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

Jian Yin*, Jialin Yu, Junkun Lei, Weiren Cui, 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. The structure we predicted was NLRP3, a challenging protein target in terms of obtaining the three-dimensional model both experimentally and computationally. The conformation of the binding pockets generated by the AI models were carefully characterized and compared with the experimental structures.

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 had hindered the development of small-molecule inhibitors. We therefore carefully evaluated the quality of the AI predicted models, and assessed the discrepancy between the binding pockets formed in AI models and the experimental structures.

Results

Evaluation of the AI-predicted models 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) (Fig. 1). For a long period of time, the only available structure that included the key subdomains of this protein was a cryo-electron microscopy (cryo-EM) model of human NLRP3 bound to NEK7, a mitotic kinase mediating NLRP3 activation (PDB code: 6NPY; Fig. 1a). A significant portion of the residues (31.8%) are missing in the HD2 region in this structure, which is near the ADP binding site, and also part of the reported binding pocket of MCC950. In addition, the complexation of the NEK7 substrate may result in a different conformation of NLRP3 than that bound to small-

molecule inhibitors. A cryo-EM structure of NLRP3 decamer bound to drug MCC950 (PDB code: 7PZC) was newly released on January 26, 2022. [8]Besides, a crystal structure with a MCC950-like inhibitor (PDB code: 7ALV) was released on October 20, 2021 [9]. But note that they had not yet been deposited to PDB when we made the structural predictions in the currently reported study.

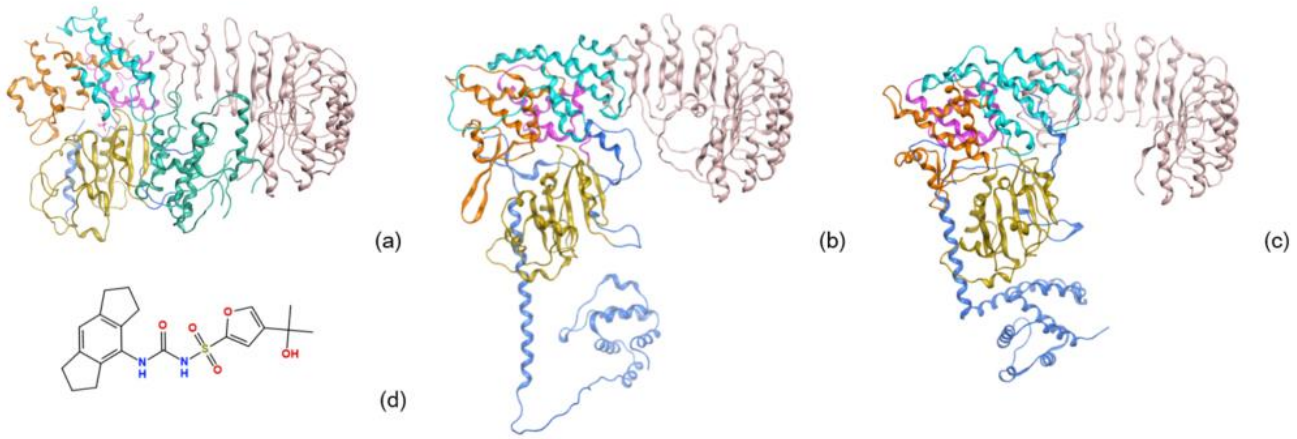


Figure 1. Previously reported cryo-EM structures of (a) NLRP3 complexed with NEK7 kinase (PDB code: 6NPY), and (b) NLRP3 bound to MCC950 (PDB code: 7PZC); the structures of NLRP3 predicted by (c) AlphaFold and (d) RoseTTAFold. Gold: NBD; Purple: HD1; Green: NEK7; Orange: WHD; Cyan: HD2; Pink: LRR; Blue: others. (d) The 2D structure of MCC950, which was reported as a selective NLRP3 inhibitor.

