https://doi.org/10.1186/s12859-023-05571-y

BMC Bioinformatics

RESEARCH Open Access

CoDock-Ligand: combined template-based docking and CNN-based scoring in ligand binding prediction

Mingwei Pang[1†], Wangqiu He[1†], Xufeng Lu[2], Yuting She[1], Liangxu Xie[1], Ren Kong[1*] and Shan Chang[1*]

[†]Mingwei Pang and Wangqiu He are both first authors.

*Correspondence: rkong@jsut.edu.cn; schang@jsut.edu.cn

[1] Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou 213001, Jiangsu, China
[2] Primary Biotechnology Inc., Changzhou 213125, Jiangsu, China

Abstract

For ligand binding prediction, it is crucial for molecular docking programs to integrate template-based modeling with a precise scoring function. Here, we proposed the CoDock-Ligand docking method that combines template-based modeling and the GNINA scoring function, a Convolutional Neural Network-based scoring function, for the ligand binding prediction in CASP15. Among the 21 targets, we obtained successful predictions in top 5 submissions for 14 targets and partially successful predictions for 4 targets. In particular, for the most complicated target, H1114, which contains 56 metal cofactors and small molecules, our docking method successfully predicted the binding of most ligands. Analysis of the failed systems showed that the predicted receptor protein presented conformational changes in the backbone and side chains of the binding site residues, which may cause large structural deviations in the ligand binding prediction. In summary, our hybrid docking scheme was efficiently adapted to the ligand binding prediction challenges in CASP15.

Keywords: Ligand binding prediction, CASP, CoDock-Ligand, Molecular docking

Introduction

Ligand binding prediction is an important issue in the structure-based drug design. With the breakthrough of AlphaFold2 (AF2) in protein structure prediction [1], a new challenge category, ligand binding prediction was added to the 15th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP15). Previously, the Community Structure–Activity Resource (CSAR) [2] disseminated experimental datasets of diverse protein–ligand complexes to improve ligand docking and scoring. After 2015, the Drug Design Data Resource (D3R) [3] replaced CSAR and released valuable benchmarking datasets containing experimentally determined binding structures and affinity data. However, unlike the CSAR and D3R challenges, the ligand binding predictions of CASP15 provide only the target sequences without structural or ligand binding site information. These predictions are much more difficult and can be classified as blind docking. Unlike the typical blind docking, CASP only provides sequence information for the receptors. Therefore, it is necessary to predict the structures of these receptors. The predicted receptor structures may deviate from the experimental structures, regardless of the main chains or side chains, further increasing the difficulty of ligand blind docking. In addition, multiple ligands binding to one specific target make the problem even more complicated in CASP15.

Several computational methods have been developed for ligand binding prediction. Well-known ligand prediction programs include DOCK [4], AutoDock [5], Vina [6], Glide [7], GOLD [8], and MDock [9]. Recently, template-based methods have been widely used to predict ligand structures. Huang et al. proposed an enhanced Virtual Screening (VS) approach, EViS [10], which integrates ligand docking, protein pocket template searching, and ligand template shape similarity calculations. Zou et al. [11] proposed a new template-guided method using dissimilar ligands as templates, which significantly outperformed traditional molecular docking methods. PocketShape [12] used the Hungarian algorithm and the Downhill simplex method to solve the problem of binding site comparison, and achieved excellent performance in distinguishing similar from dissimilar ligand binding site pairs. To enrich the AlphaFold model with ligands and cofactors, AlphaFill [13] uses sequence and structural similarities to align small molecules and ions from experimentally determined structures with AF2 predicted protein models.

In addition, convolutional neural networks (CNNs) have been used in structure-based virtual screening and scoring. Ragoza et al. [14] showed that a fully CNN scoring function (GNINA scoring function) using only spatial and atom type information as input can outperform empirical and feature-based machine learning approaches for virtual screening. Deane et al. [15] used a Densely connected CNN (DenseNet) with a transfer learning approach to produce an ensemble of protein family-specific models for virtual screening. Jones et al. [16] fused models of 3D-CNNs and spatial graph neural networks (SG-CNNs) to make more accurate predictions than the previous docking scoring and MM/GBSA rescoring.

In CASP15, we participated in the category of ligand binding prediction. Owing to the advantages of template-based modeling and the GNINA scoring function, we combined these two methods to predict the binding modes of small molecules or metal ions. For most of the CASP15 ligand systems, our fusion docking protocol achieved successful or partially successful results. Considering its robust predictive performance, our docking protocol is a good alternative for the ligand binding predictions.

Methods

3D alignment algorithm

In a previous work, we queried Protein Data Bank (PDB) [17] for template structures using sequence similarity searching. This sequence-based template search strategy has been used for protein–protein docking prediction [18, 19]. For ligand binding prediction, a structure-based 3D alignment algorithm was developed by our group and used for both pocket template searching and ligand alignment. For the pocket template searching, CA atoms in the protein pocket were set as nodes. The cutoff distance used to select CA atoms around the center of the ligand was 12 Å. For the ligand alignment, all atoms in the ligand except hydrogen were set as nodes. For the 3D alignment algorithm, a single graph representation was combined with the clique detection method.

In a single graph representation, each node represents a pairing of atoms. One from the query structure and the other from the template structure. Adjacent nodes are two nodes for which both atoms from the query and template structures are separated by equivalent distances. First, a set of fully adjacent nodes is defined as a clique, which is a completely connected subgraph. Matching is then formulated as a graph theoretical problem that attempts to find completely connected subgraphs within an undirected graph. This 3D alignment algorithm is similar to the method used in the UCSF DOCK program [4]. In contrast to the exhaustive matching algorithm in the DOCK program, we used a greedy algorithm to improve speed, search for the most similar template, and generate orientation for alignment. We calculated the similarity coefficient as $S_T = N_S/N_T$, where N_S is the number of unique atoms shared between the query structure and the template structure, and N_T is the number of unique atoms in the template. The templates with $S_T > 0.8$ were defined as the high similar templates.

