for every model and its sub-domains to quantify the confidence level in the prediction of structures, which were interpreted as follows: >90: high accuracy; 70–90: moderate to good; 50-70: low confidence; <50: disordered or unstructured. It was observed that the AF and RF models have comparable overall confidence level, with an average value of 79.9 and a standard deviation of 0.9 for AlphaFold, compared with 78.3 ± 1.2 for RoseTTAFold. It is also noteworthy that all pLDDT scores of the regions critical for drug binding are above 70, which suggest that the protein was predicted confidently at the individual domain level.

Table 1. Predicted Local Distance Difference Test (pLDDT) and predicted templated modelling (pTM) scores of the AI generated NLRP3 structures.

Domain	Model1	Model2	Model3	Model4	Model5
			AlphaFold		
NBD	83.3	85.3	86.8	83.5	85.7
HD1	82.8	84.9	85.4	82.1	84.3
WHD	70.2	71.5	74.2	71.0	71.2
HD2	76.8	78.6	79.4	77.2	79.5
LRR	84.6	85.0	86.4	85.3	84.7
Full-length	79.1 (0.58)	80.3 (0.67)	81.4 (0.71)	79.1 (0.70)	79.8 (0.71)
			RoseTTAFold		
NBD	83.4	81.6	82.5	80.7	79.5
HD1	83.5	82.3	81.3	80.7	80.3
WHD	80.0	78.4	76.4	78.7	77.3
HD2	75.2	73.1	70.2	73.0	71.1
LRR	82.3	82.9	83.1	81.9	81.8
Full-length	79.7	79.4	77.9	77.6	76.8

*pLDDT scores are reported on a 0 -100 scale. Values higher than 90: high accuracy; 70–90: moderate to good; 50-70: low confidence; <50: disordered or unstructured. pTM scores (numbers in the parentheses) are on a 0-1 scale. Numbers higher than 0.7 indicates high accuracy.

On the other hand, for multi-domain or -subdomain protein targets such as NLRP3, the predicted templated modelling (pTM) scores [2] may be better indicators of the interdomain packing accuracy. The AF package allows us to obtain the pTM scores from the fine-tuned pTM models (Table 1). The numbers near 0.7 may indicate risks of incorrect domain assembly and interdomain contacts to a modest extent, which should be of particular concern in the present test case, since the reported binding site of the known inhibitor MCC950 happens to located at the interfaces of several subdomains.

Next, we further computed the RMSD values of the predicted models with respect to the experimental structure. As shown in Table 2, the RMSD values of the HD2 and LRR domains are significantly higher than these of NBD, HD1 and WHD domains, using both AF and RF. Meanwhile, the RF models showed lower RMSD values than the AF ones, except for model 4. Nevertheless, the RMSD values of the RF models are still above 10 Å. Given the substantial dissimilarities observed between the experimental and computational structures, especially in the HD2 region near the binding site of inhibitors, the applicability of the modelled structures in hit discovery were carefully assessed, and the results are described in the following sections.

Table 2. RMSD values of the AI-predicted NLRP3 structures compared to the experimental structure of NLRP3 in 6NPY. (Unit: Å)

Domain	Model1	Model2	Model3	Model4	Model5	# of aligned atoms
			AlphaFold			
NBD	4.4 (3.7)	4.4 (3.7)	4.4 (3.7)	4.4 (3.7)	4.4 (3.7)	1178 (154)
HD1	2.7 (1.8)	2.6 (1.7)	3.0 (2.2)	2.6 (1.6)	2.7 (1.8)	505 (62)
WHD	6.0 (5.3)	6.2 (5.5)	6.1 (5.4)	5.9 (5.2)	6.1 (5.4)	780 (98)
HD2	18.5 (18.3)	18.6 (18.4)	18.5 (18.3)	18.5 (18.3)	18.4 (18.3)	587 (75)
LRR	12.1 (11.8)	12.5 (12.1)	13.5 (13.1)	11.4 (11.2)	12.8 (12.5)	2713 (359)
Full-length	14.7 (14.4)	14.8 (14.4)	14.6 (14.2)	10.5 (10.2)	14.6 (14.2)	6192 (798)
			RoseTTAFold			
NBD	4.2 (3.5)	4.3 (3.6)	4.3 (3.6)	4.2 (3.6)	4.1 (3.5)	
HD1	2.4 (1.4)	2.6 (1.5)	2.5 (1.5)	2.5 (1.5)	2.5 (1.4)	
WHD	3.2 (2.3)	3.2 (2.3)	3.3 (2.4)	3.2 (2.3)	3.2 (2.4)	*Same as AlphaFold*
HD2	16.3 (15.8)	17.6 (17.2)	17.6 (17.6)	16.4 (16.0)	16.5 (16.0)	
LRR	7.9 (7.4)	8.2 (7.7)	8.2 (7.7)	8.2 (7.7)	8.1 (7.5)	
Full-length	10.7 (10.2)	10.5 (9.9)	10.7 (10.2)	10.5 (9.9)	10.5 (9.9)	

