Predicting the protein-ligand affinity from molecular dynamics trajectories

Yaosen Min\textsuperscript{1}\textdagger, Ye Wei\textsuperscript{1}\textdagger, Peizhuo Wang\textsuperscript{1}\textdagger, Nian Wu\textsuperscript{1}, Stefan Bauer\textsuperscript{2}, Shuxin Zheng\textsuperscript{3}, Yu Shi\textsuperscript{3}, Yingheng Wang\textsuperscript{4}, Xiaoting Wang\textsuperscript{5}, Dan Zhao\textsuperscript{1}, Ji Wu\textsuperscript{1}\textasteriskcentered and Jianyang Zeng\textsuperscript{1,5}\textasteriskcentered

\textsuperscript{1}Institute of Interdisciplinary Information science, Tsinghua University, China. \\
\textsuperscript{2}Department of Intelligent Systems, KTH Stockholm, Sweden. \\
\textsuperscript{3}Microsoft Research Asia. \\
\textsuperscript{4}Department of Electrical Engineering, Tsinghua University, China. \\
\textsuperscript{5}Silexon AI Technology Co., Ltd, Nanjing, China.

*Corresponding author(s). E-mail(s): weiye@mail.tsinghua.edu.cn; wuji.ee@mail.tsinghua.edu.cn; zengjy@gmail.com;
\textdagger These authors contributed equally to this work.

Abstract

Accurate protein-ligand binding affinity prediction is essential in drug design and many other molecular recognition problems. Despite many advances in affinity prediction based on machine learning techniques, they are still limited since the protein-ligand binding is determined by the dynamics of atoms and molecules. To this end, we curated an MD dataset containing 3,218 dynamic protein-ligand complexes and further developed Dynaformer, a graph-based deep learning framework. Dynaformer can fully capture the dynamic binding modes by considering various geometric characteristics of the interaction. Our method shows superior performance on the benchmark dataset over the methods hitherto reported. Moreover, we performed virtual drug-screening on heat shock protein 90 (HSP90) by integrating our model with structure-based docking. We benchmarked our performance against the conventional knowledge-based
approach, demonstrating that our method can identify two lead compounds with submicromolar activity (0.207 µM and 0.517 µM). We anticipate that large-scale MD dataset and machine learning models will form a new synergy, providing a new route toward accelerated drug discovery.

**Keywords:** Computational biology, Graph transformer, Molecular dynamics, Drug discovery

## Introduction

Molecular recognition is of fundamental importance in biology. The interactions between proteins and small molecules regulate biological functions and systems such as enzyme catalysis and signaling pathways [1, 2]. Therefore, the research of drug design and discovery has primarily focused on manipulating the molecular binding on proteins. As the binding affinity measures the strength of such interactions, accurate affinity prediction is one of the most critical tasks in computational-aided drug discovery [3].

Recent advancements in computing have enabled protein-ligand binding affinity calculation through physics- and knowledge-based approaches [4]. The molecular docking algorithms are an essential coarse-grained physical model that efficiently predicts and evaluates the binding conformation of the small molecule. However, such models usually incorporate expert knowledge, hand-tuned features, and parameters, leading to biased results [5]. All-atom molecular dynamics (MD) is another popular computational technique to investigate the structural and dynamical properties of the biological system. It is predicated on Newtonian dynamic equations and reveals full atomic detail at a fine spatial-temporal resolution, offering accurate binding affinity prediction (i.e., alchemical free energy calculations) [6, 7]. Despite various advantages of MD, it consumes prohibitively high computational resources, rendering MD-based large-scale affinity prediction infeasible [8, 9].

Data-driven approaches are another emerging tool in the affinity prediction task [10]. Primarily benefiting from the rapidly growing amount of experimental protein-ligand complex data, data-driven approaches have achieved a high accuracy level by learning the binding modes directly from three-dimensional (3D) structures. Nonetheless, current state-of-the-art methods require sophisticated feature engineering while offering limited generalizability. More importantly, various attempts at accuracy improvement, such as topology-, spectral-, sequence- or atom pair-based methods, have mainly centered on feature engineering from the static 3D structure [11–15]. So far, none of which has taken the dynamics of protein-ligand interaction into account. In fact, all proteins are highly dynamic biomolecules, constantly interacting with their surroundings and often adopting various conformational states. As a result, proteins exist in an ensemble of conformational states. Nevertheless, the experimentally resolved structure only represents a temporally averaged...
snapshot of the most populated state, and the dynamic effects of the structure are inevitably all but lost. Furthermore, the ligand is partially exposed to solvent, therefore, it does not only interact with protein but also with the solvent (i.e., liquid water and buffer ion). With sufficient MD simulation duration, the influence of ligand-solvent interaction and other long- and short-ranged interactions (e.g., Van der Waals interaction, hydrogen bond, and any other polar/apolar interactions) on the ligand conformations are naturally embedded in the MD trajectories. Despite all the benefits, no MD-based protein-ligand dataset is as yet reported.

In this work, we curated a large-scale MD protein-ligand dataset and introduced a dynamic-graph-transformer (Dynaformer) model trained on the MD trajectories. To fully capture the dynamic binding modes, we developed the Dynaformer based on a roto-translation invariant encoding scheme, taking various interaction features into account, i.e., interatomic distances and angles between bonds. Compared to other methods, Dynaformer showed the highest accuracy on the benchmark dataset. Furthermore, we discussed several underlying thermodynamic binding mechanisms that contributed to the improved performance of Dynaformer. We also experimentally demonstrated that, compared to the molecular docking method, our model can identify new lead compounds in a more efficient manner.

Results

Thermodynamics determines the binding affinity

The driving forces that determine the association between protein and ligands are a collective consequence of various complex interactions and energy exchanges among the protein, the ligand, and the aqueous environment[16]. Consequently, as shown in Fig. 1, the state distribution of a biological system is dictated by the thermodynamic laws[17]. Eq. 1 indicates that the actual ligand binding strength is determined by the Gibbs free energy change $\Delta G$, which is a sum of enthalpy change ($\Delta H$) and entropy change ($T\Delta S$). Nonetheless, the experimentally resolved state occupies the lowest energy level $\Delta H_1$, which cannot represent the full picture of the binding mode (Fig. 1). Therefore, the best way to improve a predictive model is to introduce the dynamic effects as inductive bias into the training dataset. It is proved that both $\Delta H_i$ and $\Delta S_i$ can be recovered from an MD ensemble of protein-ligand interaction snapshots[18], which is mathematically expressed as:

$$\text{Binding affinity} \propto \Delta G = \Delta H - T \Delta S$$

$$= \frac{1}{N} \sum_{i=1}^{N} (\Delta H_i - T \Delta S_i),$$

(1)
in which $T$ represents the temperature and $N$ indicates the total number of states sampled from MD trajectories. $\Delta H_i$ and $\Delta S_i$ represent the enthalpy and entropy at the $i$-th snapshot (see Methods for detailed derivation).