GNINA scoring function

We used the GNINA scoring function to rescore the receptor-ligand complex [14, 20]. The GNINA scoring function is a CNN-based model [14]. The size of the box in this scoring function was 24 Å × 24 Å × 24 Å and was centered on the binding site with a default resolution of 0.5 Å. Each grid point stores information regarding the types of heavy atoms at that point. The ligand and protein atoms have distinct atom types, and each atom type is represented in a different channel (analogous to the RGB channels in the images) of the 3D grid. GNINA defined a total of 34 distinct types, with 16 receptor types and 18 ligand types. The new GNINA scoring function was trained using the PyTorch deep learning framework. The CNN model in the GNINA program was trained on the cross-docking and redocking datasets, and could predict both the pose score and binding affinity. The final ligand docking poses were evaluated and ranked using the GNINA scoring function.

Docking protocol

Template-based modeling and the GNINA scoring function were combined for ligand binding prediction, and the flow chart is shown in Fig. 1. For each target, the receptor structure was predicted using AF2, or that disclosed by the CASP Organizing Committee. We queried BioLiP [21] using the 3D alignment algorithm and extracted potential pocket templates for the receptor. BioLiP is a widely used databases for protein–ligand interactions, and the data were primarily collected from the Protein Data Bank (PDB). Because ligand molecules are often used as additives to solve protein structures, not all ligands present in the PDB database are biologically relevant. BioLiP uses a composite of automated and manual procedures to examine the biological relevance of ligands. Therefore, BioLiP is very useful for template-based protein–ligand docking. If an appropriate template was available, the initial ligand binding pocket was identified based on the template. The ligand conformations were then generated from the SMILES string using RDKit [22]. The ligand conformations were aligned to small molecules of the template using the 3D alignment algorithm to generate various docking poses. Finally, these binding poses were evaluated using the GNINA scoring function. Therefore, the box size in the docking protocol was the same as that of the GNINA scoring function. We provided

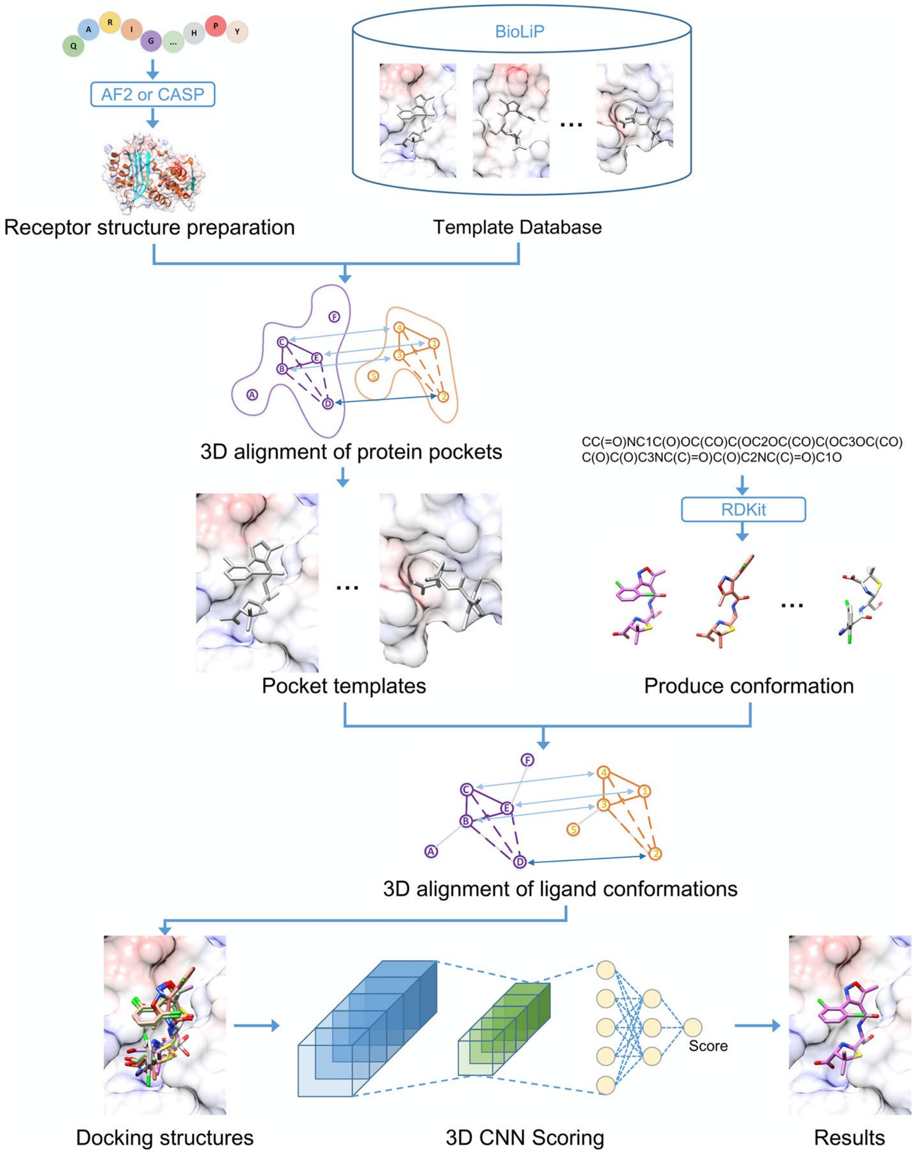
Q
A
R
I
G
...
H
P
Y
AF2 or CASP
Receptor structure preparation
BioLiP
...
Template Database
A
B
C
D
E
F
1
2
3
4
5
3D alignment of protein pockets
CC(=O)NC1C(O)OC(CO)C(OC2OC(CO)C(OC3OC(CO)
C(O)C(O)C3NC(C)=O)C(O)C2NC(C)=O)C1O
RDKit
...
Pocket templates
...
Produce conformation
3D alignment of ligand conformations
Score
Docking structures
3D CNN Scoring
Results

Fig. 1 Flow chart of the CoDock-Ligand protocol. First, the receptor pocket templates are searched in the template library using the pocket 3D alignment algorithm. Second, multiple ligand conformations are generated using RDKit. Finally, the ligand conformations are aligned to the small molecules from the pocket template, and the variety binding poses are sorted by the GNINA scoring function

a web server to users to facilitate the use of our docking program (https://codockligand.schanglab.org.cn).

