FeatureDock: Protein-Ligand Docking Guided by Physicochemical Feature-Based Local Environment Learning using Transformer

Mingyi Xue[1,2], Bojun Liu [1,2], Siqin Cao[1,2], Xuhui Huang[1,2*]

[1]Department of Chemistry, Theoretical Chemistry Institute, University of Wisconsin-Madison, Madison, WI, 53706, USA

[2]Data Science Institute, University of Wisconsin-Madison, Madison, WI, 53706, USA

[*]To whom correspondence should be addressed. E-mail: xhuang@chem.wisc.edu

Abstract

Molecular docking, the task of predicting the binding structures between a protein and a small molecule ligand, plays a significant role in structural-based drug discovery. In recent years, numerous deep learning-based methods for molecular docking have emerged. State-of-the-art approaches such as DiffDock formulate the docking problem using diffusion generative models, exhibiting superior performance than traditional docking algorithms. However, despite the strong performance of these deep learning-based docking methods in predicting binding poses, they often lack a well-defined scoring function. This limitation poses challenges in effectively distinguishing between the strong and weak inhibitors during virtual screening. To address this limitation, we introduce FeatureDock, a transformer-based deep learning framework, which can accurately predict the protein-ligand binding poses as well as achieve a strong scoring power for virtual screening. FeatureDock extracts chemical features from local environments within protein structures and utilizes a Transformer encoder to predict probability density envelopes indicating where ligands are most likely to bind in the protein pocket. We also designed a scoring function, which encodes the predicted probability density envelope, to optimize and score the ligand poses. In addition, the attention mechanism in FeatureDock's Transformer further enhances the model's interpretability by providing the attention weights of each chemical feature from the protein structures in predicting the binding probabilities. When applied to virtual screening, we demonstrated that FeatureDock outperforms DiffDock, Smina and AutoDock Vina in distinguishing strong inhibitors from weak ones for both Cyclin-Dependent Kinase 2 (CDK2, an inactivated form) and Angiotensin-converting enzyme (ACE). The performance was assessed using Kullback–Leibler (KL) divergence and area under receiver operating characteristic (AUC) evaluation metrics. We also demonstrate that FeatureDock can accurately predict the binding poses, achieving an average RMSD of 2.4 Å when compared to CDK2-ligand co-crystal structures. We anticipate that our FeatureDock holds promise to be widely applied in virtual screening to assist in drug design. FeatureDock is available at https://github.com/xuhuihuang/featuredock.

1. Introduction

Drug discovery is a complicated, costly, and time-consuming task that typically takes pharmaceutical companies over a decade and billions of dollars to bring a novel drug to market[1, 2]. The high-throughput screening (HTS) methods take considerable amount of time to evaluate a compound library that typically contains millions of compounds[3]. Considering the number of currently available compounds being ~10^8 in small molecule libraries such as PubChem[4, 5], ZINC[6, 7] and ChEMBL[8], numerous *in silico* methods evolved to virtually screen potential compounds against the target to alleviate the heavy duty in HTS.

In the computer-aided and rational drug design, docking-based methods[9-12] provide powerful tools for predicting the positions, orientations, and conformations of ligand binding. One of the critical principles of traditional docking methods is to design an accurate scoring function that can be applied to score and explore the optimal ligand binding poses. For example, DOCK[13-15] and AutoDock[16] used physics-based scoring functions consisting of terms for van der Waals (vdW) interactions, electrostatic energy, hydrogen bonding interactions, etc. These physics-based scoring functions consider inter- and intra-molecular chemical interactions, but they are often computationally expensive and suffer from the inaccurate descriptions of polarization, solvation and the entropy of protein and ligand[17, 18]. Empirical scoring functions can address the above limitations of physics-based scoring functions by fitting parameters to the experimental binding affinities using pre-selected descriptors[19, 20]. For example, AutoDock Vina[21] utilizes a fully empirical scoring function, represented as vdW-like potential, hydrophobic interactions, conformational entropy, etc. Smina[22] designed a user-friendly software interface to support custom scoring functions based on AutoDock Vina. Other derivatives of AutoDock Vina further improved the scoring function of AutoDock Vina for specific systems, e.g., AutoDock VinaXB[23] that deals with halogen bonds and Vina-Carb[24] that deals with carbohydrate systems.

The aforementioned docking methods employ genetic algorithms (GA)[13, 14, 16] or Monte-Carlo (MC) annealing[21, 23-27], combined with gradient-based optimization, to sample and refine the ligand binding poses. More recent studies have integrated deep learning models to optimize ligand poses in one-stage[28-30], significantly speeding up the docking process. For example, EquiBind[30] and TankBind[29] adopted E(3)-equivariant graph neural networks (E(3)-GNN) to effectively represent both the protein and ligand structures. They then utilized a graph matching network (GMN) to predict the accurate rigid SE(3) transformation for the ligand docking. Additionally, a fine-tuning model was included to optimize torsion angles of rotatable bonds, enabling flexible docking. Unlike EquiBind and TankBind, which treated the docking task as a regression problem, DiffDock[28] took a different approach by harnessing diffusion-based generative models. Specifically, it leveraged a denoising diffusion process to generate ligand conformations and binding poses from the initial ligand structure with randomized torsional angles and mean-removed 3D coordinates. DiffDock included a supervised model to classify whether the generated poses are close to the ground truth based on the model prediction called confidence score.

The scoring functions of docking methods are often inadequate in virtual screening, particularly when distinguishing strong binders from weak binders. This is because the scoring functions of docking methods often exhibit weak or no correlation between docking scores and binding affinities[31, 32]. This is evidenced by the low Pearson correlation coefficients (R^c) between docking scores and binding affinities for several commonly used docking software evaluated on the CASF-2016 dataset[32]. For example, AutoDock Vina, GOLD[33], MOE[34] and Glide[35, 36] achieved R^c of 0.604, [0.416, 0.617], [0.405, 0.591] and [0.467, 0.513] respectively, where the range indicates that more than one scoring function are evaluated in the software. As a result, these docking scoring functions may result in high false positives when selecting true inhibitors by considering the highest-scored compounds as inhibitors from a compound library.

Recently, several machine-learning based scoring functions[12, 32, 37] have engaged more comprehensive descriptors and more advanced neural network architectures to predict experimental binding affinities (e.g., K_d values), thus improving the scoring power. However, these machine learning methods were designed to re-score the docked pose, which could not apply to the scenario without 3D protein-ligand structures. For example, 3D CNN models like Atomic Convolutional Neural Networks (ACNN)[38] and K_{DEEP}[39] treated protein-ligand complex as atom property tensors. Furthermore, distance-aware graph neural networks (GNN) like PotentialNet[40] utilized the more efficient graph representation to encode Euclidean symmetries in structures.

In this study, we introduce a physicochemical feature-based deep-learning approach, FeatureDock, to score and pose small-molecular ligands binding to proteins. In FeatureDock, the potential binding pockets are firstly discretized into grid points, embedded using 3D-invariant FEATURE[41] representations. These representations have been shown sufficiently descriptive to extract the local chemical environment for protein-RNA binding[42]. Then, the state-of-the-art Transformer encoder[43] is adopted to predict probability density envelopes, formed by grid points in the binding pockets with high probabilities for ligands to occupy. The attention mechanism in Transformer further helps explain and visualize the importance of input features by extracting the attention weights. Using the probability density envelopes, we design a novel scoring function combined with a position optimization algorithm to screen and pose ligands in binding pockets. To demonstrate the performance of FeatureDock, we have applied it to select strong inhibitors from bioassay datasets[8] for two different protein systems: an inactivated form of Cyclin-Dependent Kinase 2 (CDK2) and Angiotensin-converting enzyme (ACE). We show that our scoring function exhibits a superior ability to differentiate between strong and weak inhibitors compared to DiffDock[28], Smina[22] and AutoDock Vina[21], as evaluated through the KL divergence and AUC evaluation. Notably, our model can accurately retrieve the binding poses of top-predicted ligands after the optimization process based on our scoring function. These binding poses closely approximate their native structures in cocrystal configurations, demonstrating a high level of proximity for CDK2.