*Numbers before and in the parentheses are the RMSD values calculated using coordinates of all heavy atoms and alpha carbons, respectively. Note that the RMSD of the subunits were computed with each set of the predicted and experimental sub-domains superposed individually.

Binding pocket Assessment for compound MCC950

MCC950 has been long identified as a selective NLRP3 inhibitor that blocks NLRP3 inflammasome activation[9, 10], but its mechanism had remained obscure until very recently. A complex structure of MCC950 bound to NLRP3 in the inactive state was recently solved using cryo-EM [8], yet the PDB structure had not been released by the time this present work was completed, which makes a good case for testing the AI-based protein prediction

techniques.

It was reported that MCC950 was bound to a cleft formed by subdomains NBD, HD1, WHD, HD2 and LRR[8], which was separated from the ATP binding site by the Walker A motif (GAAGIGKTIL)[11]. The residues adjacent to MCC950 include F575 and R578 in the HD2 subdomain, and A228 in the Walker A motif. Therefore, we measured the distance between the alpha-carbon atoms of R578 and A228 in each structure, as means to assess the formation of binding pocket for MCC950. Particularly, F575 and R578 in the experimental structure of the NLRP3-NEK7 complex were located in the solvent-exposing region, which were relatively distinct from the ATP-binding site (Figure 2a). The distance between R578 and A228 was measured as 20.3 Å. In the AF models, the R578 and A228 residues were placed facing other each, yet the distance between them was around 20 Å in each AF structure except for model 4 with a value of 14.5 Å, since the HD2 and NBD subdomains were not tightly packed together (Figure 2b and 2c). Note that AF model 4 also showed the smallest RMSD relative to the experimental structure (Table 2). By comparison, the HD2 and NBD subdomains in the RF models were more densely packed, and the R578 and A228 residues at the interfaces were close enough (~10 Å) to accommodate the inhibitors well (Figure 2d).