Results and discussion

Docking protocol test

Upon participating in the ligand prediction assessment of CASP15, we standardized the algorithm as a docking protocol. It was tested on CASF-2016 [23], and compared with the widely used AutoDock-Vina program [24]. We also compared the test results of the combination of AutoDock-Vina and GNINA scoring (AutoDock-Vina + GNINA scoring). The docking box of AutoDock Vina was defined as the center of the native ligand coordinates with dimensions of 28 Å × 28 Å × 28 Å to include the residues of the entire cavity. The exhaustiveness value was 10. The root-mean-square deviation (RMSD) was calculated for all non-hydrogen atoms in the ligand relative to the native structure. To avoid introducing biases into the docking tests, we removed the overlapping systems between BioLiP and the test set of CASF-2016. As shown in Fig. 2, CoDock-Ligand achieved better performance than AutoDock-Vina and AutoDock-Vina + GNINA scoring in terms of the success rates of the top ranking poses. For the Top1 pose, 76.5% of the systems were predicted successfully using CoDock-Ligand with RMSD ≤ 1 Å, and 83.9% of the systems were predicted successfully with RMSD ≤ 2 Å. Correspondingly, the success rates of AutoDock-Vina were 47.7% and 62.9%, respectively. Although the GNINA scoring function was the same, the success rates of AutoDock-Vina + GNINA scoring were 47.7% and 69.3%, respectively. These comparisons demonstrated the advantages of incorporating experimental data during docking.

The overview in CASP15

Considering that many targets in CASP15 are multimers with duplicate subunits, we only take one subunit as an example to show ligand binding prediction. For each target, five models were submitted with five different poses in each model. There were 25 binding poses for each ligand, and RMSD values were calculated to reference the experimental structures. The lowest RMSD values of the 25 binding poses for each target are listed in Table 1. These results are consistent with those shown in Figs. 3, 4, 5, 6, 7, 8 and 9, and are convenient for comparing the predicted ligands with their crystal structures. In Table 1, high quality predictions are defined as RMSD ≤ 2.0 Å, accepted site predictions as 2.0 Å < RMSD ≤ 5 Å and failed predictions as RMSD > 5.0 Å. The CASP only provides sequence information for the receptors. The predicted receptor structures may deviate significantly from the experimental structures, regardless of the main chains or side chains, further increasing the difficulty of ligand docking. The prediction with RMSD between 2 and 5 Å is also important, showing that the site is correct but pose is not. In previous studies [11, 25], the RMSD cutoff of 5 Å was used as the criterion.

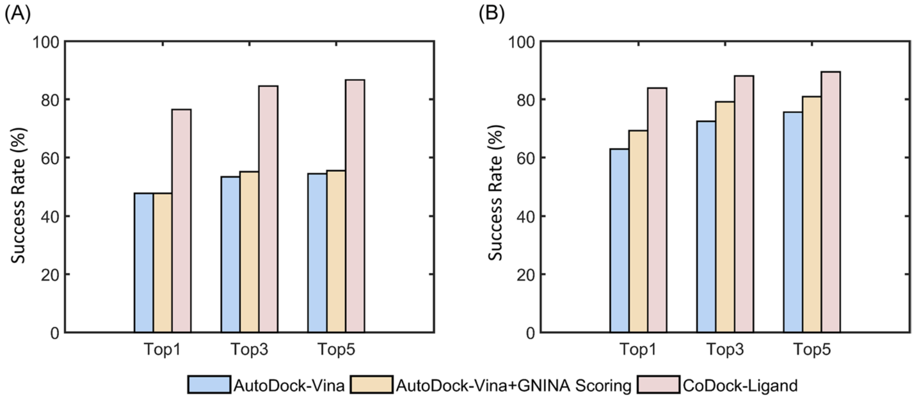

Fig. 2 Histogram of the comparison between AutoDock-Vina, AutoDock-Vina + GNINA scoring and CoDock-Ligand in the CASF-2016 dataset. **A** Success rates of three programs with RMSD ≤ 1.0 Å. **B** Success rates of three programs with RMSD ≤ 2.0 Å

Table 1 Performance of CoDock-Ligand in CASP15

CASP ID	Template	Oligomeric state	Ligand	Lowest RMSD		
				L_RMSD ≤ 2.0 Å	2.0 Å < L_RMSD ≤ 5.0 Å	L_RMSD > 5.0 Å

H1114[a]	5Y4N, 4UE3, 4KO2, 2FRV	A4B8C8	7/56	0.65(001/3NI)
				1.20(009/F3S)
				0.56(017/F3S)
				0.69(025/F3S)
				0.82(033/FCO)
				1.28(041/MG)
					3.87(049/MQ7)	

R1117	2L1V, 3FU2	A1	1/1	1.17(001/PRF)		

T1118[b]	5JJ5,6Z8A	A1	5/9	0.40(002/FE)	4.15(001/FE)
					4.35(004/LIG)
					4.73(005/LIG)
						9.13(003/LIG)

T1124	7WDW	A2	2/4	1.92(001/SAH)	2.71(003/TYR)	
R1126	5OB3, 7L0Z	A1	1/1			27.29(001/K)

T1127	3BJ7, 3BJ8, 2B4D, 2B4B, 2JEV, 2B58	A2	2/5		3.17(001/COA)
					2.16(003/EPE)	

H1135	6R16, 6R15, 6R2I	A9B3	2/12	0.08(004/K)		11.32(001/CL)

R1136	4KZD, 5OB3	A1	3/3			72.05(001/1TU)
						24.27(002/J93)
						64.57(003/K)

T1146	4Q5K, 4Q68	A1	1/1	0.58(001/NAG)		
T1152	4B8V	A2	1/1	0.81(001/NAG)		
T1158v1	5UJA, 6D3R, 6PZ9, 6PZI, 6UY0	A1	1/1	1.85(001/XPG)		
T1158v2	5UJA, 6D3R, 6PZ9, 6PZI, 6UY0	A1	1/1		3.90(001/P2E)	
T1158v3	5UJA, 6D3R, 6PZ9, 6PZI, 6UY0	A1	1/1	1.66(001/XH0)		