Fig. 1 The state ensemble determines the binding affinity. The Gibbs free energy of protein-ligand complex is a sum of entropy and enthalpy of states, which determines the binding affinity. The state ensemble follows a certain distribution that is governed by thermodynamic laws.

Dynaformer framework

The structure of molecules can be interpreted as graph-structured data, where nodes represent atoms and edges stand for the relationships between atoms, i.e., chemical bonds or non-covalent interactions (Fig. 2(B, C)). Such graph structure retains the full physical, chemical, and geometric information within the dynamic trajectories. Fig. 2(D) illustrates the basic architecture of Dynaformer, consisting of the multi-head self-attention module and a feed-forward network (see Methods). Dynaformer is directly built upon the Graphormer, which achieves state-of-the-art performances on a wide range of molecule-related tasks[19]. However, Graphormer can only handle isolated molecules with covalent bonds. Whereas in a protein-ligand complex, atoms with non-covalent interactions and their geometry in the binding site play key roles in determining the binding affinity. To capture these intricate features, Dynaformer introduces additional attention bias terms to encode the information regarding the interatomic distances and angles between bonds (Fig. 2(E)).

To mimic the in vitro aqueous environments, we developed an MD data pre-processing protocol that solvates and neutralize the protein-ligand complex in a simulation box with water molecules and ions (Fig. 2(A)). In total, 3,218
protein-ligand complexes are selected for MD simulation and the detailed selection criteria are provided in Methods. The ideal duration for the simulation is the time scale at which a biological phenomenon takes place. Nevertheless, the time scale is highly case-specific and could lead to an explosion in computation cost. In this study, we selected 10 nanoseconds (ns) for all systems for the following practical reasons: 1) Most systems reach convergence within 10 ns. 2) Characteristic binding-related events can be observed within 10 ns (discussed in the next section). As shown in Fig. 2(F), each data in the MD database contains 100 uniformly sampled frames from MD simulation. For the molecular simulation packages, we adopt the latest version of NAMD 3.0, which maximizes the performance of moderate-sized (the number of atoms is less than 1 million) MD simulations for modern GPU architectures[20].

Fig. 2 Overview of the Dynaformer framework. (A) MD simulation of a protein-ligand complex. (B) A graph representation of the protein-ligand binding state. (C) The nodes and edges in the graph represent the atoms and their covalent or non-covalent interactions. (D) One layer of Dynaformer consists of a multi-head self-attention module and a feed-forward network. (E) The 3D geometric characteristics are encoded and added to the multi-head attention module as the attention bias. (F) Each MD data contains 100 frames of the ligand and its surrounding residues within 5Å.

To build an efficient interactive graph, we only account for the protein atoms within a certain range, since the strength of interatomic interactions decays rapidly within a few Å. Hence, the relative position of the ligand to the protein is important, which can be characterized by root-mean-square deviation (RMSD) from its initial position at the binding site (see Methods). Favorable configurations usually exhibit high stability and the construction of the graph is straightforward; but for unstable configurations, it is often not possible to construct an interactive graph representation as the ligand can leave the original pocket and enters into the solvent. Therefore, to gain comprehensive insights into the configuration stability of the MD dataset, we performed RMSD analysis over all MD simulations (3,218 complexes). Fig.
Fig. 3(A) demonstrates the mean RSMD (ligand position relative to the original pocket) normalized by the number of atoms of the ligand (# atom) versus the # atom. We characterized the configuration with normalized mean RSMD lower than 0.1 as ‘stable’ and higher ones as ‘unstable’. Fig. 3(A) suggests that 90% of the configurations are stable, as indicated by their low normalized mean RSMD. Additionally, the normalized mean RSMD quickly rises when the # atom is below 37, suggesting that the configuration stability is highly correlated to the # atom. Such correlation can be explained by the fact that smaller ligands (lower # atom) generally interact less with protein residues than bigger ones and are more susceptible to the impact of the random motion of solvent molecules. Not unexpectedly, Fig. 3(B) shows that the center of the binding affinity distribution of smaller molecules (# atom < 37) is located at $pK_i = 4.2$, which is significantly lower compared to that of the bigger molecules ($pK_i = 7.8$). Fig. 3(C) showcases an example (# atom = 80, PDB ID: 2yge), where the binding is very stable (RSMD = 0.5 ± 0.2 Å). While Fig. 3(D) shows a small ligand left the original pocket (# atom < = 24, PDB ID: 4y3j) in which RMSD indicates that the ligand drifts away from the binding site after 6ns. Indeed, the experimental evidence corroborates the MD observation, showing a very low affinity ($pK_i = 2.17$). Meanwhile, Fig. 3(E) shows the Dynaformer predictions of a stable complex (normalized RSMD = 0.08, PDB ID: 4ciw, $pK_i = 4.82$). Although the affinity prediction of individual frames fluctuates from $pK_i = 6$ to 4 due to the dynamic surrounding environment of the ligand, the averaged dynamic affinity prediction ($pK_i = 4.71$) is close to the ground truth. As the affinity predictions using averaged dynamic and experimentally measured structure showed no essential difference on most configurations, we used the experimental measured structure for the model inference on the benchmark dataset. For the performance evaluation and comparison, we used the PDBBind 2016 core set as the benchmark dataset, which is a stand-alone test set for protein-ligand binding affinity prediction (see Methods). We benchmarked Dynaformer against other major methods and the results are shown in Fig. 3(F-H), which include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Pearson correlation (Pearson $r$) of best-performing models in its categories (detailed analysis see Supplementary information). It is evidenced that RSME, MAE, and Pearson $r$ of Dynaformer outperform other baseline methods by a large margin. Additionally, we selected SIGN as our baseline model for prediction comparison in the next section, as it shows the second-best performance in all three metrics.
Fig. 3 RMSD analysis and model performance evaluation. (A) Average normalized RMSD versus the ligand atom number; (B) The histogram illustrates the binding affinity distribution of ligands with different atom numbers. Blue (Red) histogram indicates ligands with less (more) than 30 atoms; (C) RSMD plot of a very stable complex (PDB ID: 2yge), where the mean RSMD remains below 1 Å. (D) RSMD plot of an unstable complex (PDB ID: 4y3j), where the mean RSMD rises sharply at 6 ns, indicating that the ligand leaves the binding site. (E) Influence of dynamic surrounding on the model prediction (PDB ID: 4ciw). The prediction of each frame fluctuates around the experimental value ($pK_i = 4.82$). (F to H) Dynaformer outperforms different methods in terms of the RMSE, MAE and Pearson $r$.