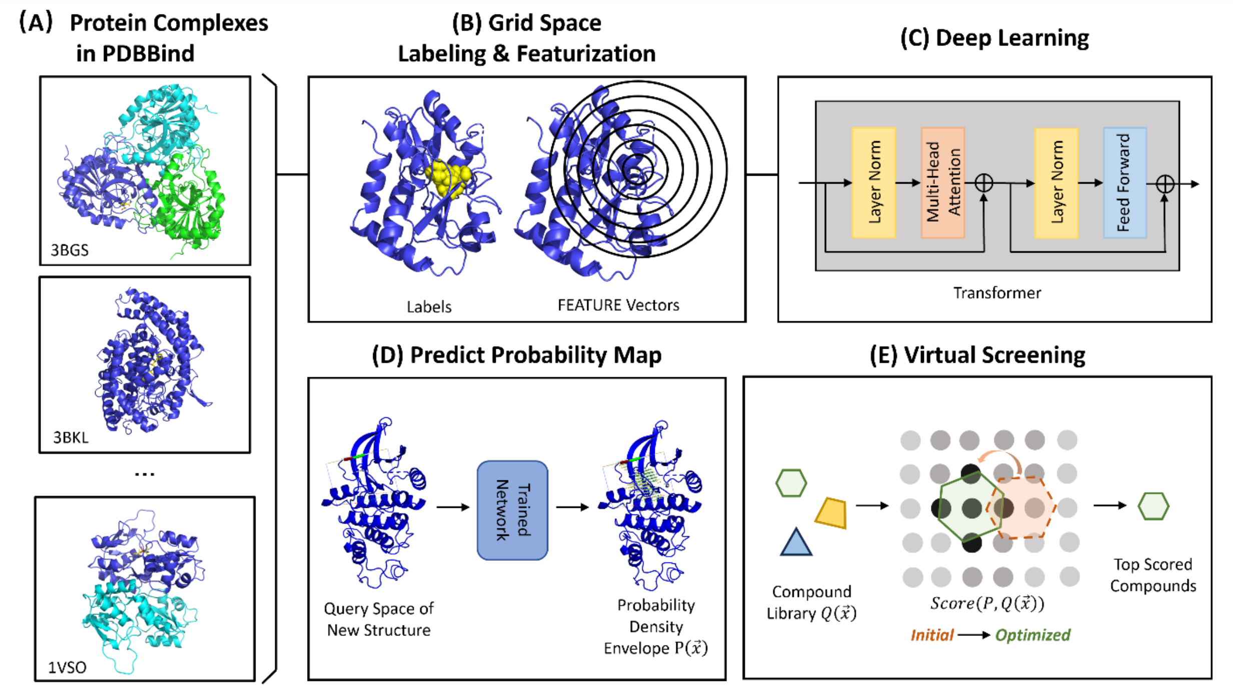

Figure 1. **FeatureDock Pipeline.** (A) Collect protein-ligand complexes from PDBBind v2020 refined set. (B) Extract and featurize protein local environments around grid points in the ligand-binding pocket and then label the grid points as either binding or non-binding with the ligand. (C) Train the Transformer Encoder to predict the ligand-binding probability of each grid point. (D) Predict the probability density envelope for the query space in *apo* proteins using the trained model. (E) Apply the predicted probability density envelope to virtual screening and pose prediction.

2. Methods

We outline the pipeline for our FeatureDock framework in Fig 1. Initially, the ligand-binding pockets of protein complexes in the PDBBind v2020 refined set[44] are discretized into grid points, embedded using 3D-invariant FEATURE representations (Fig 1A-B). We then train a Transformer encoder model as shown in Fig 1C. After training, our model predicts the binding probabilities of grid points in the query pockets of a given *apo* protein structure, forming probability density envelopes (Fig 1D). Using these probability density envelopes, we designed a scoring function combined with a position optimization algorithm to predict and score the binding poses (Fig 1E). We will describe these steps in detail below:

2.1. Dataset curation

Protein representation and data labeling: The training dataset was curated from the PDBBind v2020 refined set[44] (Fig 1A), comprising 5,316 high-quality protein-ligand complex structures that have been cleaned, fixed, and properly protonated from raw structures deposited in RCSB PDB[45]. Notably, ligands in the refined set are all non-covalently binding to the proteins. We extracted the convex hull formed by protein residues within 6Å around each ligand. This space was discretized to grid points spaced by 1 Å. The grid points too close to (< 1Å) or too far away from (> 6Å) the protein atoms were removed to reduce the dataset size because ligand atoms are less likely to show up in these regions.

Protein-ligand conformations are unstructured data compared to image tensors, thus requiring delicate data preprocessing before being fed into neural networks. Having a good protein representation is one of the determining factors to make sure that the model learns meaningful information. In our previous work focused on predicting protein-RNA interactions[42], FEATURE vectors[41] have been proven descriptive enough to capture protein local environments. Within each spheric shell, FEATURE vectors encode 80 physicochemical properties (Table S1) across various hierarchies, including atomic level, residue-level, and secondary structural level information. Since FEATURE vectors are invariant to translational, rotational and permutational changes of protein structures, there is no requirement of dataset augmentation to make models robust against these operations. Furthermore, we will show that fine-tuning of the model on a specific protein family is not necessary because the model trained on various proteins has the generalization ability to predict novel structures, as shown in the Results section 3.1.

By default, the local environment surrounding each grid point was characterized by 6 concentric shells (Fig 1B). Centered at each grid point, these shells were defined with a width of 1.25 Å each, capturing properties in a local environment up to a radius of 7.5 Å. FEATURE vectors, were then applied to these concentric shells with the 80 pre-defined physicochemical properties. This featurization resulted in a 6x80 tensor for each grid point. Grid points were assigned with labels related to the protein-ligand binding properties. Here we chose a categorical label, which tells whether there is a ligand heavy atom within 1.5 Å of the grid point (Fig 1B).

After extracting features from existing protein-ligand cocrystal structures and curating a labeled dataset for supervised learning, the Transformer encoder can be trained (Fig 1C) and then applied to predict which space is more likely occupied by ligands in a data-driven manner for novel *apo* structures (Fig 1D). Later, the predicted probability density envelope formed by the grid points and probabilities will be used in the scoring function during virtual screening (Fig 1E).

Dataset split: To avoid trivial queries and predictions during the model validation, we split the whole curated dataset (containing 4,515 structures after pre-processing) into training/validation/test set based on their sequence similarities. Structures were clustered to 1,326 structure clans (Fig S1) based on 90% MMseq2[46] protein sequence similarity fetched from RCSB PDB[45]. 10% of the protein clusters were randomly selected as the validation set and the remaining were used for training. This split method guarantees that homogeneous structures exist in either the training or validation dataset but not both. By adopting this dataset split strategy, the evaluation metrics of the validation set can better reflect the model performance and avoid overfitting. During the validating of model generality (the leave-out models), models were trained on the dataset after leaving out one certain type of protein (e.g., CDK2) and its homogeneous structures (sequence identity >= 90%) to mimic the scenario of applying the trained model to predict novel protein structures. After validating that our model could achieve good performance on predicting novel structures, we trained full models with all structure clusters and used the full-model predictions in the model application of virtual screening and pose prediction.

Data resampling for mitigating class imbalance: In this section, we discuss the class imbalance in available training protein complex structures. One source of imbalance comes from the fact that some protein families are better studied than others, leading to the structure abundance difference. For example, ACE proteases have <10 structures in the PDBBind refined dataset, CDK2 and CDK2/Cyclin have ~20 structures, while many other proteins have only one single structure (Fig S1). Another source comes from the fact that grid points with negative labels (non-binding regions) are more abundant than those with positive labels (binding regions) because we selected a relatively large pocket compared to the regions that the ligand occupy in the step of dataset curation. We adopted two data resampling strategies to address the class imbalance issue. The first is to sample each protein cluster uniformly instead of sample each protein structure uniformly. The model can therefore have better generalization capability by sampling local environments from rare protein clusters more frequently. The second is to balance the abundance of different labels by under-sampling negative datapoints and over-sampling positive datapoints.