T1158v4	5UJA, 6D3R, 6PZ9, 6PZI, 6UY0	A1	4/4	0.79(001/ATP)
				1.24(002/ATP)
				0.14(003/2MG)
				0.65(004/2MG)		

T1170	6CHS	A6	3/9	1.20(001/ADP)
				0.77(007/MG)
					2.45(004/AGS)	

H1171	6CHS	A6B1	3/9	1.52(001/ADP)
				1.57(007/MG)
					2.48(004/AGS)	

H1172	6CHS	A6B2	3/9	1.06(007/MG)	2.09(001/ADP)
					2.38(004/AGS)	

T1181	5W6H, 4OJ6, 4OJ5, 4OJP, 4OJO	A3	6/8		4.45(006/ZN)	10.63(001/OAA)
						16.33(002/OAA)
						36.99(005/ZN)
						5.69(007/ZN)
						27.23(008/CA)

T1186	1FCM, 4R1G	A1	1/1	1.05(001/LIG)		
T1187	–	A2	1/2			14.36(001/NAG)

T1188	2YBT, 6BT9	A1	5/5		3.96(001/DW0)
					2.05(004/CD)
						6.02(002/DW0)
						11.01(003/CD)
						15.42(005/CO)

[a] In these targets, our group achieved successful results in ligand binding predictions

[b] In these targets, our group achieved partially successful results in ligand binding predictions

For some target systems, complex templates with high similarity scores were identified, including H1114, R1117, T1124, T1127, H1135, T1146, T1152, T1158, T1170, H1171, H1172 and T1186. For ligands identical to those in the complex template, structure-based alignment was directly used to obtain the ligand position in the predicted

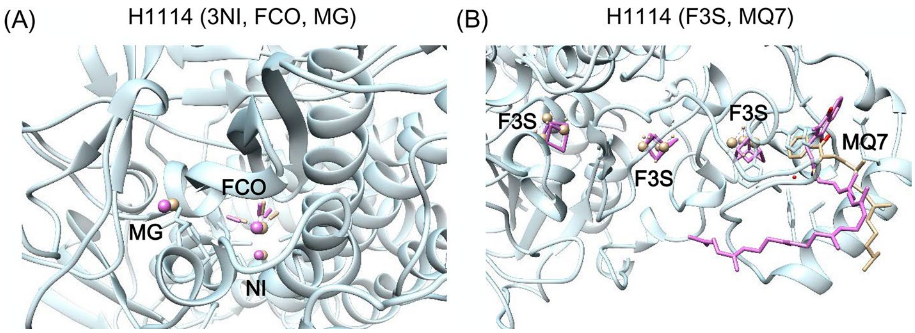

Fig. 3 Ligand prediction of H1114. The receptor protein and ligands of the crystal structure are colored light blue and orange, respectively. The predicted ligand structures are colored pink. **A** Ligands of Ni ion, FCO, and Mg ion. **B** Ligands of F3S and MQ7

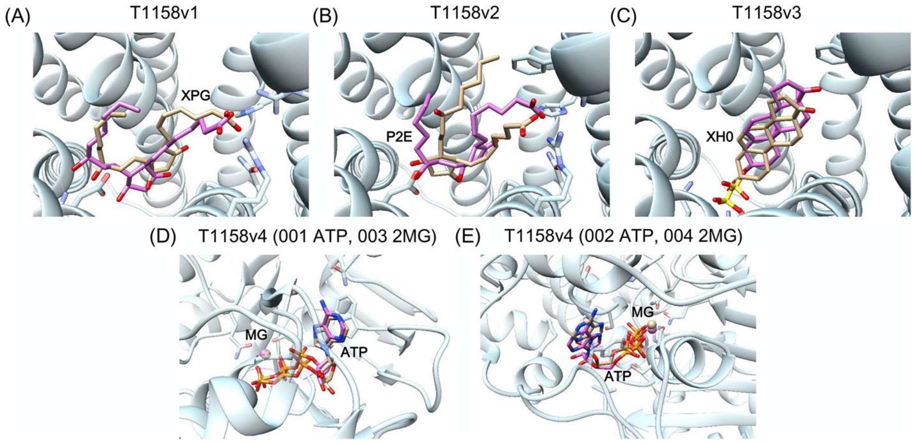

Fig. 4 Prediction of four different ligand binding with T1158. The receptor protein and ligands of the crystal structure are colored light blue and orange, respectively. The predicted ligand structures are colored pink. **A** Ligand of T1158v1. **B** Ligand of T1158v2. **C** Ligand of T1158v3. **D** Ligands of 001 ATP and 003 Mg ion in T1158v4. **E** Ligands of 002 ATP and 004 Mg ion in T1158v4

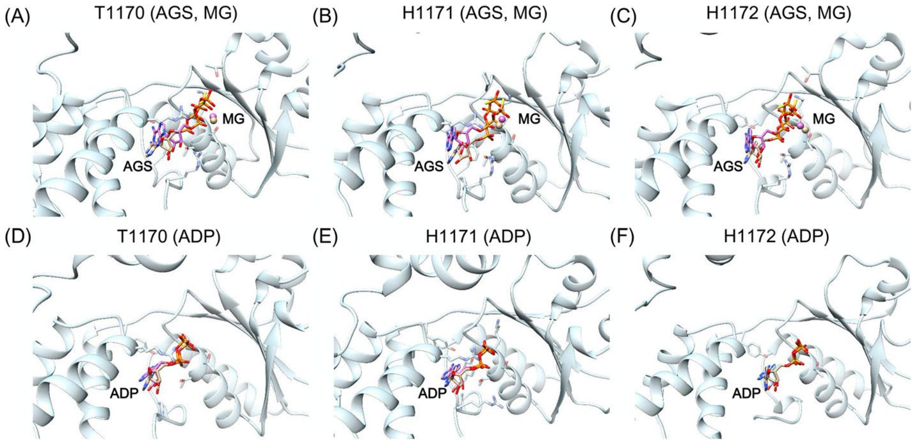

Fig. 5 Ligand prediction of T1170-H1172. The receptor protein and ligands of the crystal structure are colored light blue and orange, respectively. The predicted ligand structures are colored pink. **A**, **D** are ligands of T1170. **B**, **E** are ligands of H1171. **C**, **F** are ligands of H1172