Molecular dynamics improves affinity prediction

The improved performance is likely due to Dynaformer’s ability to learn the dynamic binding modes from MD dataset. Fig. 4 displays three cases from the PDBBind 2016 core set, in which Dynaformer shows high accuracy where SIGN deviates considerably. The detailed analysis shows that the dynamic binding process in these cases significantly differs from its static counterpart. These case studies unequivocally suggest that our MD-based dataset contains more information regarding the binding process and Dynaformer is able to learn from it.

The first case (PDB ID: 2v7a) is shown in Fig. 4(A), ITK kinase domain in complex with a ligand, demonstrating the entropy and enthalpy
compensation[21]. Here, SIGN underestimated the binding affinity by a large margin of more than 40%. While both static and dynamic data contain enthalpy information, the dynamic data contained extra entropic information (the ligand flexibility), which had been learned by Dynaformer and results in better accuracy. As it can be seen in the zoom-in figure of Fig. 4(A), the head of the ligand (pyrrolo[3,4-c]pyrazole group) was tightly buried in the pocket, whose contacts with protein are highly conserved during the MD simulation. The two nitrogen atoms of the pyrazole group interacted with the carbonyl oxygen of GLU-316 and with the amide nitrogen of MET-318, whereas the nitrogen of the amide group forms a hydrogen bond with the carbonyl oxygen of MET-318. In addition, the pyrrolopyrazole core held hydrophobic contact with ALA-269, VAL-256 and LEU-248. These favorable non-covalent interactions led to a large negative enthalpy change term. Whilst the Root Mean Square Fluctuation (RMSF) plot (Fig. 4(B)), which measures the fluctuation of a group of atoms around their average positions, indicated that the N-methylpiperazine group exhibits less contact with protein and wiggles randomly due to its exposure to the solvent, resulting in a positive entropy change term. Consequently, this ligand showed a strong affinity towards protein ($pK_i = 8.3$).

The second case (PDB ID: 3udh) is demonstrated in Fig. 4(C), where the protease endothiapepsin as the receptor interacted with 1-methyl-L-histidine derived from a novel fragment library[22]. Various conformers were observed during the MD simulation and the ligand drifted away from the original pocket after a few femtoseconds (fs). Such drift-away event was demonstrated by a large deviation (more than 4 Å RMSD) of the ligand from its initial position (Fig. 4(D)). Consequently, such molecules should have a relatively low binding affinity. Indeed, such phenomena were commonly observed in the MD simulation of low-affinity protein-ligand complex (e.g., 4y3j and 3gv9), but SIGN prediction ($pK_i = 5.9$) could deviate as much as 200% from ground truth, indicating that models can only learn such knowledge by training on the MD trajectories.

Fig. 4(E) demonstrates the Dynaformer’s ability to handle the affinity cliff, by which the ligands with similar conformation possess distinct binding affinity. As shown in Fig. 4(F-H), three thiophenes bind to PTP1B (Protein tyrosine phosphatase 1B) with different potency[23]. 2-phenylthiophene, as the structural core of the three compounds, adopts a nearly identical binding pose to occupy the active site, and tail groups (phenyl ring) entered a new subsite (Supplementary information). The RMSD plot in Fig. 4(I and J) indicates the ligand from 2qbp exhibits high conformational stability at the binding site ($\text{RSMD} < 1\text{Å}$), likely due to the Pi-stacking interaction with residue PHE-182 and hydrogen bonds with residue LYS-120 and ARG-221. On the other hand, 2qbq’s ligand left the original active site after a very short time (around 1 nanosecond) while the structural core of the ligand of 2qbr still occupies the main active site thanks to the $\pi$-stacking interaction with residue TYR-48, albeit with less stability ($\text{RSMD} > 3\text{Å}$). These structurally similar but dynamically different ligands provide distinct binding strength ($pK_i$ of 2qbp = 8.4,
As shown in Fig. 4(E), Dynaformer provides more accurate predictions ($pK_i$ of 2qbp = 7.5, 2qbq = 7.0, 2qbr = 5.9) in comparison with SIGN. More importantly, it differentiates the rank of the binding strength (2qbp > 2qbq > 2qbr), thus providing more reliable guidance toward lead optimization.

Fig. 4 Various dynamic binding mechanisms. (A and B) The entropy and enthalpy compensation and the ligand RMSD plot; The head of the ligand is tightly buried (enthalpy contribution) in the pocket (pyrrollo[3,4-c]pyrazole group) while the N-methylpiperazine group wiggles randomly (entropy contribution) caused by its exposure to the solvent. (C and D) Flexible ligand movement; the protease endothiapusin interacted with 1-methyl-L-histidine. Various binding poses had been observed during the MD simulation, and the ligand drifts away from the original pocket only after a few fs. (E-H) Same conformation, different dynamics; Three thiophenes bind to PTP1B (Protein tyrosine phosphatase 1B) with different potency. 2-phenylthiophene, as the structural core of the three compounds, adopts nearly identical conformation to occupy the active site. However, the tail groups of these compounds entered different subsites, resulting in completely different dynamics. (I and J) ligand pocket RMSD, indicating that three ligands had considerably different dynamics. Notably, the ligand from 2qbp exhibited high conformational stability whereas 2qbr displays high flexibility. such difference was also observed in the interaction diagram. PS, HP, and HB stand for Pi-stacking interaction, hydrophobic interaction, and hydrogen Bond respectively.

Nevertheless, the spatial-temporal patterns of the binding process can be difficult to learn in some cases. Fig. 5(A) shows an example, in which Thrombin as a serine protease is involved in blood coagulation and an inhibitor with a bis(phenyl) methane moiety bind to the active site of thrombin with
remarkable potency (PDB ID: 2zda, $pK_i = 8.4$)[24, 25]. In this case, both the predictions of SIGN and Dynaformer ($pK_i = 5.9$ and $6.4$) did not agree with the experimental result. A plausible explanation might be that the protein with high conformational flexibility violates the assumption that binding does not render significant protein structural change (see Methods), thereby making accurate calculation and prediction of the binding affinity very difficult. This is evidenced in the protein RMSD plot, in which the configuration is not stabilized even after 10 ns (Fig. 5(C)). The RMSF plot also indicates that a few protein side chains show a high degree of freedom (Fig. 5(B)). Consequently, the ligand experienced a drastic change in surroundings, as evidenced by the difference between ligand position at frame 100 and frame 1, and affinity prediction was very difficult in such a scenario. Nevertheless, protein stability could serve as an important criterion for identifying the outliers in the dataset and further improve the accuracy of learning models.

**Fig. 5** The protein with conformational flexibility. (A) The protein demonstrates high structural flexibility in MD simulation. (B) The protein RMSF plot indicates that two side chains undergo a significant conformational change. (C) The RMSD plot of the protein shows that the protein is not stabilized after 10 ns.