2.2. Neural network architectures

FeatureDock used the Transformer encoder (Fig 1C) to predict the ligand binding probabilities of grid points because it consistently outperformed Feed Forward Neural Network and ResNet across various model sizes. To elucidate that Transformer encoders could achieve better performance, we compared different neural network architectures at different model sizes, and evaluated their performances by the loss, area under ROC (AUC), F1 Score and Matthews correlation coefficient (MCC) on the validation set.

Transformer encoder

Transformer[43] was first proposed in 2018 for natural language processing (NLP) tasks. It pioneered the use of attention mechanism to capture the temporal correlations in token sequences, enabling the efficient parallel computing compared to traditional recurrent neural networks, long short-term memory (LSTM)[47], etc. The success of Transformer has also inspired significant advancements in the field of computer vision (CV). Various transformer-based models, e.g., vision transformer (ViT)[48] and swin transformer[49], have been developed, outperforming most of traditional CNNs in image classification tasks. Compared to convolutional operations that capture local spatial correlation, the attention mechanism in Transformer calculates all-to-all correlations among the input features. Transformer has also gained its popularity in the drug discovery field. For example, Transformer has been adopted to encode protein sequences and generate SMILES of potential drugs, or to study drug-target interactions (DTI)[50].

We leveraged the Transformer encoder to predict the ligand-binding probability of each grid point based on the physicochemical properties of its local environment. Specifically, we conceptualized each physicochemical property as analogous to a word (Fig 2). Values of the same property in different concentric spheritic shells (total of 6 shells) were aggregated into a single word (consisting of 6 characters in each word). The class token[48, 51] was concatenated with the 80 properties to form an 81-word sentence $\boldsymbol{h_0}$ for the classification purpose. The 81-word sentence was additionally encoded with positional embeddings and fed to the encoder blocks. The final embedding of the class token can be viewed as the representation of the grid point after propagating information of input properties. It was then fed into the Softmax layer which outputs the binding probability.

Here we denote the input of each encoder block (Fig S3 and Eq. 1) as $\boldsymbol{h_i} \in \mathbb{R}^{\boldsymbol{l \times d}}$ and the output as $\boldsymbol{h_{i+1}} \in \mathbb{R}^{\boldsymbol{l \times d}}$, which will then serve as the input of the next block. l is the sequence length ($l = 81$), and d is the dimension of each word (i.e., chemical property) in the sequence. In each block, the input $\boldsymbol{h_i}$ is projected to the query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), value ($\boldsymbol{V}$) matrices, parameterized by $\boldsymbol{W_Q}, \boldsymbol{W_K}, \boldsymbol{W_V} \in \mathbb{R}^{\boldsymbol{d \times d'}}$. The attention mechanism utilizes $\boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{\boldsymbol{l \times d'}}$ to calculate a word-to-word weight matrix $\boldsymbol{S} \in \mathbb{R}^{\boldsymbol{l \times l}}$, where $\boldsymbol{S}(j, k)$ represents the weight of word k contributing to word j. The updated hidden state $\boldsymbol{h'_{i+1}} \in \mathbb{R}^{\boldsymbol{l \times d'}}$ (d' is the dimension of the hidden state) is a weighted combination of the value matrix $\boldsymbol{V}$, given the attention weight matrix $\boldsymbol{S}$. The final output of the block $\boldsymbol{h_{i+1}}$ will be calculated using $\boldsymbol{h'_{i+1}}$ and $\boldsymbol{h_i}$ via residual connection. The attention weight from physicochemical properties to the class token can be used to further enhance model interpretability by visualizing contributing weights (see Fig 4 for examples). In implementation, we utilized multi-head attention that projects the queries, keys and values h times parallelly, and concatenated their outputs (see Fig S3 for details). When calculating the contribution of features to ligand binding, we extracted and averaged the weight matrices of attention heads in the final encoder block.

$$
\begin{gathered}
\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} = \boldsymbol{h_i W_Q}, \boldsymbol{h_i W_K}, \boldsymbol{h_i W_V} \\
\boldsymbol{S} = \ Softmax\left(\frac{\boldsymbol{QK^T}}{\sqrt{d_k}}\right) \\
\boldsymbol{h'_{i+1}} = \boldsymbol{SV} \\
\boldsymbol{h_{i+1}} = f(\boldsymbol{h'_{i+1}}) + \boldsymbol{h_i}
\end{gathered}
\tag{1}
$$

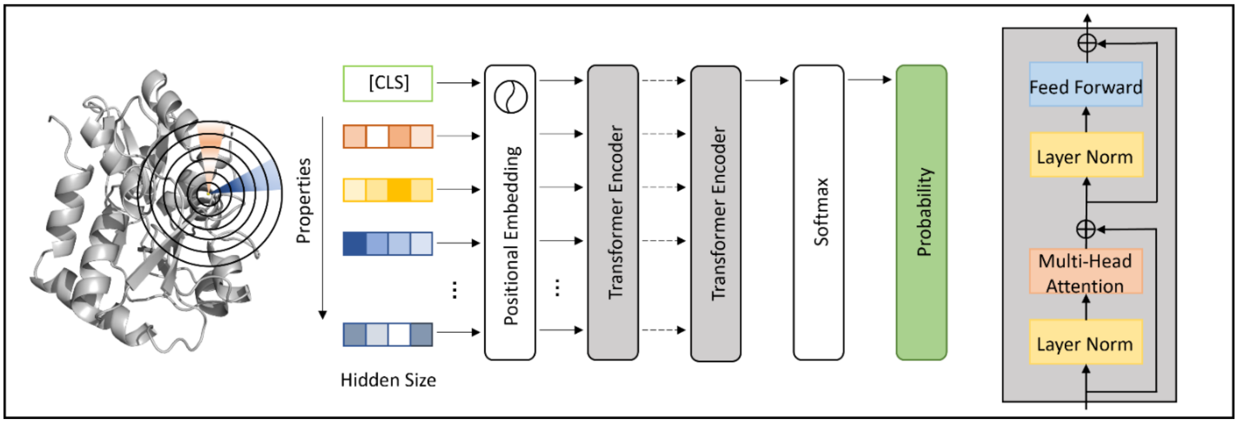

Figure 2. Transformer encoder architecture.

Benchmark Model 1: Feed Forward Neural Network (FNN)

Feed forward neural network (FNN) is a stack of fully connected layers with different number of neurons. Each layer performs a linear combination of input features followed by a nonlinear activation function performed on each feature. At the end of the neural network, a Softmax layer is applied to provide the probabilities of assigning each input data to different categories. In our scenario (Fig S2A), the 6x80 input features corresponding to each grid point are serialized as a 480-dimensional vector $\boldsymbol{h_0}$, which is then passed into FNN.

Benchmark Model 2: ResNet

Convolutional neural networks (CNNs) use convolutional operations to capture localized patterns in images and have exhibited robustness in image classification tasks. For example, pioneering CNN architectures like AlexNet[52] achieved a prominent top-1 accuracy of 63.3% and showed a much lower top-5 error rate by at least 10% compared to the other models in 2012 ImageNet contest. Subsequently, in 2014, ResNet[53] was introduced, revolutionizing CNNs by introducing residual connections. The residual connection enables the identity mapping across different convolutional blocks, significantly enhancing the robustness of larger deep learning models. Furthermore, residual connections alleviated the gradient vanishing/explosion issue that otherwise may occur in standard neural network models.

As shown in Fig S2B, the 6x80 features were reshaped to a 1x6x80 tensor $\boldsymbol{h_0}$ and fed into ResNet blocks. In each ResNet block, the input state $\boldsymbol{h_i} \in \mathbb{R}^{1x6x80}$ was filtered by the 1x2x80 convolutional kernels and converted to the hidden state $\boldsymbol{h_{i+1}}$with the same dimension. This kernel shape can capture spatial correlation among properties in two adjacent concentric shells and allows the information from outer shells to pass to inner shells when more residual blocks are added on.