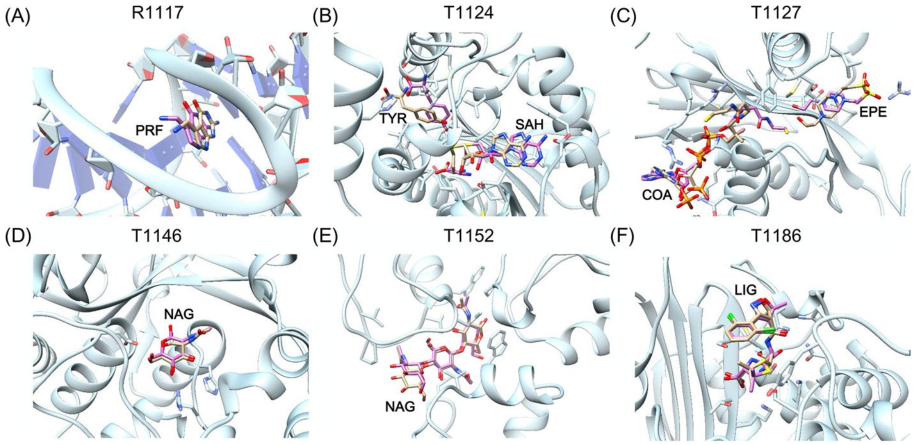

Fig. 6 Successful ligand prediction of six systems. The receptor protein and ligands of crystal structure are colored light blue and orange, respectively. The predicted ligand structures are colored pink. **A** Ligand of R1117. **B** Ligands of T1124. **C** Ligands of T1127. **D** Ligand of T1146. **E** Ligand of T1152. **F** Ligand of T1186

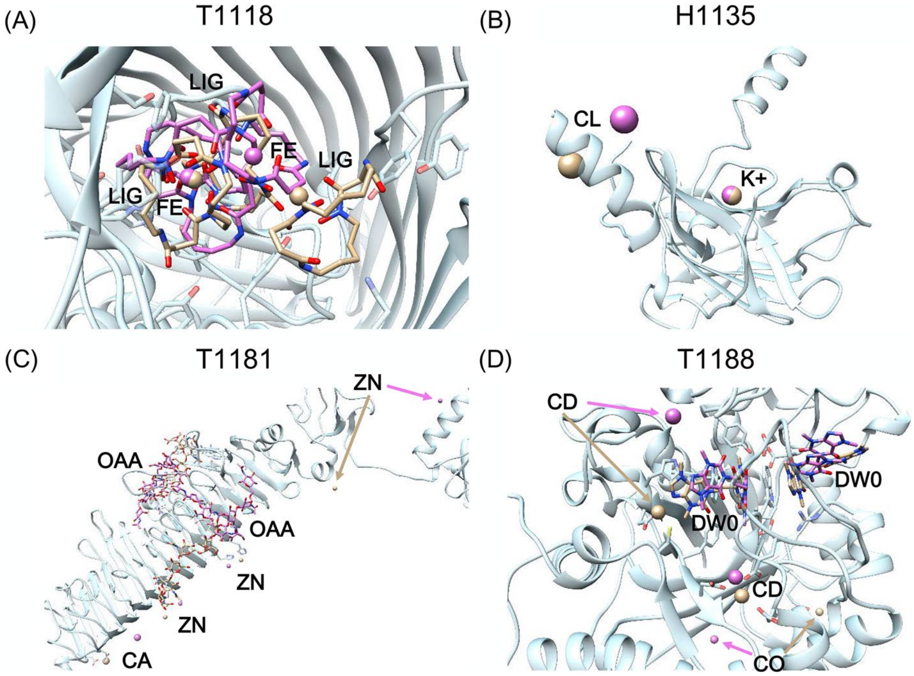

Fig. 7 Partial successfully ligand prediction of four systems. The receptor protein and ligands of crystal structure are colored light blue and orange, respectively. The predicted ligand structures are colored pink. **A** Ligands of T1118. **B** Ligands of H1135. **C** Ligands of T1181. **D** Ligands of T1188

target. For example, the ligand of R1117 and metal ions of H1114 were predicted in this manner. For the ion ligands, a simple coordinate transformation was used for docking prediction. For ligands chemically similar to those in the complex templates, template guided docking protocol was used to obtain the target-ligand complex structure. It was applied for the ligand predictions of T1124, T1158 v1, T1158 v2, and T1152. For target systems without appropriate complex templates, such as T1181 and T1187, traditional docking was performed using Glide [7]. Previous study compared Glide and GNINA

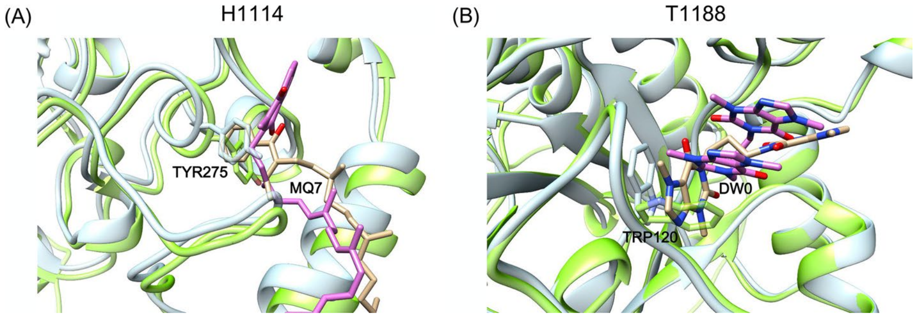

Fig. 8 Analysis the effect of the side chain on ligand prediction. The crystal structures of receptor protein and ligand are colored light blue and orange, respectively. The predicted structures of receptor protein and ligand are colored light green and pink, respectively. **A** Prediction of MQ7 in H1114. The side chain of TYR275 leads to the incorrect conformation of MQ7. **B** Prediction of DW0 in T1188. The side chain of TRP120 leads to the orientation change of the aromatic ring in DW0