Given the difficulties in obtaining the structure of the multi-subdomain NLRP3 protein both experimentally and theoretically, we designed the prediction of NLRP3 protein conformation using both AF and RF as a test case. 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 computed the RMSD values of the predicted models with respect to one of the monomers in the experimental structure of the NLRP3 decamer. As shown in Table 2, the RMSD values of the LRR subdomain structures predicted using AF and RF were significantly higher than those of other subdomains. Meanwhile, the structure predicted by the RF model performs better on WHD, while the structure predicted by the AF model is more similar to the experimental results on HD1. Nonetheless, all models have RMSD values above 10 Å. Given the significant differences observed between the experimental and computational structures, especially in the HD2 region near the inhibitor binding site, we carefully evaluated the applicability of the modeled structures in small molecule drug discovery, and the results are described in the following sections.

Table 2. RMSD values of the AI-predicted NLRP3 structures relative to one of the monomers in 7PZC (Unit: Å).

Domain	Model1	Model2	Model3	Model4	Model5	# of aligned atoms
			AlphaFold			
NBD	2.5 (1.7)	2.6 (1.7)	2.5 (1.7)	2.6 (1.8)	2.6 (1.8)	1227 (154)

HD1	2.4 (1.5)	2.4 (1.5)	2.7 (2.0)	2.2 (1.3)	2.4 (1.5)	505 (62)
WHD	5.7 (4.9)	5.9 (5.2)	5.8 (5.1)	5.7 (4.9)	5.8 (5.1)	873 (107)
HD2	4.4 (3.6)	4.4 (3.5)	4.1 (3.1)	4.0 (3.1)	4.2 (3.4)	921 (110)
LRR	11.1 (10.4)	11.8 (11.1)	9.5 (8.9)	4.4 (3.8)	11.5 (10.8)	2945 (385)
Full-length	25.2 (24.8)	26.6 (26.1)	18.8 (18.3)	10.6 (10.2)	17.4 (17.0)	8048 (1010)
			RoseTTAFold			
NBD	2.9 (1.9)	2.8 (1.9)	2.8 (1.9)	2.8 (1.9)	2.8 (1.9)	
HD1	2.3 (1.3)	2.4 (1.4)	2.4 (1.3)	2.4 (1.3)	2.3 (1.3)	
WHD	3.5 (2.1)	3.5 (2.1)	3.5 (2.1)	3.4 (2.0)	3.4 (2.1)	*Same as*
HD2	8.0 (7.5)	5.3 (4.3)	5.2 (4.2)	7.2 (6.6)	7.1 (6.5)	*AlphaFold*
LRR	10.9 (10.6)	10.9 (10.6)	11.0 (10.7)	10.9 (10.7)	11.1 (10.7)	
Full-length	13.0 (12.5)	14.8 (14.4)	23.8 (23.3)	21.9 (21.3)	21.8 (21.5)	

*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.

Characterization of the Binding pocket of compound MCC950

MCC950 has been long identified as a selective NLRP3 inhibitor that blocks NLRP3 inflammasome activation [10, 11] (Fig. 1d), but its mechanism had remained obscure until very recently. The newly reported experimental structure showed 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)[12]. The residues adjacent to MCC950 include F575, R578 and E629 in the HD2 subdomain, A228 in the Walker A motif of NBD, as well as I411 and Y443 in HD1 (Fig. 2 and 3a). Therefore, we measured the distances between the α-carbons of key residue pairs in the small molecule binding site of each AI model, as well as the distances between side chains, to assess the potential for the formation of the MCC950 binding pocket. In particular, the carbamide and sulfonyl group of

MCC950 engaged hydrogen bonds with A228 and R578, located in different subdomains across the binding pocket (Fig. 4a). As a result, the distance between A228 and R578 residues was considered as one of the key indicators for the formation of a suitable binding site.

The results showed that R578 and its adjacent F575 in the experimental structure of the NLRP3-NEK7 complex were located in the solvent-exposing region, relatively distant from the ATP-binding site (Fig. 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 (Fig. 2b and 2c). Note that AF model 4 also showed the smallest RMSD relative to the experimental structure (Table 2). In contrast, the HD2 and NBD subdomains in the RF models were more densely packed, and the R578 and A228 residues at the interfaces appeared to be close enough (~12 Å) to accommodate the inhibitors well (Fig. 2d). Indeed, the corresponding distance in the experimental structure 7PZC is 12.7 Å.