**Lead discovery with Dynaformer**

As a proof-of-concept, we experimentally validated that Dynaformer is capable of discovering potent lead compounds by re-scoring the docked molecular structures. The original docking score often has a large number of false positives in the top-scored lists[26]. As a result, one needs to perform hundreds of experiments to identify the true lead compounds. Here, we demonstrate that, by re-scoring the docked poses, Dynaformer could more efficiently deliver the lead compounds.

Specifically, we selected Heat Shock Protein 90 (HSP90) as our target protein. Hsp90 is an important chaperone protein involved in pathways of many refractory diseases, including cancer, neurodegenerative diseases and viral
infections. A common strategy to inhibit HSP90 functioning is to design small molecules targeting its ATP binding pocket to suppress the ATPase activity in the N-terminal domain. Because studies of inhibitors against HSP90 are relatively abundant, there are 18 protein-ligand structures associated with ATPase activity in the training data with $pK_i$ ranging from 3.84 to 8.32. All of these structures could be categorized as the known inhibitor scaffolds summarized in the literature[27, 28].

Subsequently, we performed docking on HSP90 against the ChemBridge DIVERSet-EXP Library containing 50,000 small molecules that cover a diverse pharmacophore space. The docked conformers were obtained using Autodock Vina, a widely used open-source docking software[29, 30]. The reference protein-ligand structure (PDB ID: 2xdl) was chosen from the PDBBind 2016 core set for identification of the binding pocket. The lead discovery protocol can be summarized as follows: 1) The ordinary docking was performed using the Autodock Vina; 2) molecules were re-scored and re-ranked based on the Dynaformer prediction; 3) The top-ranked binders are then clustered by scaffold and visually inspected for prioritization. To ensure the fairness of prioritization, we created a pool consisting of molecules among the top 10% from both Autodock Vina and Dynaformer. From the pool, 20 molecules are chosen by visual inspection without knowing their origin. Finally, the Surface Plasmon Resonance (SPR) experiments were conducted for each of the 20 molecules to measure the inhibition constant $K_i$. The details and results of both dry lab and wet lab experiments are stated in the Supplementary Information.

Fig. 6(A and G) show the structures, experimentally measured $K_i$ and the ranking of the top 10 molecules. Numbers in red indicate the Dynaformer rank is better than Autodock Vina and vice versa. The $K_i$ comparative statistics are summarized in Fig.6(F). It is evident that Dynaformer can identify the lead molecules more efficiently by assigning them higher ranks. Specifically, the Dynaformer ranked the top 3 molecules at 0.82%, 3.71%, and 0.21% respectively, surpassing Autodock Vina by a large margin. In particular, molecule 1, 3, 4 and 6 show $K_i$ of 0.207µM, 1.70µM, 2.50µM and 4.50µM without known precedent in the literature, which can serve as favorable candidates for further lead optimization. The possible binding mode is shown in Fig. 6(B, D and E), which is consistent with findings from rational drug design[31, 32]. For example, a previous study shows that 1) the hydrogen bond networks between inhibitors and residues (ASP-93, THR-184, and LYS-58) are crucial for stable binding; 2) the occupancy by the aromatic groups of lipophilic pocket bottom made up of MET-98, PHE-138, TYR-139, VAL-150, and VAL-186 leads to enhanced affinity and selectivity[27, 28, 33, 34]. In addition, we found some molecules resemble previously reported scaffolds. Molecule 2, 7, 10 and 11 are inhibitors containing the resorcinol-like scaffold from natural products, the possible binding mode of 2 is shown in the Fig. 6(C). Molecule 5 and 9 share a similar backbone that belongs to purine scaffold while molecule 8 is similar to an aminotriazine compound reported previously in a fragment-based screening[28, 35]. Therefore, the wet lab results showed that Dynaformer is
capable of delivering lead compounds with good binding affinity and novel scaffolds.

Fig. 6 HSP90 inhibitors from virtual screen and experimental results. (A) Structures of top 10 compounds. Numbers in red (blue) indicate the higher Dynaformer (Autodock Vina) rank. It is noteworthy that all top-3 compounds were identified by Dynaformer. (B-E) The possible binding mode from docked poses of compounds 1-4, where lead 1(B), 3(D) and 4(E) are arguably considered as novel scaffolds. (F) The $K_i$ comparative statistics of both methods. (G) Experimentally measured $K_i$ and the ranking of the top 10 compounds. It can be seen that Dynaformer is able to identify the lead compounds more efficiently by assigning them a higher ranks.

**Conclusion**

In this work, we presented a graph transformer model to learn the underlying physico-chemical principles of protein-ligand interaction from the MD trajectory. For the first time, we constructed a large-scale dynamic protein-ligand dataset, which encompasses various dynamic effects. We also designed a graph model that encodes geometric-chemical characteristics of the binding process.
As a result, our model outperforms the current state-of-the-art on the benchmark dataset significantly. Here, we manually curated an MD protein-ligand dataset with 3,218 trajectories, but it represents only a small fraction of the available database. Besides, not all protein types are included using the current study, i.e., the membrane protein requires a bi-lipid layer rather than an aqueous environment and many metalloprotein simulations are beyond the capacity of any MD simulation software due to the highly complicated nature of protein-metal interaction. In the future, one might develop a more efficient way to automate the large-scale MD simulation pipeline. Moreover, the MD itself suffers from some drawbacks, such as the inherently approximate nature of the force field calculation and insufficient sampling of MD trajectories. Further progress on the force field accuracy of MD might help improve the quality of the MD dataset.

By re-scoring the docked poses with our model, we showcased the effectiveness of machine learning-based lead compound discovery. We validated that our model can considerably reduce the experimental cost compared to the traditional knowledge-based method. Nevertheless, as the current virtual drug-screening relies on molecular docking, and the ligand database, it is limited both in scoring power and search space. One possible method for further development is to replace structure-based docking with the generative model to establish a de novo virtual screening pipeline, as the generative model could potentially provide new molecules which do not exist in the ligand library. Finally, by fusing molecular dynamics and machine learning, we create not only a symbiotic relationship between the two scientific fields but also provide better tools for describing biological host-guest problems.