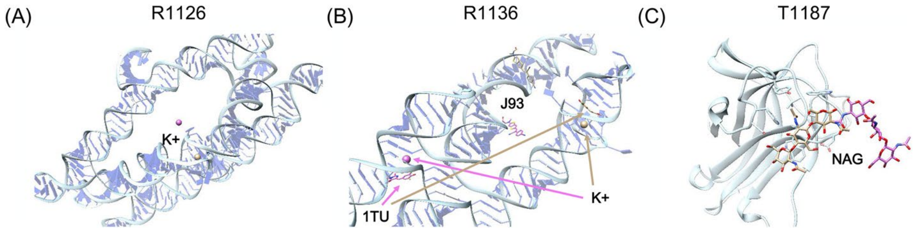

Fig. 9 Failed ligand prediction of three systems. The receptor protein and ligands of crystal structure are colored light blue and orange, respectively. The predicted ligand structures are colored pink. **A** Ligands of R1126. **B** Ligands of R1136. **C** Ligands of T1187

on the CASF-2016 dataset [25], and demonstrated that Glide performed slightly better than GNINA. When no acceptable template structures were found, Glide was used for docking.

Successfully predicted targets

H1114

H1114 is the [NiFe]-hydrogenase Huc from *Mycobacterium smegmatis* (PDB: 7UUS) [26]. The Huc catalytic subunits form an octameric complex containing 56 metal cofactors and small molecules. For this system, 5Y4N, 4UE3, 4KO2, and 2FRV were selected as the templates. Six metal cofactors and one small-molecule compound, MQ7, were predicted for each subunit. As shown in Fig. 3, the binding poses of six metal cofactors were predicted correctly with RMSD < 2.0 Å. For the MQ7 ligand, the binding site was correct, but the lowest RMSD was 3.87 Å.

T1158

T1158 v1 to v4 are MRP4-mut E1202Q protein in four states, binding to four different ligands, respectively. It is a multidrug resistance-associated protein that transport compounds out of cells. 5UJA, 6D3R, 6PZ9, 6PZI, and 6UY0 were identified as highly similar templates and used for ligand binding prediction. Based on the protein function and ligand binding in the template structures, it was deduced that the binding site was located in the center channel of the protein. As shown in Fig. 4, the ligands in v1, v3, and v4 states were predicted correctly with RMSD < 2.0 Å. For the P2E ligand in v2 state, the binding site was predicted correctly, but the lowest RMSD was 3.9 Å.

T1170-H1172

T1170-H1172 are RuvAB branch migration motor in complex with the Holliday junction (PDBs: 7PBR, 7PBL, and 7PBP) [27]. These systems bind to the same ligands, including AGS, ADP, and Mg ion. The only difference is that H1171 and H1172 are in different multimer states. To simplify the analysis, we compared only the ligands in one of these states. For these systems, 6CHS was selected as the template. As shown in Fig. 5, AGS, ADP, and Mg ion were correctly predicted in these three systems. Most of ADP and Mg ion were predicted correctly with RMSD < 2 Å. However, AGS has more rotatable bonds than ADP, and its lowest RMSDs in these three systems were 2.45, 2.48 and 2.38 Å, respectively.

Other six systems

The four systems, R1117, T1146, T1152, and T1186, are relatively simple, with only one ligand to be predicted. R1117 is a PreQ1 class I type III riboswitch, and we selected 2L1V and 3FU2 as the templates. For the T1146 system, 4Q5K and 4Q68 were used as the templates. T1152 is a clostridium thermocellum CtCBM50 structure in complex with beta-1,4-GlcNAc trisaccharide (PDB: 7R1L), and we used 4B8V as the template. T1186 is a beta-lactamase with dicloxacillin, and we selected 1FCM and 4R1G as the templates. As shown Fig. 6A, D–F, the ligand structures of these four systems were correctly predicted, and the lowest RMSDs were 1.17, 0.58, 1.52 and 1.05 Å, respectively.

The other two systems, T1124 and T1127, were more complex than the above four systems, and both had two ligands to be predicted. T1124 is an L- and D-tyrosine O-methyltransferase from the marformycin biosynthesis pathway of Streptomyces drozdowiczii, with SAH and L-tyrosine bound (PDB: 7UX8) [28]. For this system, 7WDW was used as the template. For the T1127 system, 3BJ7, 3BJ8, 2B4D, 2B4B, 2JEV, and 2B58 were selected as the templates. As shown Fig. 6B, C, the lowest RMSDs of the ligands in these two systems were close to 2.0–3.0 Å.

Partial successfully predicted targets

Four systems

T1118 is an outer membrane FoxA with (2:3) Fe-bisucaberin bound. According to the CASP information, the FoxA structure in complex with another ligand, nocardamine, has been solved (PDB: 6Z8A) [29]. For this system, 5JJ5 and 6Z8A were used as the templates. As shown in Fig. 7A, the binding sites of the ligand were correctly predicted, but the orientations of the two ligands (a Fe ion and a bisucaberin) were far away from the crystal structure. The RMSDs were 4.15 and 9.13 Å, respectively.

The oligomeric state of H1135 was A9B3, and we took one subunit for analysis (see Fig. 7B). For this system, 6R16, 6R15, and 6R2I were used as the templates. The K ion was functionally relevant and its predicted deviation was less than 1.0 Å. Although the pose of Cl ion was incorrect, it was not functionally relevant.

T1181 is a trimer, and one subunit was shown for the ligand binding prediction in Fig. 7C. For this system, 5W6H, 4OJ6, 4OJ5, 4OJP, and 4OJO were used as the templates. The ligand OAA is a polysaccharide molecule containing 47 rotatable bonds. Thus, only one Zn ion were predicted with RMSD < 5.0 Å, but the ligands OAA and Ca ion were incorrectly predicted.

T1188 had five ligands to be predicted, including two DW0 ligands and three metal cofactors. For this system, 2YBT and 6BT9 were used as the templates. As shown in Fig. 7D, the poses of two DW0 and a Cd ion in T1188 were basically correct, and the RMSDs were 3.96, 6.02 and 2.05 Å, respectively. However, the lowest RMSDs of the other two metal ions, Cd and Co ions were 11.01 and 15.42 Å, respectively.