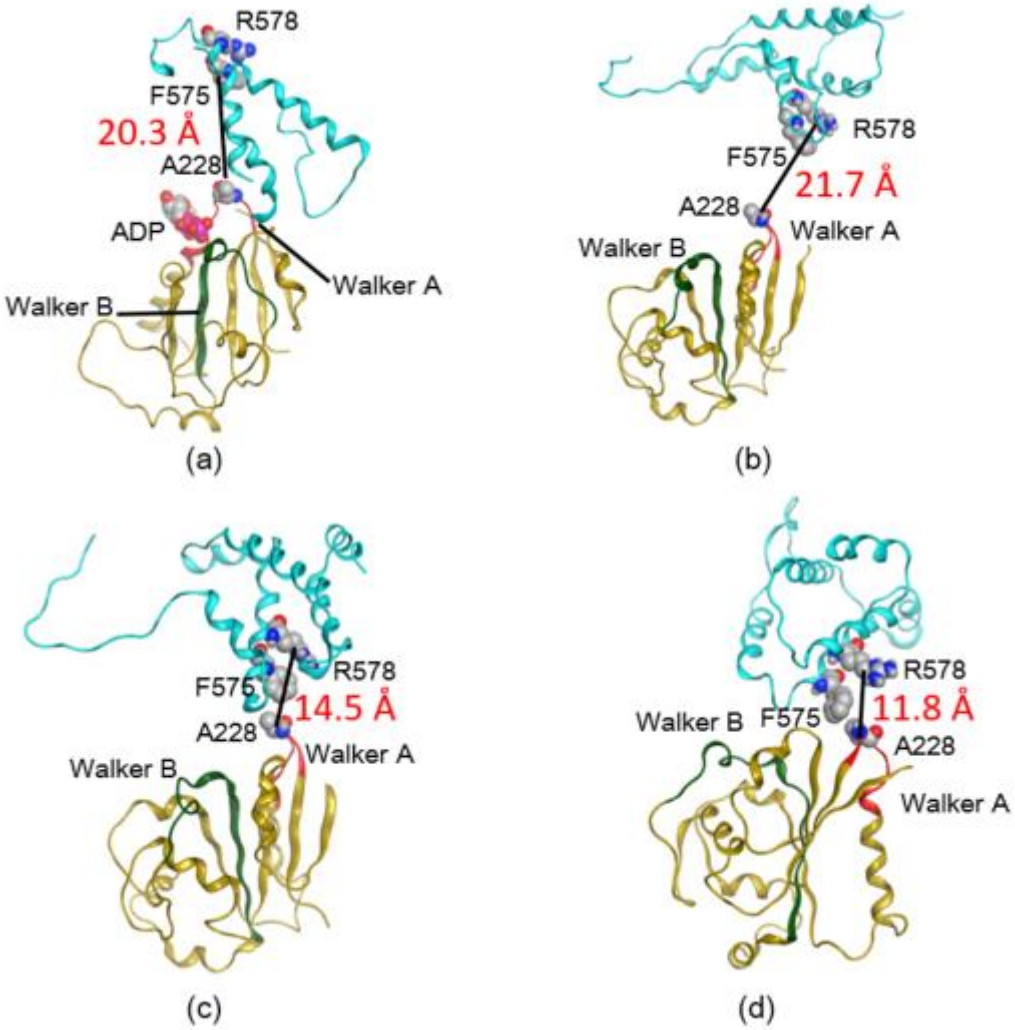
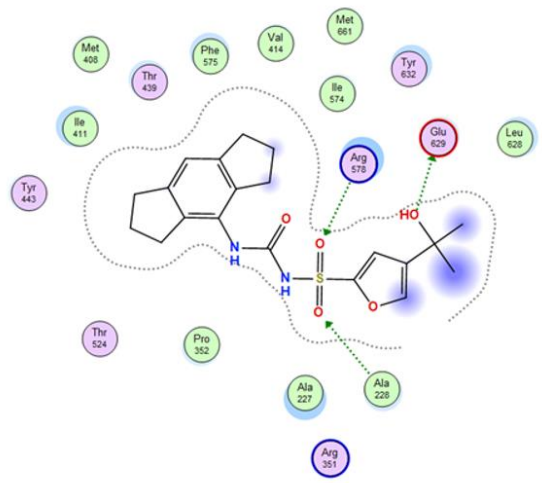