2.3. Hyperparameters and fine-tuning

We used cross-entropy loss as the objective function of our models. The training hyperparameters are summarized in Table S2. The learning rate was set as 0.01, with a decay factor of 0.5 when the validation loss reached a plateau, and the minimum threshold was 10^{-6}. The optimizing method was AdamW optimizer[54] with betas (0.9, 0.999). To better reflect the true model capacity, no weight decay was applied during model capacity testing and architecture selection procedures. The best model was recorded based on the validation loss to avoid overfitting. We trained each model by 5 random above-mentioned training/validation splits to get validation statistics and select the best model architecture with the most suitable capacity.

The default training epochs was set to 30 in the leave-out model training and 50 in the full model training. The default batch size was 5000, which meant sampling 5 structures from each structure cluster and sampling 1000 class-balanced datapoints from each structure. Benefitted from this dynamic resampling process, the model tended to iterate over different structures of each structure cluster as the training went on. The models were trained on NVIDIA A100 SXM4 80GB.

2.4. Scoring functions and virtual screening

We encoded the grid-point coordinates $\boldsymbol{p}$ together with ligand coordinates $\boldsymbol{q}$ in our scoring function (Eq. 2). $\boldsymbol{p}$ has the shape of $(N, 3)$, where N represents the number of grid points. $\boldsymbol{P}(\boldsymbol{p})$ represents the predicted probabilities of the grid points. $\boldsymbol{q}$ has the shape of $(M, 3)$, where M means the number of heavy atoms in the given compound. $\boldsymbol{T}$ is the transformation matrix that incorporates translational and rotational operations in the three-dimensional Euclidean space. The scoring function rewards compound poses that can occupy space of higher predicted probability. In the following equations, $\boldsymbol{p_i}$ represents the coordinate of the ith grid point, and $\boldsymbol{q_j}$ represents the coordinate of jth ligand heavy atom.

$$\text{Scoring Function}= \frac{\mathbf{1}}{|\boldsymbol{q}|}\boldsymbol{\Sigma}_{\boldsymbol{q_j}}\frac{\mathbf{1}}{|\boldsymbol{N_p}(\boldsymbol{Tq_j})|}\boldsymbol{\Sigma}_{\boldsymbol{p_i}\in \boldsymbol{N_p}(\boldsymbol{Tq_j})}\boldsymbol{P}(\boldsymbol{p_i})\boldsymbol{e}^{-\|\boldsymbol{p_i}-\boldsymbol{Tq_j}\|^2} \quad (2)$$

$$\text{Objective Function}= -\frac{\mathbf{1}}{|\boldsymbol{q}|}\boldsymbol{\Sigma}_{\boldsymbol{q_j}}\frac{\mathbf{1}}{|\boldsymbol{p}|}\boldsymbol{\Sigma}_{\boldsymbol{p_i}}\boldsymbol{P}(\boldsymbol{p_i})\boldsymbol{e}^{-\|\boldsymbol{p_i}-\boldsymbol{Tq_j}\|^2} \quad (3)$$

The binding position of any compound can be optimized by L-BFGS-B to minimize the objective function (Eq. 3). It is worth noting that the objective function considers distance-weighted probabilities of all grid points given a certain probability cutoff, while the scoring function rescales the objective function by only considering the neighboring (<= 1.5 Å) grid points $\boldsymbol{N_p}(\boldsymbol{Tq_j})$. We used $probability\_cutoff = 0.50$ in the objective function to allow a larger compound explorable region and used a different probability cutoff to re-score the optimized poses for different protein systems. During scoring, we suggest using the best probability cutoff evaluated on AUC when the bioassay results are available for the query protein. Please refer to Results section 3.3 for the details of the above-mentioned optimization and scoring method in selecting strong inhibitors from compound libraries for inactive CDK2 and ACE. When the bioassay results are unavailable, it is suggested to choose the probability cutoff based on the pocket properties, which will be further elaborated in the Discussion section.

To predict binding poses using the probability density envelope, we first generated up to 20 conformers for each compound using RDKit (with pairwise RMSDs between these conformations are larger than 1Å after molecular alignment). *useExpTorsionAnglePrefs* and *useBasicKnowledge* were enabled during sampling, and the generated conformers were further optimized by the MMFF[55] force field. Each ligand conformer was initialized by 500 random rotations. An utmost of 10,000 optimization results per compound were finally clustered by DBSCAN[56, 57] method based on their RMSDs to reduce the number of aligned poses. The proposed clusters were sorted by the average score of cluster members. The purposes of massive initial samplings followed by a clustering process are to avoid potential local minima during L-BFGS-B optimization, and to reduce the number of proposed binding poses. We combined sampling and clustering of three different random seeds to make the top-proposed clusters more robust.

2.5. Preparation of query structures and compound libraries

Compound libraries used in virtual screening were fetched from the CHEMBL[8] database. The compounds with IC_{50} less than 10 nM are considered as strong inhibitors and those with IC_{50} values greater than 10 μM are considered as weak inhibitors. To mimic a blind virtual screening, we further removed compounds that share > 95% fingerprint Tanimoto similarity with ligands in the PDBBind cocrystal structures to curate compound libraries different from native ligands used in training. This setting helps validate the model's ability to select novel compounds that are different from existing ligands from a pre-selected drug-like compound library.

An inactive form of CDK2

The cleaned *apo* structure of 1B38 was used as the reference protein structure of CDK2 during virtual screening. The query space (Fig S6A) used in our method is formed by grid points within 5 Å away from native ligands of 11 pre-aligned CDK2 structures.

We filtered compounds with molecular weight greater than 400 Da from CHEMBL301 which consists of CDK2 bioassay results with IC_{50} values. Compounds in CHEMBL301 with bioassay descriptions explicitly reporting to be tested on CDK2/Cyclins are removed because they bind to a different protein conformation caused by cyclin activation. The above filtering rules result in a 147-compound virtual screening library for the CDK2. The molecular property distributions of strong inhibitors and weak inhibitors in this library are similar (Fig S6B).

ACE

The cleaned *apo* structure of 3BKL was used as the reference protein structure of ACE. The query space (Fig S7A) used is formed by grid points within 5 Å away from native ligands of 4 pre-aligned ACE structures that share a similar scaffold (2OC2, 3BKK, 3BKL and 6F9U), but it is large enough to cover all 7 native ACE ligands in the cocrystal structure dataset.

We filtered compounds with molecular weight greater than 400 Da and less than 600 Da from CHEMBL1808 which consists of ACE bioassay results with IC_{50} values. We further used the following criteria to filter out compounds less likely to be drugs and make the compound properties match in the meanwhile (Fig S7B): logP <= 5, number of hydrogen donors <= 5, number of hydrogen acceptors <= 10, number of rotatable bonds <= 5. The above filtering rules result in a 94-compound virtual screening compound library.

2.6. Parameters used in the benchmark docking programs

Protein structures were prepared using AutoDockTools1.5.7 for Smina and AutoDock Vina. The query box of inactive CDK2 has the size of 18 Å x16 Å x16 Å, centered at [1.38, 26.15, 9.40][58] (Fig S4). The query box of ACE has the size of 18 Å x22 Å x24 Å, centered at [43, 45, 44]. These two docking boxes are large enough to cover all ligands from cocrystal structures in the curated dataset. *sdf* files of compounds are generated from SMILES using RDKit. *pdbqt* files used in docking softwares were converted from *sdf* files using *openbabel*[59]. The parameters used in Smina and in AutoDock Vina are *exhaustiveness* = 8, *num_modes* = 20.

The default parameters of DiffDock were used to generate compound conformations and positions during the virtual screening: *inference_steps* = 20, *samples_per_complex* = 40, *batch_size* = 10 *actual_steps* = 18, *no_final_step_noise* = True.