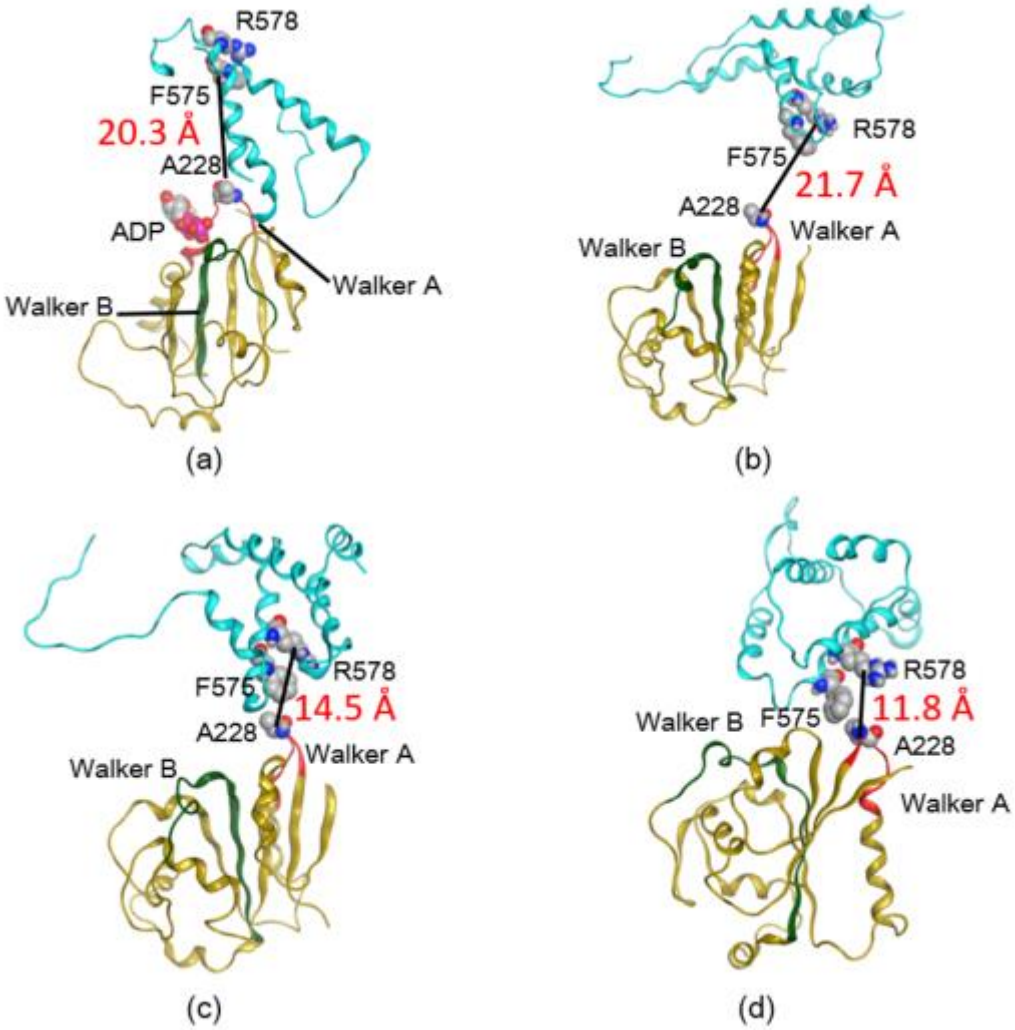

Figure 2. The positions of three key residues that constitute the binding pocket of MCC950 in (a) the NLRP3-NEK7 complex structure (PDB code: 6NPY), (b) model 1 generated by AlphaFold, (c) model 4 by AlphaFold; and (d) model 1 by RoseTTAFold. The distance between the alpha-Carbon atoms of R578 and A228 was 20.3 Å, 21.7 Å, 14.5 Å and 11.8 Å in (a), (b), (c) and (d), respectively. Red: the Walker A motif (GAAGIGKTIL); Green: the Walker B motif previously suggested as the binding region of MCC950 (RILFMDGFDELQGAFDEHI) [11]; Gold: NBD; Cyan: HD2. Other domains omitted for clarity.

Binding poses of MCC950 and CY-09

To further evaluate the feasibility of AI-predicted structures for drug discovery, we docked MCC950 into the reported binding cavity of NLRP3, using the structures of both AF and RF models. It was found that only RF model 1 provided the most plausible binding interactions for MCC950 (Figure 3a). In particular, one of the oxygen atoms in the amide moiety of MCC950 formed a hydrogen bond with R578, and another oxygen in the sulfonyl group hydrogen bonded with one of the alanine residues residing in the Walker A motif. After a 100 ns molecular dynamics (MD) simulation without applying any backbone or sidechain restraints, the complexation between the protein and the compound were refined, and a few remarkable changes in binding modes were observed. First, the amide oxygen in MCC950 engaged bidentate hydrogen bonding interactions with R578, which were further stabilized by the salt bridge between R578 and E629. Secondly, besides the hydrogen bond formed with A228 in the Walker A motif, the sulfonyl oxygen of the ligand also formed an intramolecular H-bond with the amide nitrogen. Lastly, the terminal hydroxyl group of the compound acted as an H-bond donor when interacting with D662, which has not been reported before. It is noteworthy that an H-bond network involving the surrounding Y632, S658, T659 residues may be crucial for the stabilization of this interaction.

The docking pose of another known NLRP3 inhibitor, CY-09[12] is also reported here (Figure 4), showing similar binding modes with MCC950. We speculate that an alternative binding mode is also possible in which the thiorhodanine ring in CY-09 flips 180 degrees without interrupting the H-bonding interactions with A228 and R5778, and there is a certain possibility

that the terminal carboxylate group would engage H-bonds with the D662-centered sub pocket rich of residues with polar side chains.