Methods

Thermodynamics of binding

In the following, we show that the free energy $\Delta G$ for protein-ligand binding, which determines the binding affinity, can be more accurately described using molecular dynamics trajectories than the 3D static counterpart. As it was shown in Ref.[16, 36], $\Delta G$ can be expressed as the sum of two components, the vacuum binding free energy and the solvation free energy difference:

$$RT\ln(K_i) = \Delta G = \Delta G_{\text{vacuum}} + \Delta G_{\text{solv}}$$

where $R$ and $T$ stand for the universal gas constant and temperature respectively; $K_i$ is the thermodynamic equilibrium constant of the binding process; $\Delta G_{\text{vacuum}}$ is the binding free energy between protein and ligand in the vacuum. This energy can be further split into two parts:

$$\Delta G_{\text{vacuum}} = \langle E_{pl}^{\text{bind}} \rangle - T\Delta S$$
where $\langle E_{\text{bind}}^{\text{pl}} \rangle$ is the ensemble-averaged protein-ligand interaction energy and the term $-T\Delta S$ is the entropic contribution. Based on the interaction entropy scheme [36], $\Delta G_{\text{vacuum}}$ can be expressed as:

$$
\Delta G_{\text{vacuum}} = -RT \ln \left( \frac{\int dq_w dq_p dq_1 e^{-\beta(E_p + E_l + E_{\text{bind}}^{\text{pl}} + E_w + E_{pw} + E_{lw})}}{\int dq_w dq_p dq_1 e^{-\beta(E_p + E_l + E_w + E_{pw} + E_{lw})}} \right)
$$

$$
= -RT \ln \left( \frac{1}{e^{\langle \beta E_{\text{bind}}^{\text{pl}} \rangle}} \right) = RT \ln \langle e^{\beta E_{\text{bind}}^{\text{pl}}} \rangle
$$

$$
= RT \ln \left( e^{\langle \beta E_{\text{bind}}^{\text{pl}} \rangle} (e^{\beta E_{\text{bind}}^{\text{pl}}} - \langle E_{\text{bind}}^{\text{pl}} \rangle) \right)
$$

$$
= \langle E_{\text{bind}}^{\text{pl}} \rangle + RT \ln \langle e^{\Delta E_{\text{bind}}^{\text{pl}}} \rangle,
$$

in which $\beta$ is equal to $1/RT$ and $E_{\text{bind}}^{\text{pl}}$ stands for the interaction energy of protein and ligand (e.g., hydrogen and hydrophobic interaction, Van der Waals interaction, Pi-Pi interactions, long-range electrostatic interactions, etc.); $E_{pw}$, $E_{lw}$ are the interaction energies of protein-water, ligand-water; $E_p$, $E_l$, $E_w$ are internal energies of the protein, ligand, and waters, respectively. The enthalpy terms $\Delta H_i$ equals to $E_{\text{pl}}^i$, $E_{\text{pl}}^i$ is the interaction energy between protein and ligand at state $i$. Meanwhile, the entropy term $\Delta S_i$ equals $e^{\Delta E_{\text{pl}}^i}/kT$, where $\Delta E_{\text{pl}}^i$ represents the deviation of state $i$ and the average interaction energy ($k$ is the Boltzmann constant). Equation 4 proves that the binding free energy can be evaluated using the ensemble average. Such ensemble calculation can be performed by averaging over the MD simulations:

$$
\langle E_{\text{bind}}^{\text{pl}} \rangle = \frac{1}{T} \int_0^T E_{\text{pl}}^{\text{bind}}(t) dt = \frac{1}{N} \sum_{i=1}^N E_{\text{pl}}^{\text{bind}}(t_i),
$$

$$
\langle e^{\beta \Delta E_{\text{pl}}^{\text{bind}}} \rangle = \frac{1}{N} \sum_{i=1}^N e^{\beta \Delta E_{\text{pl}}^{\text{bind}}(t_i)},
$$

where $N$ is the number of MD frames.

Finally, $\Delta G_{\text{solv}}$ is the free energy difference between the protein-ligand complex and the sum of stand-alone protein and ligand systems. In principle, computing $\Delta G_{\text{solv}}$ would require three independent MD simulations of protein, ligand and complex. In order to simplify the calculation, one typical assumption is that no significant conformational changes occur upon binding, so that $\Delta G_{\text{solv}}$ can be approximated using one single frame. Therefore, compared to the 3D static structure, which is essentially a snapshot of the most populated states of the complex. An ensemble of uncorrelated snapshots collected from an equilibrated molecular dynamics simulation can take both enthalpy and entropy into account, presenting a more accurate input feature for predicting protein-ligand binding affinity.
Datasets preparation

**PDBBind dataset.** PDBBind is a public protein-ligand binding benchmark dataset curated from Protein Data Bank (PDB)[37, 38]. PDBBind is a comprehensive crystal structure collection of target proteins, ligands and corresponding experimentally measured binding affinity data. PDBBind dataset consists of three parts: general set, refined set and core set. Structures from PDBBind version 2019 are used to conduct our study. Explicitly, The general set contains totally 21,382 complexes. The refined set is a subset of the general set containing 4,852 complexes with better quality. The core set is a stand-alone test set consisting of 285 high-quality complexes for testing the accuracy of binding affinity prediction. The general set and refined set are updated annually, while the core set is not. Therefore, we evaluate our model and all baseline models on the latest version, PDBBind core set 2016.

**MD dataset.** We choose complexes with relatively high quality in PDBBind refined set as the initial state of MD. We used web-based CHARMM-GUI to build complex systems and prepare their inputs for MD simulation [39, 40]. The MD is performed using NAMD, which is a parallel molecular dynamics program designed for high-performance simulation of large biomolecular systems. The MD simulation is based on the CHARMM36 force field. Each complex is solvated in a truncated periodic TIP3P water box, and the minimum distances from the surfaces of the box to the complex atoms are set to 10 Å. Counter ions are added to neutralize systems, and the initial configuration is decided using a short Monte Carlo simulation. CHARMM ligand topology and parameter file is generated using ParamChem service. The missing residue is modelled using GalxyFill. We used A semi-isotropic Langevin-piston method (piston period of 50 fs) with piston decay of 25 fs and Langevin temperature coupling with a friction coefficient of 1 ps$^{-1}$ to control the temperature and pressure, respectively. The simulation time step was set to 2 fs in conjunction with the SHAKE algorithm to constrain the covalent bonds involving hydrogen atoms. Simulation temperature is maintained at 303.15 K. The simulation time step was set to 2 fs in conjunction with the SHAKE algorithm to constrain the covalent bonds involving hydrogen atoms. Simulation temperature is maintained at 303.15 K. For a stable protein-ligand complex, 2 ns is sufficient for computing free energy calculation. On the other hand, ligand-drift-away events (or other events) become more obvious if MD duration is longer. Therefore, 0.5 ns NVT (constant volume and temperature) is set for equilibrating stage before 10 ns NPT (constant pressure and temperature) dynamics is performed. In practice, such procedure can fail on several occasions: 1) Complex small molecules that lead to failed force field parametrization (e.g., Glycans etc). 2) The protein structure contains too many missing residues. 3) specific type of metallic ion, due to its complex electronic structure (e.g., Fe, Mn, Co etc.). After removing failed cases, totally 3,218 MD simulation are performed, each containing 100 frames.
Model architecture

Dynaformer is a Transformer-based model that learns the protein-ligand binding features from molecular dynamics data. On top of Graphormer\cite{19}, we propose a roto-translation invariant encoding scheme to learn the interaction energy $E_{\text{bind}}^{\text{pl}}$ which associates with the binding affinity. On a higher-level, the Dynaformer predicts the binding affinity given the collective interaction patterns (from molecule dynamics) of a ligand at the active binding site. Given the protein-ligand interaction graph as $G = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{v_0, v_1, v_2, \ldots, v_n\}$ and $\mathcal{E} = \{e_{i,j} | i, j \in \mathcal{V}\}$, where $n$ represents the number of atoms and $v_i, e_{ij}$ represents the input atom or edge features, the design of Dynaformer is described as follows. Note that $v_0$ is a virtual node for aggregating information from all other nodes to predict the binding affinity.