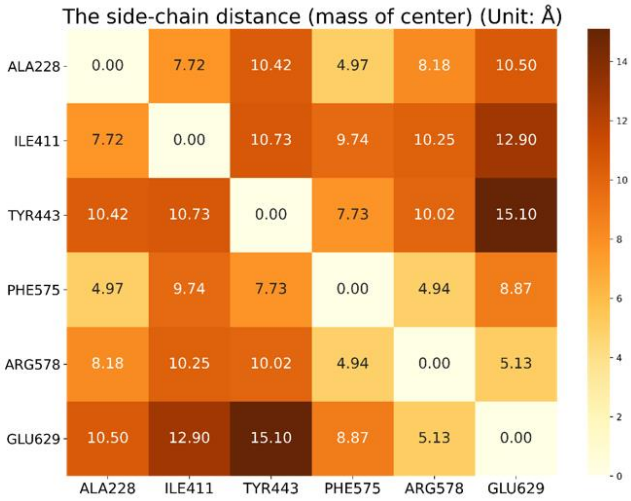
Figure 2. The positions of three key residues that constitute the binding pocket of MCC950 in (a) the NLRP3-NEK7 complex structure (PDB code: 6NPY), (b) model 1 generated by AlphaFold, (c) model 4 by AlphaFold; and (d) model 1 by RoseTTAFold. The distance between the α-carbon atoms of R578 and A228 was 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) [12]; Gold: NBD; Cyan: HD2. Other domains omitted for clarity.

In addition to the hydrogen bonds formed between the sulfonamide group of MCC950 and the protein, the tricyclic head group of the small molecule had hydrophobic interactions

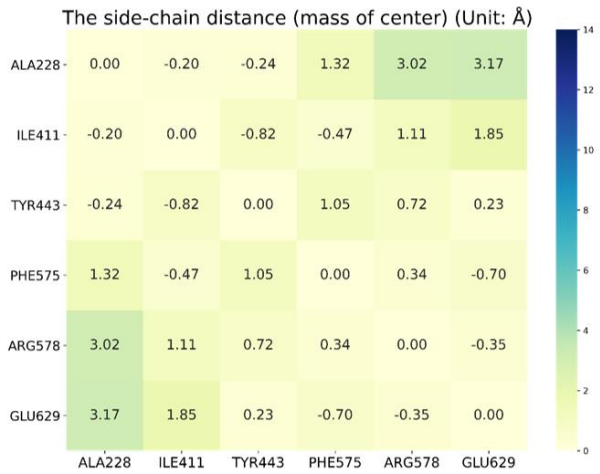
with a subpocket including protein residues I411, Y443, and F575. In addition, the terminal carboxylate group engaged H-bonds with G629 (Fig. 4a). We therefore compared the pairwise distances of key residues in the MCC950 binding pocket for all AF and RF models with the corresponding ones in 7PZC, as a finer measure of binding pocket structure predictions. It was observed that, on average, RF model 1 had the smallest discrepancy from the experimental results among all models (Fig. 3). AF model 4 was less different than other AF models, but still, the looseness of the structure among subdomains was shown.