3. Results

3.1. Model training and selection

We chose Transformer encoders in our FeatureDock, as it outperforms other neural network architectures, including FNN and ResNet (see details of these architecture in Methods section 2.2). The model performance was evaluated by cross-entropy loss, area under ROC (AUC), F1 Score and Matthews correlation coefficient (MCC) on the validation set. Transformer encoder consistently achieved the lowest loss across all tested model capacities (Fig 3A and Table S3) compared to FNN and ResNet. Especially, the validation loss saturated at the Transformer model with 20 encoder blocks (500k parameters, loss=0.269, see Table S4), indicating that the model with 500k parameters is large enough to fit the dataset. Additionally, the standard deviations of Transformer models are notably smaller than those of FNN and ResNet.

We further show that the fine-tuning on a specific protein family was not necessary in our framework. This is supported by the evidence that CDK2 leave-out model further optimized on kinase proteins did not show significant performance improvement over the general model (Table S5) when testing on the CDK2 dataset.

Prediction of probability density envelopes for the ligand binding: We first demonstrate the performance of FeatureDock on CDK2 by training a model excluding CDK2 and its 90% homogeneous structures. FeatureDock can generate a probability density envelope by collecting the predicted binding probabilities of all grid points within the query pocket (see Fig S4 for the pocket information). Regions predicted with higher probability suggest a greater likelihood of being occupied by the ligand (Fig 3C), making the probability density envelope a valuable tool for ligand design and ligand binding pose prediction. To assess the precision and robustness of our method's predicted probability density envelope with regards to ligand occupancy, we analyzed 18 pre-aligned structures (Fig S4) of CDK2 (the inactive form) and CDK2/Cyclin (the active form). Fig 3B illustrates the distances between ligand heavy atoms and grid points associated with various predicted binding probability ranges. Notably, grid points with high predicted binding probabilities (> 0.8) are very close to the actual locations of ligand atoms, with an average distance of 0.62 Å. This finding indicates that our model's predicted high binding probability regions closely correspond to the actual locations of the ligand heavy atoms. Thus, our model shows great promise for predicting ligand poses by aligning them with the predicted probability density envelope.

We also applied FeatureDock to predict binding probability density envelopes on different conformational states of the same protein. As shown in Fig S5, the probability density envelopes of the inactive and active CDK2 exhibit substantial differences in the region proximal to Tyr15, located at the glycine-rich loop (G-loop, 11-16). Consequently, an allosteric ligand binding near Tyr15 may hinder the structural transition from the inactive to the active state, thereby inhibiting the CDK2 activation. This prediction is consistent with the reported ANS allosteric binding site[60] located between the ATP binding pocket and C-helix which can inhibit cyclin binding.

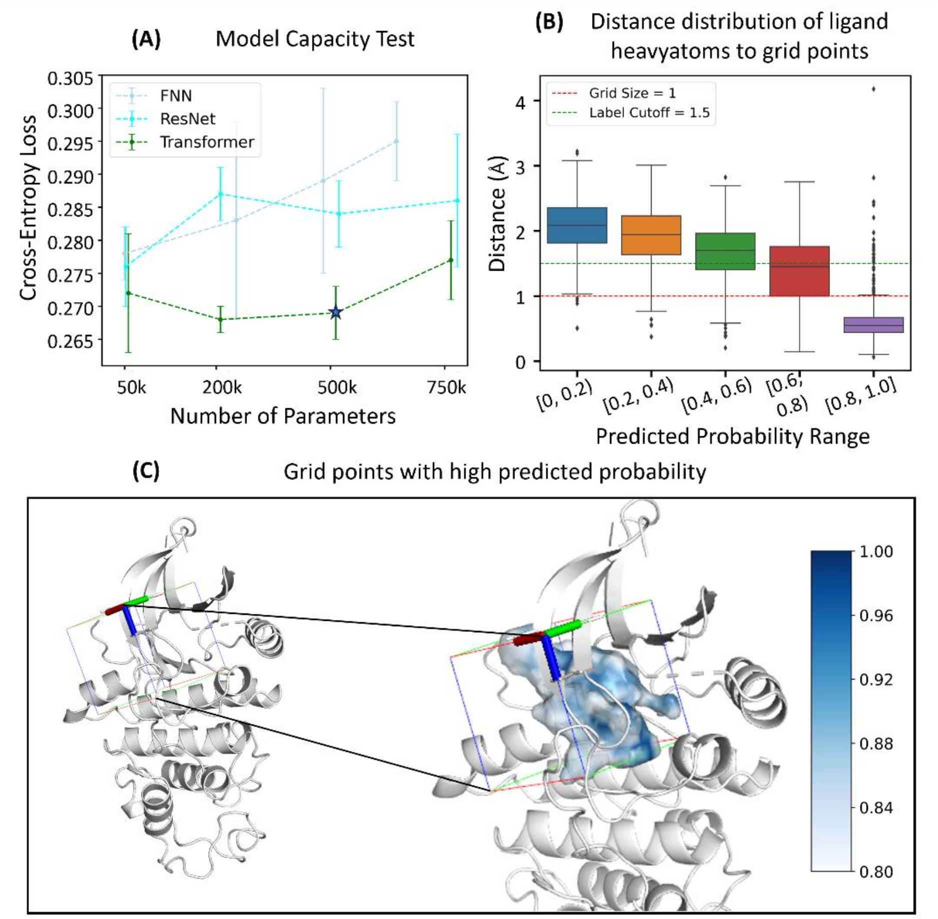

Figure 3. **Performance and explainability of CDK2 predictions**. (A) Comparisons of cross-entropy loss of the validation set of three different architectures: FNN, ResNet, and Transformer. (B) Distance distributions of true ligand atoms to grid points. (C) Query box and the probability density envelope of 1B38 (inactive CDK2). The darker blue regions have higher probabilities.

3.2. Identifications of chemical features contributing most to ligand binding

Our FeatureDock tool generates attention maps for each input chemical feature using the attention weights from the Transformer architecture (see Fig S3 and Eq. 1). Each attention map records the contributions of a specific chemical feature in predicting the binding probabilities of all grid points in the binding pocket. For example, Fig 4B visualizes the attention maps of two representative chemical features (hydrophobicity and polar residues) of grid points in the binding pocket for an inactive form of CDK2 structure (PDB ID: 1B38[61]). In Fig 4A, we report the average attention weights per given amino acid surrounding the binding pocket. Among all the hydrophobic residues, Phe82 and Phe80 contribute the most to the ligand binding, with contributions of 2.6% and 2.0%, respectively (top panel of Fig 4A). The attention weights from our FeatureDock tool is consistent with the previous results obtained from the pharmacophore model[62]. In this pharmacophore model, hydrophobic pharmacophores containing Phe82 and Phe80 emerged as frequently occurring features based on multicomplex-based analysis constructed on 124 CDK2 cocrystal structures. For polar amino acids surrounding the binding pocket, Gln85 and Gln131-Asn132 are predicted by our

FeatureDock to mostly contribute to the ligand binding (bottom panel of Fig 4A). Among these residues, Gln85 has also been predicted by the previous pharmacophore modeling study[62] as a favorable interacting residue for pharmacophore binding. Moreover, previous Molecular Dynamics (MD) simulation studies[63, 64] also pinpointed Gln85 as a key residue in designing ligands selectively binding to CDK2. Specifically, replacing the chemical groups on the ligand near Gln85-Lys89 with more electronegative ones enhances ligand selectivity for CDK2 over CDK7. This previous finding supports our observation that Gln85 contributes as large as 3.9% to the ligand binding. In addition, while Gln131 and Asn132 currently show low statistical occupancy as hydrogen bond acceptors or donors in the pharmacophore model[62], we anticipate that designing ligands to interact with these two residues via favorable hydrogen bonding interactions could improve the CDK2 ligand selectivity, as both of these residues show high contributions in our attention analysis (bottom panel of Fig 4A).