**Enthalpic motivation of Dynaformer.** As mentioned before, The interaction energy of protein and ligand $E_{\text{bind}}^{\text{pl}}$ is directly associated with the binding affinity through Equation 4. $E_{\text{bind}}^{\text{pl}}$ mainly consist of the non-covalent bonds between them, which can be fully represented by distances between atoms and angles between bonds. To encode the $E_{\text{bind}}^{\text{pl}}$, we exploit Gaussian Basis Function (GBF), which can be defined as Equation 7, where $\mu$ and $\sigma$ are learnable scalars and $x$ are embedded scalars representing the encoded distances $d_{\chi(i,j)}$ or angles $s_{\psi(i,j)}$. Two functions $d_{\chi(i,j)}$ and $s_{\phi(i,j)}$ have the form of $\cdot:\mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ that encode the distances or angles and the corresponding atomic features. Hence, the inter-atomic energy (i.e., the enthalpic contribution) can be encoded by Equation 8,9:

$$\text{GBF}(x) = \frac{1}{2\pi\sigma} \exp \left[-\frac{(x - \mu)^2}{2\sigma^2}\right],$$

$$d_{\chi(i,j)} = \text{GBF}([v_i|d_{ij}|v_j] W_\chi),$$

$$s_{\psi(i,j)} = \text{GBF} \left( \left[v_i | \sum_k \angle ij k | v_j \right] W_\psi \right),$$

where $W_\chi$ and $W_\psi$ are the parameter matrices for linear transformation.

**Dynaformer layer.** In each encoder layer, Dynaformer has two components, a multi-head self-attention module (MHA) and a feed-forward network (FFN). The encoded distances and angles are fused into MHA module which is used to update atom representation and the energy contribution of each atom. In each head of MHA, the attention is calculated as Equation 10,11, where $Q = HW_Q, K = HW_K, V = HW_V$ are linear transformed atom representations denoted by $H = [h_0, \ldots, h_n]$. The attention module is used for updating atomic representation by a weighted summation from all atoms. Similar to $d_{\chi(i,j)}$ and $a_{\psi(i,j)}$, $s_{\phi(i,j)}$ is an additional attention bias, which are learnable scalars indexed by function $\phi(i,j) = [v_i|\text{SPD}_{ij}|v_j] W_\psi$ where $\text{SPD}_{ij}$ is the shortest path distance between node $i$ and $j$ on the interaction graph. The other term $c_{ij} = \frac{1}{N} \sum_{n=1}^{N} e_n w_n^T$ encodes the chemical bonds on the shortest
path between node $i$ and $j$, where $w_n$ denotes the weight that projects $e_n$ to a scalar.

$$\text{Attention}(Q, K, V) = \text{softmax}(A) \cdot V$$  \hspace{1cm} (10)$$

$$A_{ij} = \left( h_i W_Q \right) \left( h_j W_K \right)^T \frac{1}{\sqrt{d_K}} + d_{\chi(i,j)} + a_{\psi(i,j)} + s_{\phi(i,j)} + c_{ij}$$  \hspace{1cm} (11)$$

**Predictions through the virtual node.** A virtual node is a meta node that collects information for affinity prediction from all other nodes. In every layer of Dynaformer, other nodes (atoms) gradually aggregate information to the virtual node through Equation 10, and at the last layer, the representation of the virtual node is exploited to predict the binding affinity by a linear layer. Further, to extend the receptor field of the binding pocket, three fingerprint vectors are calculated and fused into the virtual node[41–43]. By design, such a scheme allows the Dynaformer to automatically learn the effective structural factors that strongly associate with binding free energy through dynamic trajectories, and the frames in the trajectories also serve as a rational data augmentation approach in the model training.

**Ablation study.** Details can be found in the Supplemental Information.

**Evaluation Metrics**

**Binding affinity.** In this work, the binding affinity of a protein-ligand pair is the dissociation constant ($K_d$) or more narrowly defined, inhibition constant ($K_i$). Both $K_d$ and $K_i$ are equilibrium constants that describe how potent a ligand is binding to a protein, as shown below. In this paper, we use the negative logarithm of $K_d$ or $K_i$, i.e. $-\log_{10}K_d$ or $-\log_{10}K_i$ ($pK_d$ or $pK_i$) to represent the binding affinity.

$$[PL] \xrightleftharpoons{K_d/K_i} [P] + [L]$$

**Pearson $r$.** Also known as the Pearson correlation coefficient, Pearson $r$ measures the linear relationship between two sequences of data $\{x_i\}$ and $\{y_i\}$. It is defined as:

$$r(x, y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

where $N$ is the number of samples, $\bar{x}, \bar{y}$ are the mean values of $x$ and $y$, respectively. $r$ varies between $[-1, 1]$ with 0 indicating no correlation and $-1$ or 1 indicating exact negative or positive linear relationship. This metric is widely used in the virtual screen where rank order is crucial. Note that Pearson $r$ is invariant to the scale nor location of the two variables, which means it does not suggest the error between prediction and groundtruth.

**MAE, MSE and RMSE.** As a complementary part of Pearson $r$, we use
MAE, MSE and RMSE to determine the average error of binding affinity prediction. They are defined as:

$$\text{MAE}(x, y) = \frac{1}{N} \sum_{i=1}^{N} |x_i - y_i|$$

$$\text{MSE}(x, y) = \text{RMSE}^2(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2$$

where $N$ is the number of samples, $\{x_i\}$ and $\{y_i\}$ are the predictions and the groundtruth binding affinity.