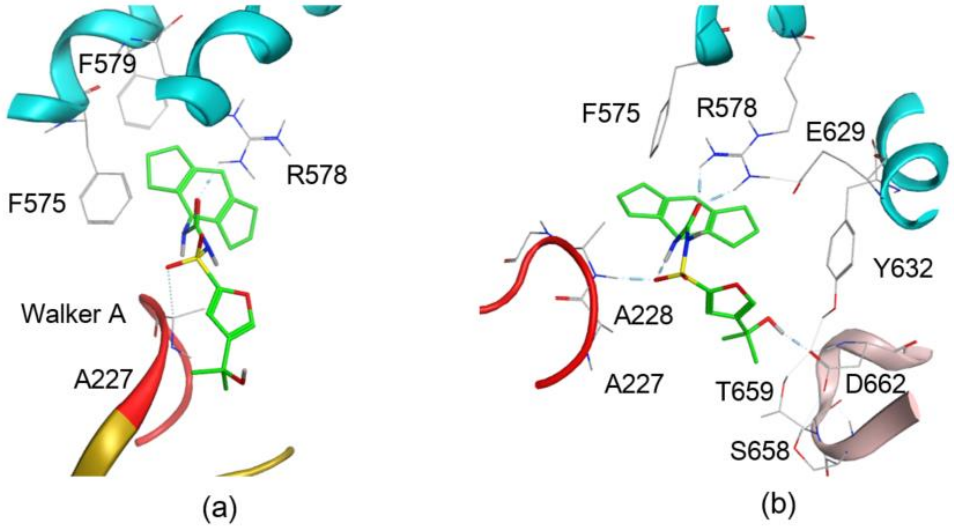

Figure 3. Binding poses of MCC950 based on RF model 1 (a) before and (b) after the refinement using molecular dynamics simulation. Green: MCC950; Red: the Walker A motif; Gold: NBD; Cyan: HD2; Pink: LRR; other domains omitted for clarity.

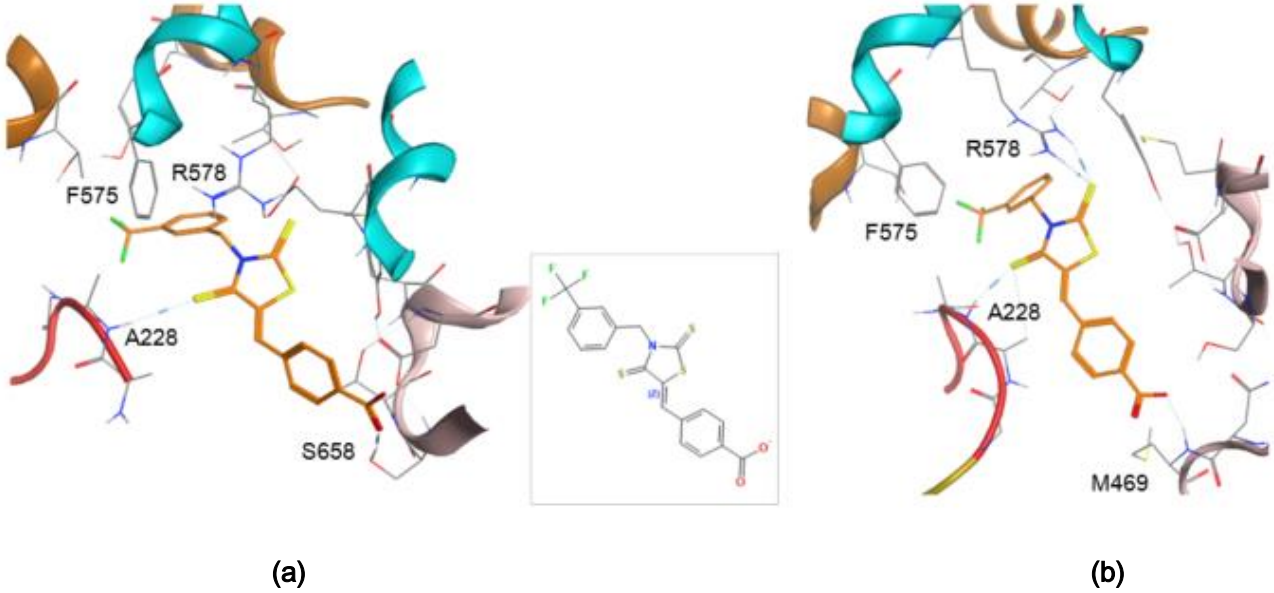

Figure 4. Binding poses of CY-09 based on the protein structure of refined RF model 1 (a) before and (b) after the refinement using molecular dynamics simulation. Red: the Walker A motif; Orange: WHD; Cyan: HD2; Pink: LRR; other domains omitted for clarity.

Discussion

Despite the use of protein structure templates during both training and predicting procedures, AlphaFold and RoseTTAFold are intrinsically more advanced than the conventional template-based methods such as homology modelling and threading. With the neural network architecture, the AI methods are able to learn deeply from the coevolution of protein sequences, and detect the underlying correlations between pairs of amino acids through the construction of multiple sequence alignment (MSA) and pair representation, instead of merely patching the templates together. Therefore, the AI methods can accurately predict protein structures using templates with sequence identity and structural coverage lower than 30%. Indeed, this current study demonstrated that the structures predicted by the AI methods did not heavily rely on the only existing experimental structure, and could be a good complementary method for experimental techniques including X-ray crystallography, NMR and cryo-EM technique.