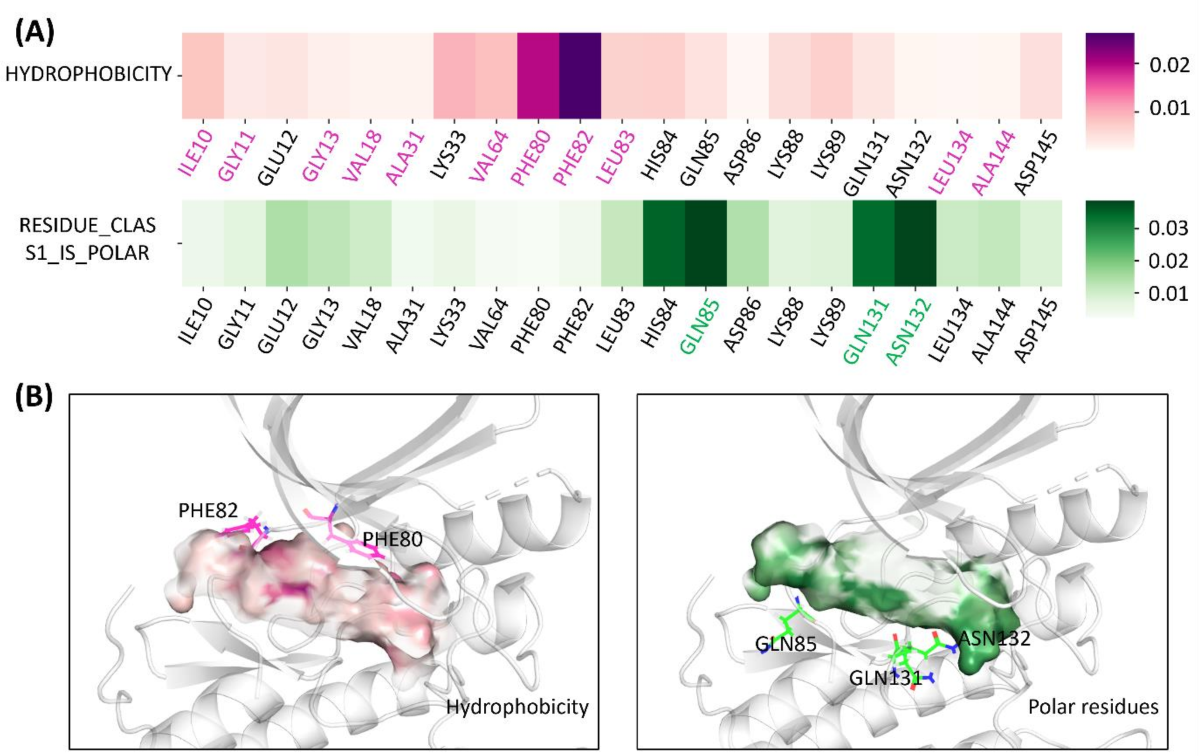

Figure 4. Model interpretability visualized by attention weight of different input properties. (A) Heatmap of property contribution to the amino acids of interest by averaging attention weights of grid points around each amino acid. The picked residues frequently form contact with ligands in cocrystal structures. The text of hydrophobic residues and polar residues are colored in magenta and green, respectively. The range of attention weights is [0, 1]. (B) The attention weight map of different properties of the spatial grid points. Regions in darker color have higher contribution values in the attention analysis.

3.3. FeatureDock outperforms DiffDock, Smina and AutoDock Vina in differentiating strong and weak inhibitors

As discussed in the Introduction section, distinguishing between strong and weak inhibitors remains challenging for docking algorithms during virtual screening, as all inhibitors tend to fit reasonably well in the binding pockets. In this section, we demonstrate that our FeatureDock tool outperforms traditional docking methods (Smina[22] and AutoDock Vina[21]), as well as the recently developed diffusion-based generative model (DiffDock[28]) in this task for an inactivated form of CDK2 and the ACE receptor. In FeatureDock, we defined a scoring function based on Eq. 2, which assesses the agreement between the locations of ligand atoms and the predicted probability density envelopes. To optimize the ligand's position and orientation in the binding pocket, we implemented an optimization procedure using the L-BFGS-B algorithm to minimize the objective function defined in Eq. 3. For additional details on our optimization protocol and the preparation of query structures and compound libraries for virtual screening, please refer to Methods sections 2.4 and 2.5. To evaluate the performance of virtual screening, we used two metrics: KL divergence and AUC. A higher KL divergence indicates that the model can more effectively distinguish strong and weak inhibitors. Similarly, a higher AUC demonstrates the scoring function’s effectiveness in ranking strong inhibitors above weak ones. The parameters used in benchmark docking methods are detailed in Methods section 2.6.

Inactivated form of CDK2: We compiled a library containing 147 compounds followed by the filtering rules described in the Methods section 2.5. As shown in Fig 5A, FeatureDock achieved the largest KL divergence (0.67) in differentiating strong inhibitors from weak inhibitors, outperforming DiffDock (0.39), Smina (0.04), and Vina (0.04). The peaks of score distributions of FeatureDock showed better separation compared to DiffDock in Fig 5A. These results suggest that we can use the FeatureDock scoring function to select reliable candidates which achieve higher scores in the virtual screening. Moreover, the peak of strong inhibitors is mixed with that of weak inhibitors in both Smina and AutoDock Vina, which explains the high false positive ratios caused by the scoring functions used in these two docking methods. Furthermore, as shown in Fig 5B, the AUC showed that FeatureDock (0.74) and DiffDock (0.76) can rank strong inhibitors higher than the weak ones, while the binding affinity scores used by Smina (0.43) and Vina (0.43) ranked strong and weak inhibitors in this library worse than random guess. These two-evaluation metrics both support the effectiveness of FeatureDock’s scoring function (defined in Eq. 2) during virtual screening.

In FeatureDock, we simultaneously score compounds and optimize their binding poses using the L-BFGS-B optimization algorithm. In Fig 5C, we show the optimal ligand poses obtained from FeatureDock overlaid with their predicted probability density envelopes for one strong inhibitor (CHEMBL402158) and two weak inhibitors (CHEMBL3421971 and CHEMBL3798066) of CDK2. In the FeatureDock pose of the strong inhibitor (CHEMBL402158), its 1H-indazole group occupies the highest probability regions ($p >= 0.95$) within the hydrophobic pocket near Phe80 and Phe82 of CDK2. The remaining portion of this strong inhibitor extends into adjacent regions with similarly high probability values ($0.90 <= p < 0.95$). As a result, CHEMBL402158 exhibits a high score in FeatureDock, aligning with its experimentally measured strong IC_{50} of 7nM (Left panel of Fig 5C). For the weak inhibitor CHEMBL3421971 (Middle panel of Fig 5C), its benzene and piperazine groups also occupy the highest probability region ($p>=0.95$) near the hydrophobic residues Phe80 and Phe82. However, the linker atom connecting these function groups to the rest of the compound bypasses a region with substantially lower probabilities (as indicated by the red-labelled atom in the Middle panel of Fig 5C). As a result, FeatureDock assigns a lower score to

CHEMBL3421971, consistent with its experimentally measured IC_{50} of approximately 13.7μM. As for the other weak inhibitor, CHEMBL3798066, a significant portion also occupies a relatively low probability region (atoms labelled in red in the Right panel of Fig 5C), resulting in a weak inhibitor (IC_{50} = 11μM). These examples indicate that FeatureDock provides not only an accurate scoring function, but also meaningful optimal ligand poses, making it promising for virtual screening.