**RMSD and RMSF.** The RMSD measures the average deviation of a matching set of atoms between two structures. It can be expressed as:

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta_i^2}$$

where $\delta_i$ is the euclidean distance between the atom $i$ and the reference position. Whereas the RMSF is the fluctuation around an average, per atom or residue, over a sequence of structures (e.g., from a trajectory).

$$\text{RMSF} = \sqrt{\langle (x_i - \bar{x}_i)^2 \rangle}$$

where $x_i$ is the coordinates of the particle $i$, and $\bar{x}_i$ is the ensemble average position of $i$. Where the RMSD quantifies how much a structure diverges from a reference over time, the RMSF can measure the mobility of parts of the system. While RMSD is frequently calculated using the initial state, the RMSF should be calculated to an average structure of the simulation. An area of the structure with high RMSF values frequently diverges from the average, indicating high mobility. When RMSF analysis is carried out on proteins, it is typically restricted to backbone or alpha-carbon atoms; these are more characteristic of conformational changes than the more flexible side-chains.

**References**

[1] Wodak, S.J., Janin, J.: Structural basis of macromolecular recognition. Advances in protein chemistry 61, 9–73 (2002). https://doi.org/10.1016/s0065-3233(02)61001-0

[2] Deremble, C., Lavery, R.: Macromolecular recognition. Current opinion in structural biology 15(2), 171–175 (2005). https://doi.org/10.1016/j.sbi.2005.01.018
[3] Cournia, Z., Allen, B., Sherman, W.: Relative binding free energy calculations in drug discovery: Recent advances and practical considerations. Journal of Chemical Information and Modeling 57(12), 2911–2937 (2017) https://doi.org/10.1021/acs.jcim.7b00564. https://doi.org/10.1021/acs.jcim.7b00564. PMID: 29243483

[4] Mayr, L.M., Bojanic, D.: Novel trends in high-throughput screening. Current Opinion in Pharmacology 9(5), 580–588 (2009). https://doi.org/10.1016/j.coph.2009.08.004. Anti-infectives/New technologies

[5] Bender, B.J., Gahbauer, S., Luttens, A., Lyu, J., Webb, C.M., Stein, R.M., Fink, E.A., Balius, T.E., Carlsson, J., Irwin, J.J., Shoichet, B.K.: A practical guide to large-scale docking. Nature Protocols 16(10), 4799–4832 (2021). https://doi.org/10.1038/s41596-021-00597-z

[6] Woo, H.-j., Roux, B.: Calculation of absolute protein–ligand binding free. Proceedings of the National Academy of Sciences 102(19), 6825–6830 (2005)

[7] Frenkel, Daan, B.S.: Understanding molecular simulation: from algorithms to applications. Vol. 1. Elsevier (2001)

[8] Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D.A., Cheatham, T.E.: Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. Accounts of Chemical Research 33(12), 889–897 (2000). https://doi.org/10.1021/ar000033j

[9] Gilson, M.K., Zhou, H.-X.: Calculation of protein-ligand binding affinities. Annual Review of Biophysics and Biomolecular Structure 36(1), 21–42 (2007) https://doi.org/10.1146/annurev.biophys.36.040306.132550. https://doi.org/10.1146/annurev.biophys.36.040306.132550. PMID: 17201676

[10] Lo, Y.-C., Rensi, S.E., Torng, W., Altman, R.B.: Machine learning in chemoinformatics and drug discovery. Drug Discovery Today 23(8), 1538–1546 (2018). https://doi.org/10.1016/j.drudis.2018.05.010

[11] Zheng, L., Fan, J., Mu, Y.: OnionNet: A Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein-Ligand Binding Affinity Prediction. ACS Omega 4(14), 15956–15965 (2019) arXiv:1906.02418. https://doi.org/10.1021/acsomega.9b01997

[12] Liu, X., Feng, H., Wu, J., Xia, K.: Persistent spectral hypergraph based machine learning (PSH-ML) for protein-ligand binding affinity prediction. Briefings in Bioinformatics 22(5), 1–11 (2021). https://doi.org/10.1093/bib/bbab127
[13] Li, S., Zhou, J., Xu, T., Huang, L., Wang, F., Xiong, H., Huang, W., Dou, D., Xiong, H.: Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 975–985 (2021) arXiv:2107.10670. https://doi.org/10.1145/3447548.3467311

[14] Ballester, P.J., Mitchell, J.B.O.: Europe PMC Funders Group Europe PMC Funders Author Manuscripts A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking 26(9), 1169–1175 (2012). https://doi.org/10.1093/bioinformatics/btq112.A

[15] Rube, H.T., Rastogi, C., Feng, S., Kribelbauer, J.F., Li, A., Becerra, B., Melo, L.A.N., Do, B.V., Li, X., Adam, H.H., Shah, N.H., Mann, R.S., Bussemaker, H.J.: Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. Nature Biotechnology (2022). https://doi.org/10.1038/s41587-022-01307-0

[16] Du, X., Li, Y., Xia, Y.L., Ai, S.M., Liang, J., Sang, P., Ji, X.L., Liu, S.Q.: Insights into protein–ligand interactions: Mechanisms, models, and methods. International Journal of Molecular Sciences 17(2), 1–34 (2016). https://doi.org/10.3390/ijms17020144

[17] Bissantz, C., Kuhn, B., Stahl, M.: A Medicinal Chemist’s Guide to Molecular Interactions. Journal of Medicinal Chemistry 53(14), 5061–5084 (2010). https://doi.org/10.1021/jm100112j

[18] Zhou, H.-X., Gilson, M.K.: Theory of Free Energy and Entropy in Noncovalent Binding. Chemical Reviews 109(9), 4092–4107 (2009). https://doi.org/10.1021/cr800551w

[19] Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.Y.: Do Transformers Really Perform Bad for Graph Representation? Advances in Neural Information Processing Systems 34(NeurIPS), 28877–28888 (2021) arXiv:2106.05234

[20] Phillips, J.C., Hardy, D.J., Maia, J.D.C., Stone, J.E., Ribeiro, J.V., Bernardi, R.C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M.C.R., Radak, B.K., Skeel, R.D., Singharoy, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., Kalé, L.V., Schulten, K., Chipot, C., Tajkhorshid, E.: Scalable molecular dynamics on CPU and GPU architectures with NAMD. The Journal of chemical physics 153(4), 44130 (2020). https://doi.org/10.1063/5.0014475

[21] Modugno, M., Casale, E., Soncini, C., Rosettani, P., Colombo, R., Lupi, R., Rusconi, L., Fancelli, D., Carpinelli, P., Cameron, A.D., Isacchi, A.,
Moll, J.: Crystal Structure of the T315I Abl Mutant in Complex with the Aurora Kinases Inhibitor PHA-739358. Cancer Research 67(17), 7987–7990 (2007). https://doi.org/10.1158/0008-5472.CAN-07-1825

[22] Huschmann, F.U., Linnik, J., Sparta, K., Ühlein, M., Wang, X., Metz, A., Schiebel, J., Heine, A., Klebe, G., Weiss, M.S., Mueller, U.: Structures of endothiapepsin-fragment complexes from crystallographic fragment screening using a novel, diverse and affordable 96-compound fragment library. Acta Crystallographica Section F: Structural Biology Communications 72, 346–355 (2016). https://doi.org/10.1107/S2053230X16004623