In the above successful and partial successfully prediction targets, we also determined the reasons why some ligands were located in the correct binding pocket but had a large RMSD value. Because the receptor structure was predicted, the side chain orientations or main chain conformations of the receptor were different between the predicted structure and the experimental structure. In some cases of CASP15, subtle rotamer rearrangements of side chains greatly affect the docking results and lead to incorrect predictions of ligand poses, especially for those with π–π interactions. As shown in Fig. 8, the side chain orientations of TYR275 and TRP120 led to remarkable deviations in the ligand predictions for MQ7 in H1114 and DW0 in T1188. For the T1188 system, the RMSD value was still larger than 5.0 Å despite the correct predicted site of DW0.

Failed prediction targets

Three systems

The target structures of R1126 and R1136 are RNA molecules in the Traptamer and Apta-FRET forms, respectively. The ligand of R1126 is K ion, and those of R1136 are 1TU, J93, and K ion, respectively. We found some templates with similar ligands for these two systems. However, we used the receptor structures provided by the CASP for RNA targets. Actually, the final evaluation shows that the backbone deviations between these predicted receptor structures and the experimental structures were 52.637 and 54.508 Å, respectively. Thus, the ligands could not be aligned correctly by our modeling method (see Fig. 9A, B).

Another failed target is T1187, which is a tobacco lectin Nictaba in complex with triacetylchitotriose (PDB: 8AD2). For this system, our docking protocol did not identify any appropriate complex templates, so traditional docking was performed using Glide. As shown in Fig. 9C, the predicted binding site deviated from the actual binding site, and the lowest RMSD was 14.36 Å.

Here we proposed a template based docking protocol, CoDock-Ligand, and applied in CASP15 Ligand prediction category. An atom-type based 3D align algorithm was designed to capture potential pocket templates and perform ligand alignment. Combined with GNINA scoring function, CoDock-Ligand achieved better performance than AutoDock-Vina in terms of the success rates of ligand poses. Our group (CoDock) was ranked as No. 1 in the ligand prediction category, showing remarkable accuracy in receptor-ligand complex structure prediction. However, the receptor structure showed significant impact for the ligand binding predictions. Such as the ligand predictions for MQ7 in H1114 and DW0 in T1188, subtle side-chain conformation changes of pocket residues made the bad predictions for ligand poses. For RNA-ligand interactions, poor RNA structure predictions led to final failures for ligand binding predictions. It is necessary to consider the conformation flexibility of proteins and further understand of RNA interactions [30–34], which may help to improve the ligand binding predictions in the future.

In recent years, ODE-based theoretical modeling studies have been widely applied on gene/protein signaling networks [35–37]. Combine with these methods, the docking models could contribute to understand regulatory mechanisms and find potential therapeutic targets in diseases. In addition, some new computational methods, such as graph convolutional neural networks, have been used to predict interactions, which provide valuable insights into genetic markers and related diseases [34, 38, 39]. These models will be helpful to improve the docking protocol and further provide atomic details for interactions between bio-molecules.

Conclusion

Our docking protocol combines template-based modeling and the GNINA scoring function for receptor-ligand structure prediction. The template-based modeling method adopts a structure-based 3D alignment algorithm developed by our group that can accurately identify templates from the structure database. This method captured similar templates for most targets, and CoDock-Ligand achieved successful or partially successful predictions for these systems in CASP15. Meanwhile, we analyzed the failed systems to determine the reasons. If there was a remarkable backbone deviation between the predicted receptor structure and the experimental structure, such as the RNA-ligand systems R1126 and R1136, our docking protocol failed in the ligand binding prediction. Because the receptor structures are predicted in the ligand prediction assessment of CASP15, the side chain conformations, especially the orientations of the aromatic rings have a significant impact on ligand binding predictions. Therefore, the flexibility of the receptor should be considered in the docking protocol, and the precise receptor structure greatly contributes to ligand binding predictions.

Abbreviations

GNINA	A fork of SMINA that allows the utilization of grid-based CNN models as scoring functions
CASP15	The 15th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction
CNN	Convolutional neural network
CSAR	The Community Structure–Activity Resource
D3R	Drug Design Data Resource
VS	Virtual Screening
DenseNet	Densely connected CNN
SG-CNNs	Spatial graph neural networks
PDB	Protein Data Bank

Acknowledgements

We thank the organizers and the evaluation team of CASP for their effort, the experimental groups for providing targets, and Prof. Cun-Xin Wang and Prof. Xiao-qin Zou for helpful discussion.

Author contributions

Methodology, MP, WH, XL, RK and SC; Supervision, LX, RK and SC; Validation, MP, YS and SC; Writing—original draft, MP, WH, RK and SC.

Funding

This work was supported by National Natural Science Foundation of China (62373172, 22003020), the fund of Changzhou Sci. and Tech. Program (CE20205033).

Availability of data and materials

The data of CASP15 experiment can be downloaded from http://predictioncenter.org/casp15/index.cgi. The CoDock-Ligand web server, with the online documentation, is publicly available at https://codockligand.schanglab.org.cn.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 21 September 2023 Accepted: 16 November 2023
Published online: 23 November 2023