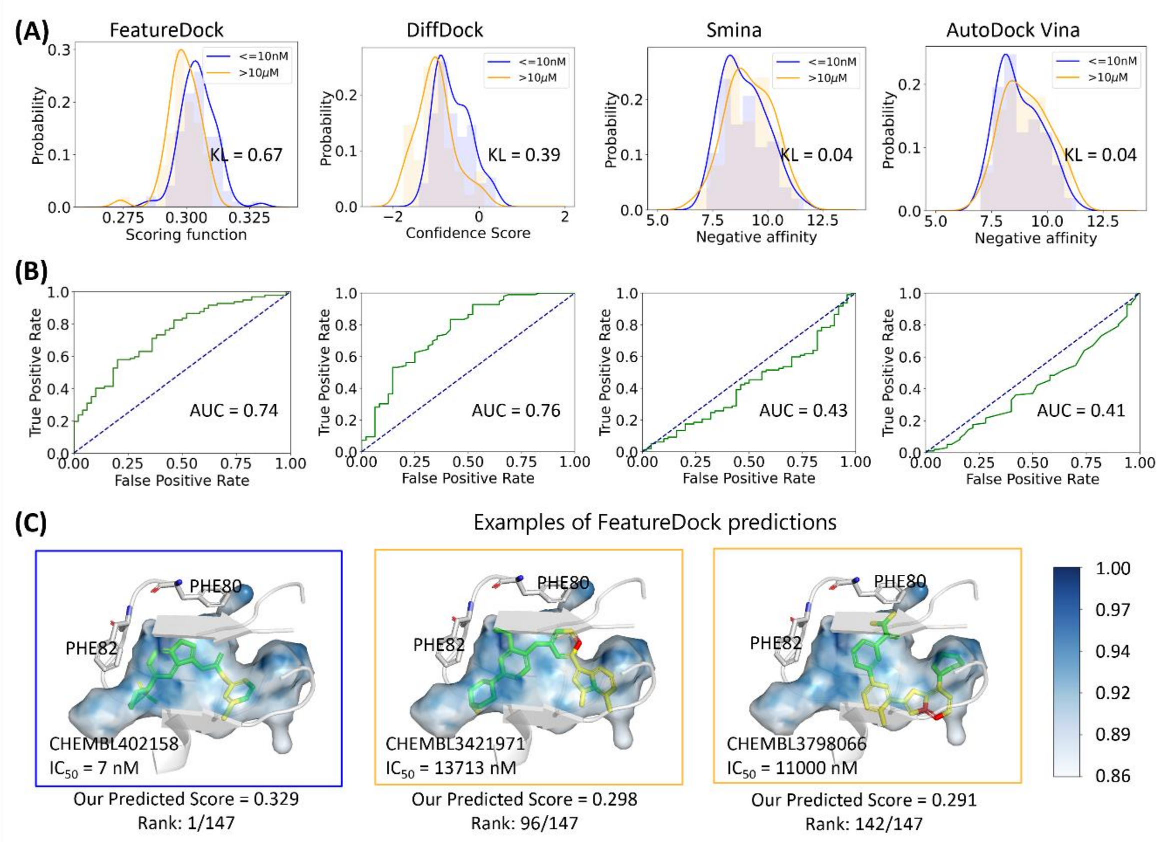

Figure 5. Probability density envelope guided virtual screening of an inactivated form of CDK2. (A) KL divergence of score distribution on strong inhibitors and weak inhibitors from a filtered CDK2 bioassay dataset which contains 147 compounds. (B) AUC of scoring functions on the 147-compound library. (C) True positive (the blue boxed) and true negative (the orange boxed) examples of predicted poses overlaid with the predicted probability density envelope. The colorbar represents colors of different probability values. Compound atoms in the regions of $p>=0.95$, $0.90<=p<0.95$ and $0.80<=p<0.90$ are colored in green, yellow, and red respectively. Surrounding protein structures are shown in white ribbon, with important binding residues Phe80 and Phe82 highlighted in white sticks.

To further assess the accuracy of pose prediction from FeatureDock, we validated its predicted poses by comparing to the co-crystal structures for 11 native ligands of CDK2 (the inactivated form) that have cocrystal structures in PDBBind. The posing performance is evaluated by root mean square distance (RMSD) between the native pose and the predicted pose. An RMSD below 2 Å is usually considered to be a good prediction in docking methods. After pose optimization described in Methods section 2.4, the RMSD converged after picking 4 top-scored poses for each ligand in FeatureDock. The average and median RMSD achieved by our method are 2.4 Å and 2.1 Å, which is significantly better than randomly choosing poses centered to the predicted regions while not optimized via soft alignment based on the probability density envelope. Examples of predicted poses in Fig 6 right panel show the alignment performance.

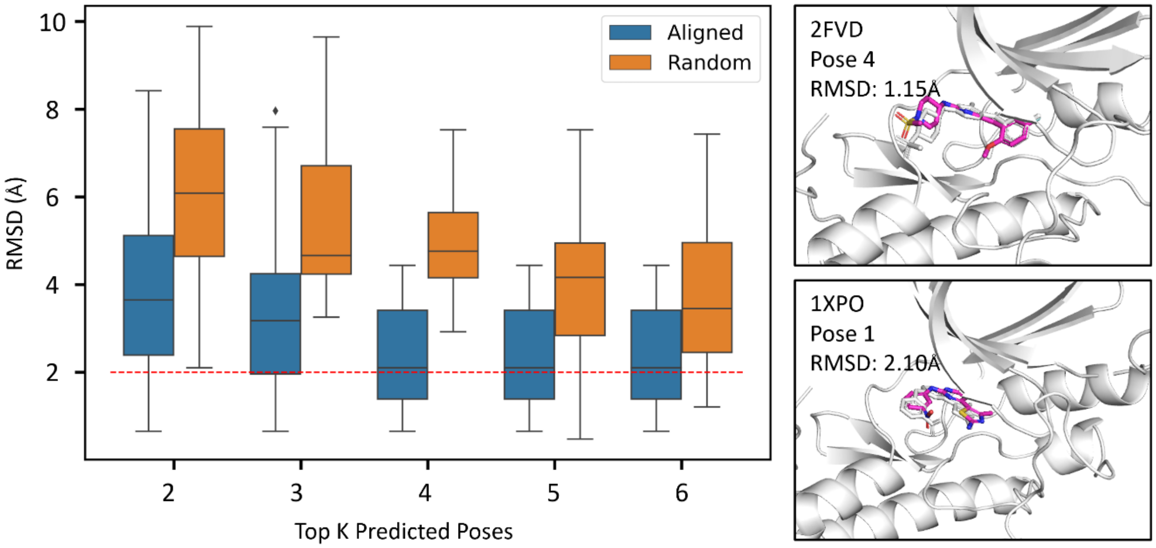

Figure 6. Pose prediction from FeatureDock on CDK2. The query space of each protein refined by protein residues within [1, 6] Å away from the heavy atoms of its native ligand. The native rotamer was used in the posing process. Left panel: Distribution of root mean square distances (RMSDs) between predicted poses and native poses. Right panel: Examples of predicted poses. The ground truths and the predictions are presented in white sticks and magenta sticks, respectively.

ACE receptor: We followed the filtering rules described in Methods section 2.5 and compiled a compound library containing 94 compounds. On ACE, FeatureDock also achieves the highest KL divergence (0.30) (Fig 7A) and AUC (0.69) (Fig 7B), outperforming DiffDock, Smina and AutoDock Vina. A relatively higher probability cutoff of 0.90 was used to re-score the compound occupancies for ACE based on the result of hyperparameter scanning (Fig S9). For the details of how to select the proper probability cutoff as a scoring hyperparameter for different protein systems, please refer to the Methods section 2.4. We visualized one true positive prediction and two true negative prediction to analyze the performance of scoring and pose prediction. As depicted in Fig 7C, the true positive prediction CHEMBL100413 occupied the regions of p>=0.95 close to Tyr520, which is one of the lisinopril-binding residues[65, 66]. Furthermore, the predicted pose of CHEMBL100413 exhibits the correct orientation to form the hydrogen bond with Tyr520, which explains its high potent to inhibit ACE. The true negative prediction CHEMBL5220968 occupied the same region of p>=0.95 except that the compound sticks out of the probability envelope due to the compound geometric constraint, leading to a lower score. Another true negative prediction CHEMBL5177969 achieved an even lower score because it totally missed the regions of highest probabilities.