[23] Wilson, D.P., Wan, Z.K., Xu, W.X., Kirincich, S.J., Follows, B.C., Joseph-McCarthy, D., Foreman, K., Moretto, A., Wu, J., Zhu, M., Binnun, E., Zhang, Y.L., Tam, M., Erbe, D.V., Tobin, J., Xu, X., Leung, L., Shilling, A., Tam, S.Y., Mansour, T.S., Lee, J.: Structure-based optimization of protein tyrosine phosphatase 1B inhibitors: From the active site to the second phosphotyrosine binding site. Journal of Medicinal Chemistry 50(19), 4681–4698 (2007). https://doi.org/10.1021/jm0702478

[24] Carlson, H.A.: Protein flexibility and drug design: how to hit a moving target. Current opinion in chemical biology 6(4), 447–452 (2002). https://doi.org/10.1016/s1367-5931(02)00341-1

[25] Nilsson, M., Hämäläinen, M., Ivarsson, M., Gottfries, J., Xue, Y., Hansson, S., Isaksson, R., Fex, T.: Compounds binding to the S2-S3 pockets of thrombin. Journal of Medicinal Chemistry 52(9), 2708–2715 (2009). https://doi.org/10.1021/jm8011849

[26] Lyu, J., Wang, S., Balius, T.E., Singh, I., Levit, A., Moroz, Y.S., O’Meara, M.J., Che, T., Alგaa, E., Tolmachova, K., et al.: Ultra-large library docking for discovering new chemotypes. Nature 566(7743), 224–229 (2019)

[27] Janin, Y.L.: Atpase inhibitors of heat-shock protein 90, second season. Drug discovery today 15(9-10), 342–353 (2010)

[28] Li, L., Wang, L., You, Q.-D., Xu, X.-L.: Heat shock protein 90 inhibitors: an update on achievements, challenges, and future directions. Journal of medicinal chemistry 63(5), 1798–1822 (2019)

[29] Eberhardt, J., Santos-Martins, D., Tillack, A.F., Forli, S.: Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. Journal of Chemical Information and Modeling 61(8), 3891–3898 (2021)

[30] Trott, O., Olson, A.J.: Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry 31(2), 455–461
[31] Kung, P.-P., Funk, L., Meng, J., Collins, M., Zhou, J.Z., Johnson, M.C., Ekker, A., Wang, J., Mehta, P., Yin, M.-J., et al.: Dihydroxylphenyl amides as inhibitors of the hsp90 molecular chaperone. Bioorganic & medicinal chemistry letters 18(23), 6273–6278 (2008)

[32] Murray, C.W., Carr, M.G., Callaghan, O., Chessari, G., Congreve, M., Cowan, S., Coyle, J.E., Downham, R., Figueroa, E., Frederickson, M., et al.: Fragment-based drug discovery applied to hsp90. discovery of two lead series with high ligand efficiency. Journal of medicinal chemistry 53(16), 5942–5955 (2010)

[33] Stebbins, C.E., Russo, A.A., Schneider, C., Rosen, N., Hartl, F.U., Pavletich, N.P.: Crystal structure of an hsp90–geldanamycin complex: targeting of a protein chaperone by an antitumor agent. Cell 89(2), 239–250 (1997)

[34] Yoshimura, C., Nagatoishi, S., Kuroda, D., Kodama, Y., Uno, T., Kitade, M., Chong-Takata, K., Oshiumi, H., Muraoka, H., Yamashita, S., et al.: Thermodynamic dissection of potency and selectivity of cytosolic hsp90 inhibitors. Journal of Medicinal Chemistry 64(5), 2669–2677 (2021)

[35] Huth, J.R., Park, C., Petros, A.M., Kunzer, A.R., Wendt, M.D., Wang, X., Lynch, C.L., Mack, J.C., Swift, K.M., Judge, R.A., et al.: Discovery and design of novel hsp90 inhibitors using multiple fragment-based design strategies. Chemical biology & drug design 70(1), 1–12 (2007)

[36] Duan, L., Liu, X., Zhang, J.Z.H.: Interaction entropy: A new paradigm for highly efficient and reliable computation of protein-ligand binding free energy. Journal of the American Chemical Society 138(17), 5722–5728 (2016). https://doi.org/10.1021/jacs.6b02682

[37] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic acids research 28(1), 235–242 (2000). https://doi.org/10.1093/nar/28.1.235

[38] Wang, R., Fang, X., Lu, Y., Yang, C.-Y., Wang, S.: The PDBbind Database: Methodologies and Updates. Journal of Medicinal Chemistry 48(12), 4111–4119 (2005). https://doi.org/10.1021/jm048957q

[39] Jo, S., Kim, T., Iyer, V.G., Im, W.: CHARMM-GUI: a web-based graphical user interface for CHARMM. (2008). https://doi.org/10.1002/jcc.20945

[40] Lee, J., Cheng, X., Swails, J.M., Yeom, M.S., Eastman, P.K., Lemkul, J.A., Wei, S., Buckner, J., Jeong, J.C., Qi, Y., Jo, S., Pande, V.S., Case,
D.A., Brooks, C.L., MacKerell, A.D., Klauda, J.B., Im, W.: CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. Journal of Chemical Theory and Computation 12(1), 405–413 (2016). https://doi.org/10.1021/acs.jctc.5b00935

[41] Sánchez-Cruz, N., Medina-Franco, J.L., Mestres, J., Barril, X.: Extended connectivity interaction features: improving binding affinity prediction through chemical description. Bioinformatics 37(10), 1376–1382 (2021)

[42] Ballester, P.J., Mitchell, J.B.: A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 26(9), 1169–1175 (2010)

[43] Rayka, M., Firouzi, R.: Gb-score: Minimally designed machine learning scoring function based on distance-weighted interatomic contact features. ChemRxiv (2022). https://doi.org/10.26434/chemrxiv-2022-f2tqx

Supplementary information. The supplementary material is available in the attachment.

Acknowledgements. We thank Hao Zhang and Tristan Bereau for the insightful discussions.

Author contributions. Y.W. conceived the idea; J.W. and J.Y.Z. planned the study; Y.M., Y.W., P.W., Z.X., and J.W. designed the machine learning model; Y.M. and P.W. implemented the algorithm; Y.W. developed the molecular dynamics protocol; Y.W., P.W., and Y.M. prepared the dataset; D.Z. and X.W. performed the biological experiments; Y.W. and N.W. designed and produced final figures; all authors participated discussion and commented on the results.

Competing interests. Authors declare that they have no competing interests.

Code availability. The code is available at https://github.com/Minys233/Dynafomer