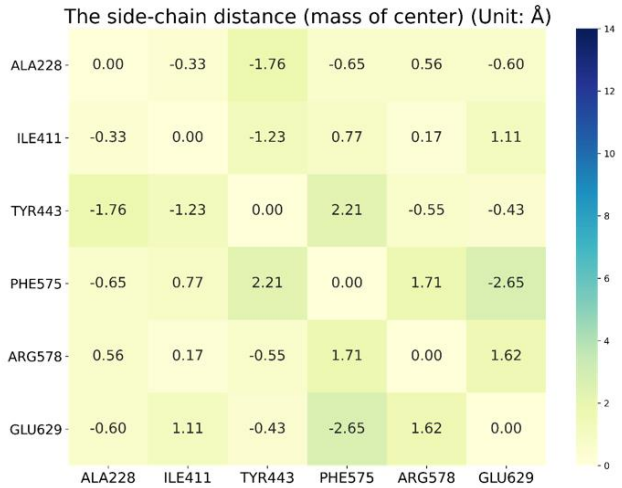
(a)



(b)



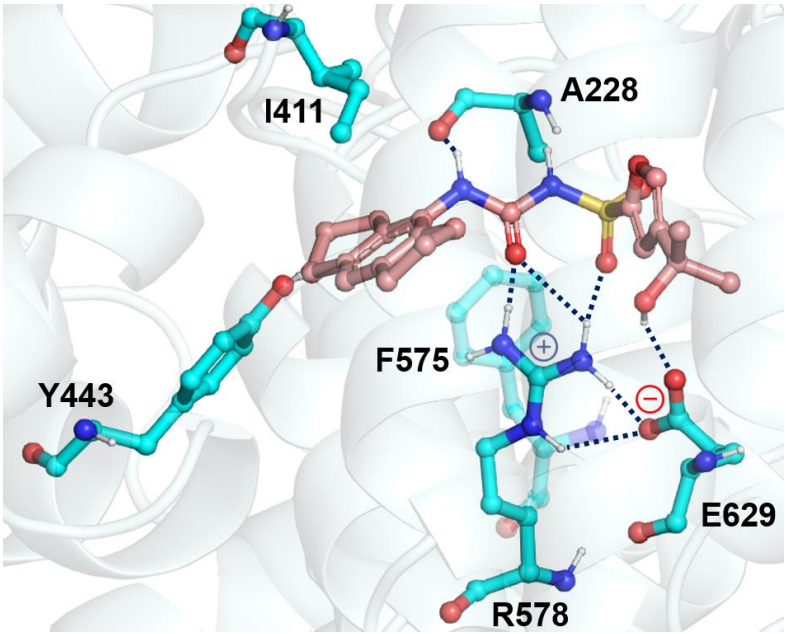
(c)



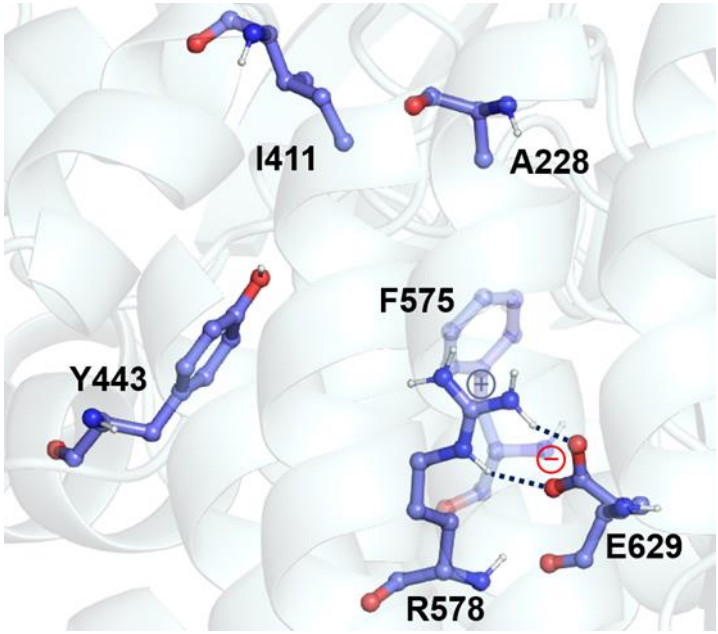
(d)

Figure 3. (a) The major interactions between MCC950 and NLRP3 in 7PZC; (b) the side-chain distances between pairs of key residues in the binding pocket of MCC950 in 7PZC; the distances between pairs of key residues in the binding pocket of MCC950 in (c) AF model 4 and (d) RF model 1 relative to 7PZC.