Besides the possible conformational change caused by the bound molecules, determining the binding modes of the small molecule inhibitors based on AI structures in the present work is a particularly challenging case, for the NLRP3 protein is composed of multiple subdomains, and the putative binding pocket is located at interdomain regions. It was shown that both AlphaFold and RoseTTAFold can make predictions confidently at the subdomain level. Nevertheless, RoseTTAFold performed better on domain assemblies in the current case, and generated a structure that can well reproduce the reported binding pocket after the refinement

using molecular dynamics. Our study suggested that one should be careful when using AI protein structure prediction tools for drug targets comprising of two or more structural domains, especially when the drug binding site is formed at the interfaces of subunits. In practice, AI predicted models that can provide connection with the results of affinity selection experiment may be regarded as more reliable. The effort of validating AI generated protein structures based on affinity selection data are currently ongoing.

References

[1]. Marrone, T.; Briggs, J. M.; McCammon, J. A., Structure-based drug design: computational advances. *Annu. Rev. Pharmacol. Toxicol.* **1997,** *37*, 71-90.

[2]. Jumper, J. M.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D. A.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021,** *596*, 583 - 589.

[3]. Baek, M.; Dimaio, F.; Anishchenko, I. V.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A. M.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D., Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021,** *373*, 871 - 876.

[4]. Sherman, W.; Day, T.; Jacobson, M.; Friesner, R. A.; Farid, R., Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006,** *49 2*, 534-53.

[5]. Schaeffer, R. D.; Kinch, L.; Kryshtafovych, A.; Grishin, N. V., Assessment of domain interactions in the fourteenth round of the Critical Assessment of Structure Prediction. *Proteins: Struct. Funct. Bioinform.*

[6]. Sharif, H.; Wang, L.; Wang, W. L.; Magupalli, V. G.; Andreeva, L.; Qiao, Q.; Hauenstein, A. V.; Wu, Z.; Núñez, G.; Mao, Y.; Wu, H., Structural mechanism for NEK7-licensed activation of NLRP3 inflammasome. *Nature* **2019**, 1-6.

[7]. He, Y.-q.; Hara, H.; Núñez, G., Mechanism and Regulation of NLRP3 Inflammasome Activation. *Trends Biochem. Sci.* **2016,** *41 12*, 1012-1021.

[8]. Hochheiser, I. V.; Pilsl, M.; Hagelueken, G.; Moecking, J.; Marleaux, M.; Brinkschulte, R.; Latz, E.; Engel, C.; Geyer, M., Cryo-EM structure of the NLRP3 decamer bound to the cytokine release inhibitory drug CRID3. *bioRxiv* **2021**.

[9]. Tapia-Abellán, A.; Angosto-Bazarra, D.; Martínez-Banaclocha, H.; de Torre-Minguela, C.; Cerón-Carrasco, J. P.; Péreš Sánchez, H.; Aróstegui, J. I.; Pelegrín, P., MCC950 closes the active conformation of NLRP3 to an inactive state. *Nat. Chem. Biol.* **2019,** *15*, 560 - 564.

[10]. Coll, R. C.; Robertson, A. A. B.; Chae, J. J.; Higgins, S. C.; Muñoz-Planillo, R.; Inserra, M. C.; Vetter, I.; Dungan, L. S.; Monks, B. G.; Stutz, A.; Croker, D. E.; Butler, M. S.; Haneklaus, M.; Sutton, C. E.; Núñez, G.; Latz, E.; Kastner, D. L.; Mills, K. H. G.; Masters, S. L.; Schroder, K.; Cooper, M. A.; o Neill, L. A. J., A small-molecule inhibitor of the NLRP3 inflammasome for the treatment of inflammatory diseases. *Nat. Med.* **2015,** *21*, 248-255.

[11]. MacDonald, J. A.; Wijekoon, C. P.; Liao, K.-C.; Muruve, D. A., Biochemical and structural aspects of the ATp binding domain in inflammasom-forming human NLRP proteins. *IUBMB Life* **2013,** *65*.

[12]. Jiang, H.; He, H.; Chen, Y.; Huang, W.; Cheng, J.; Ye, J.; Wang, A.; Tao, J.; Wang, C.; Liu, Q.; Jin, T.; Jiang, W.; Deng, X.; Zhou, R., Identification of a selective and direct NLRP3 inhibitor to treat inflammatory disorders. *J. Exp. Med.* **2017,** *214* (11), 3219-3238.