References

1. 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(7873):583–9.
2. Smith RD, Dunbar JB Jr, Ung PM-U, Esposito EX, Yang C-Y, Wang S, Carlson HA. CSAR benchmark exercise of 2010: combined evaluation across all submitted scoring functions. J Chem Inf Model. 2011;51(9):2115–31.
3. Gathiaka S, Liu S, Chiu M, Yang H, Stuckey JA, Kang YN, Delproposto J, Kubish G, Dunbar JB, Carlson HA, et al. D3R grand challenge 2015: evaluation of protein–ligand pose and affinity predictions. J Comput Aided Mol Des. 2016;30(9):651–68.
4. Ewing TJA, Kuntz ID. Critical evaluation of search algorithms for automated molecular docking and database screening. J Comput Chem. 1997;18(9):1175–89.
5. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem. 2009;30(16):2785–91.
6. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–61.
7. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem. 2004;47(7):1739–49.
8. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen. J Mol Biol. 1997;267(3):727–48.
9. Huang S-Y, Zou X. An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials. J Comput Chem. 2006;27(15):1866–75.
10. Zhang W, Huang J. EViS: an enhanced virtual screening approach based on pocket-ligand similarity. J Chem Inf Model. 2022;62(3):498–510.
11. Xu X, Zou X. Dissimilar ligands bind in a similar fashion: a guide to ligand binding-mode prediction with application to CELPP studies. Int J Mol Sci. 2021;22(22):12320.
12. Li S, Cai C, Gong J, Liu X, Li H. A fast protein binding site comparison algorithm for proteome-wide protein function prediction and drug repurposing. Proteins Struct Funct Bioinform. 2021;89(11):1541–56.
13. Hekkelman ML, de Vries I, Joosten RP, Perrakis A. AlphaFill: enriching AlphaFold models with ligands and cofactors. Nat Methods. 2023;20(2):205–13.
14. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-ligand scoring with convolutional neural networks. J Chem Inf Model. 2017;57(4):942–57.
15. Imrie F, Bradley AR, van der Schaar M, Deane CM. Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. J Chem Inf Model. 2018;58(11):2319–30.
16. Jones D, Kim H, Zhang X, Zemla A, Stevenson G, Bennett WFD, Kirshner D, Wong SE, Lightstone FC, Allen JE. Improved protein-ligand binding affinity prediction with structure-based deep fusion inference. J Chem Inf Model. 2021;61(4):1583–92.
17. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Iype L, Jain S, Fagan P, Marvin J, et al. The protein data bank. Acta Crystallogr Sect D Biol Crystallogr. 2002;58:899–907.
18. Kong R, Liu R-R, Xu X-M, Zhang D-W, Xu X-S, Shi H, Chang S. Template-based modeling and ab-initio docking using CoDock in CAPRI. Proteins Struct Funct Bioinform. 2020;88(8):1100–9.
19. Lensink MF, Brysbaert G, Mauri T, Nadzirin N, Velankar S, Chaleil RAG, Clarence T, Bates PA, Kong R, Liu B, et al. Prediction of protein assemblies, the next frontier: the CASP14-CAPRI experiment. Proteins Struct Funct Bioinform. 2021;89(12):1800–23.
20. McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, Sunseri J, Koes DR. GNINA 1.0: molecular docking with deep learning. J Cheminformatics. 2021;13(1):43.
21. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Res. 2012;41(D1):D1096–103.
22. RDKit: Open-source cheminformatics. https://www.rdkit.org.

23. 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. 2019;59(2):895–913.
24. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and python bindings. J Chem Inf Model. 2021;61(8):3891–8.
25. Stärk H, Ganea O, Pattanaik L, Barzilay DR, Jaakkola T. EquiBind: geometric deep learning for drug binding structure prediction. In: Proceedings of the 39th international conference on machine learning. 2022. p. 20503–20521.
26. Grinter R, Kropp A, Venugopal H, Senger M, Badley J, Cabotaje PR, Jia R, Duan Z, Huang P, Stripp ST, et al. Structural basis for bacterial energy extraction from atmospheric hydrogen. Nature. 2023;615(7952):541–7.
27. Wald J, Fahrenkamp D, Goessweiner-Mohr N, Lugmayr W, Ciccarelli L, Vesper O, Marlovits TC. Mechanism of AAA+ ATPase-mediated RuvAB–Holliday junction branch migration. Nature. 2022;609(7927):630–9.
28. Wu K-L, Moore JA, Miller MD, Chen Y, Lee C, Xu W, Peng Z, Duan Q, Phillips GN Jr, Uribe RA, et al. Expanding the eukaryotic genetic code with a biosynthesized 21st amino acid. Protein Sci. 2022;31(10): e4443.
29. Normant V, Josts I, Kuhn L, Perraud Q, Fritsch S, Hammann P, Mislin GLA, Tidow H, Schalk IJ. Nocardamine-dependent iron uptake in pseudomonas aeruginosa: exclusive involvement of the FoxA outer membrane transporter. ACS Chem Biol. 2020;15(10):2741–51.
30. Hu H, Feng Z, Lin H, Cheng J, Lyu J, Zhang Y, Zhao J, Xu F, Lin T, Zhao Q, et al. Gene function and cell surface protein association analysis based on single-cell multiomics data. Comput Biol Med. 2023;157: 106733.
31. Zhang L, Yang P, Feng H, Zhao Q, Liu H. Using network distance analysis to predict lncRNA–miRNA interactions. Interdiscip Sci Comput Life Sci. 2021;13(3):535–45.
32. Meng R, Yin S, Sun J, Hu H, Zhao Q. scAAGA: single cell data analysis framework using asymmetric autoencoder with gene attention. Comput Biol Med. 2023;165: 107414.
33. Gao H, Sun J, Wang Y, Lu Y, Liu L, Zhao Q, Shuai J. Predicting metabolite–disease associations based on auto-encoder and non-negative matrix factorization. Brief Bioinform. 2023;24(5):bbad259.
34. Wang W, Zhang L, Sun J, Zhao Q, Shuai J. Predicting the potential human lncRNA–miRNA interactions based on graph convolution network with conditional random field. Brief Bioinform. 2022;23(6):bbac463.
35. Li X, Zhang P, Yin Z, Xu F, Yang Z-H, Jin J, Qu J, Liu Z, Qi H, Yao C, et al. Caspase-1 and Gasdermin D afford the optimal targets with distinct switching strategies in NLRP1b inflammasome-induced cell death. Research. 2022;2022:9838341.
36. Li X, Zhong C-Q, Wu R, Xu X, Yang Z-H, Cai S, Wu X, Chen X, Yin Z, He Q, et al. RIP1-dependent linear and nonlinear recruitments of caspase-8 and RIP3 respectively to necrosome specify distinct cell death outcomes. Protein Cell. 2021;12(11):858–76.
37. Xu F, Miao D, Li W, Jin J, Liu Z, Shen C, Zhang J, Shuai J, Li X. Specificity and competition of mRNAs dominate droplet pattern in protein phase separation. Phys Rev Res. 2023;5(2): 023159.
38. Wang T, Sun J, Zhao Q. Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism. Comput Biol Med. 2023;153: 106464.
39. Sun F, Sun J, Zhao Q. A deep learning method for predicting metabolite–disease associations via graph neural network. Brief Bioinf. 2022;23(4):bbac266.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.