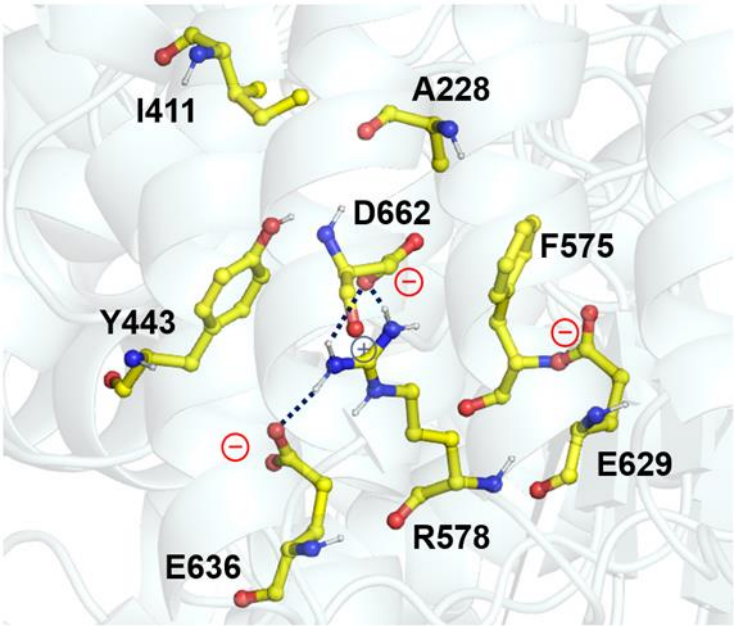
We performed molecular docking of MCC950 for each model of AF and RF. However, even the top performing ones can only produce partially correct binding modes. We also explored the use of molecular dynamics (MD) simulations to tune the side chain orientation with and without small molecule binding, to take account of the induced fit effect and dynamic effects. However, even in these attempts of using refined structures, the binding modes of the small molecules were significantly different from the experimental structures. We speculated that this may be caused by the following reasons. First, the assembly of subdomains was difficult to optimize by MD simulations because conformational changes involving the main chain typically require long convergence times. Secondly, a salt bridge interaction between R578 and E629 was observed in both 7PZC and AF models (Fig. 4), while R578 formed a salt bridge with E636 and D662 in RF models. The orientation of the R578 side chain was disrupted by the electrostatic interaction of other negatively charged residues in RF models. In addition, the walker A moiety in the NBD subdomain where A228 resides is easily displaced during MD simulations, in the absence of small molecule binding.



(a)



(b)



(c)

Figure 4. Important interactions between NLRP3-MCC950 and protein residues in (a) 7PZC, (b) AlphaFold model 4 and (c) RoseTTAFold model 1.

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, accurately predicting the protein structure in the present work is a particularly challenging case, for the NLRP3 protein is composed of multiple subdomains, and the small-molecule 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, whereas AlphaFold produced a salt bridge interaction more similar to that in the reported binding pocket. 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.

Supplementary materials

Raw structures generated by AlphaFold and RoseTTAFold. When querying templates for the protein target, we used the PDB downloaded on May 14, 2020 and the PDB70 cluster database downloaded on

May 13, 2020, both before the release dates of 7PZC and 7ALV.

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. et al. Structure of the NLRP3 decamer bound to the cytokine release inhibitor CRID3. Nature 604, 184–189 (2022). https://doi.org/10.1038/s41586-022-04467-w.

[9]. Carien Dekker, H. M., Michael Wright, Andreas Boettcher, Alexandra Hinniger, Nicola Hughes, Sandra Kapps-Fouthier, Jörg Eder, Paulus Erbel, Nikolaus Stiefl, Angela Mackay, Christopher J. Farady, Crystal Structure of NLRP3 NACHT Domain With an Inhibitor Defines Mechanism of Inflammasome Inhibition. J. Mol. Biol. **2021**, 433 (24),167309. DOI: 10.1016/j.jmb.2021.167309

[10]. 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.

[11]. 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.

[12]. 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*.