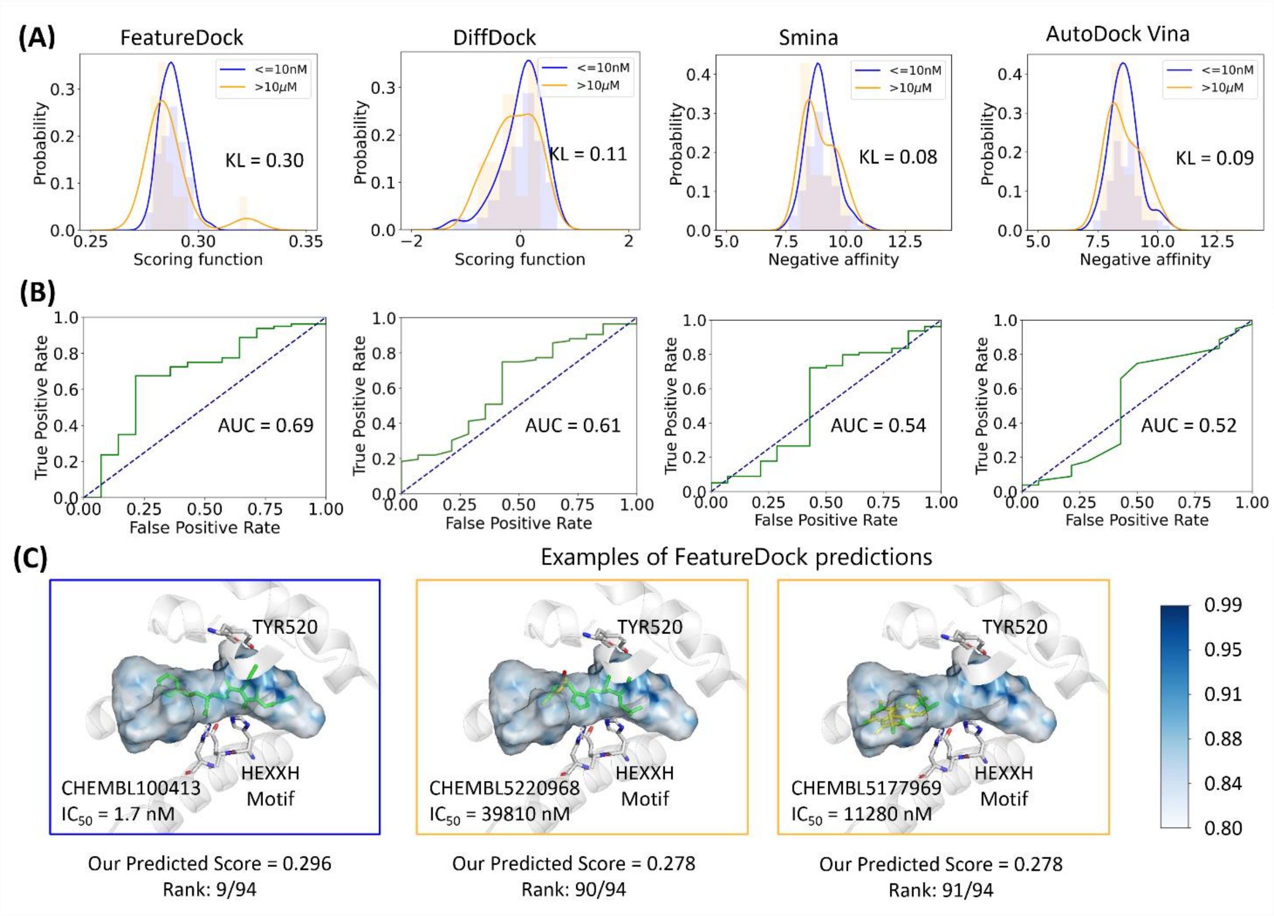

Figure 7. Probability density envelope guided virtual screening of ACE. (A) KL divergence of score distribution on strong inhibitors and weak inhibitors from a filtered ACE bioassay dataset which contains 94 compounds. (B) AUC of scoring functions on the 94-compound library. (C) True positive (the blue boxed) and true negative (the yellow boxed) examples of predicted poses overlaid with the predicted probability density envelope. Compound atoms in the regions of $p>=0.95$, $0.90<=p<0.95$ and $p<=0.80$ are colored in green, yellow, and red respectively. The colorbar represents colors of different probability values. Surrounding protein structures are shown in white ribbon. The active site (NEXXH motif) and the Tyr520 residue are shown as white sticks.

4. Discussion

FeatureDock predicts the binding poses and evaluates the binding potency of compounds based on their alignment with the probability density envelopes. Therefore, it provides more information about the ligand poses and orientations than other deep learning methods, including K_{DEEP}[39] and PotentialNet[40] that only predict protein-ligand binding affinity values and rely on 3-dimensional protein-ligand complexes as the input.

In FeatureDock, we utilize the L-BFGS-B optimization algorithm to find the ligand pose that best fits to the predicted probability density envelope, and we acknowledge that this optimization process could be trapped into local energy minima. To address this issue, we perform independent optimization runs initiated from different compound rotamers and using different random seeds.

This strategy has successfully identified the correct binding pose for CDK2 (see Fig 6). In the future, to further improve the efficiency of pose sampling and optimization, one could incorporate E3-equivariant diffusion models to generate binding positions based on the probability density envelopes.

When applying FeatureDock to virtual screening, it is important to select a proper probability cutoff in the scoring function. We recommend choosing its value by scanning the probability cutoff as a hyperparameter on the compound bioassay data and pick the best cutoff that achieves the highest AUC. Intuitively, for more rigid pockets that do not allow multiple binding modes, a higher probability cutoff can help refine the search space, focusing only on regions containing grid points with confident ligand binding predictions. For example, a higher cutoff value was chosen for ACE than CDK2 in the Results section 3.3, which is consistent with the observation that the ACE ligand binding pocket has lower flexibility compared to that of CDK2, as evaluated by higher B-factor values for ACE (Fig S11). When the bioassay data is not available, we would suggest using a balanced probability cutoff (e.g. $cutoff = 0.90$) to guarantee that high probability regions are prioritized to be occupied, while remaining a probability density envelope that can represent and cover the binding pocket adequately. This cutoff has shown effective in the virtual screening of both CDK2 and ACE (Fig S8 and S9).

One challenge facing deep learning-based docking methods is their tendency to generate "invalid" ligand poses[67]. This occurs because these methods often produce poses with unphysical chemical bonds or steric clashes between the protein and ligand, as they do not explicitly include physical interactions during training and pose prediction. To address this challenge, we introduced a post-analysis step to identify and remove unphysical poses generated by FeatureDock and DiffDock (Fig S14). In this step, we utilized AutoDock Vina for local optimization to identify steric clashes and other unphysical interactions that result in unfavorable affinity scores. Specifically, we rejected poses that with affinity scores larger than 0. It is important to note that we used docking affinity only to identify steric clashes, not to re-pose the conformations generated by FeatureDock or DiffDock. In the future, one could directly introduce a physical interaction term to avoid steric clashes in the objective function of the pose optimization process in the FeatureDock.

5. Conclusions

In this study, we introduced FeatureDock, a feature-based Transformer framework for predicting and scoring protein-ligand binding poses. FeatureDock extracts the physicochemical features of the protein's local environment and utilizes a transformer encoder to predict probability density envelopes for given protein pockets. These predicted probability density envelopes help determine the binding preference of the ligand at grid points within the protein pockets. Compared to FNN and ResNet architectures, Transformer encoders demonstrate superior performance, achieving higher and more stable prediction accuracy throughout comprehensive model capacity tests. In addition, the attention mechanism in the Transformer encoder further shows potential for assisting rational drug design by elucidating the contribution of each chemical feature to the final predicted ligand binding probabilities. Using the predicted probability density envelopes, we designed a custom scoring function and a ligand-position optimization algorithm to score and predict the protein-ligand binding poses. In virtual screening of drug-like compounds, FeatureDock outperforms DiffDock, Smina and AutoDock-Vina in distinguishing strong inhibitors from weak ones for both the inactivated form of CDK2 and ACE systems. FeatureDock also effectively predicts their binding poses accurately. We expect that FeatureDock holds potential to be widely applied in virtual screening, aiding in drug design.

6. Acknowledgements

XH acknowledges the support from the Hirschfelder Professorship Fund, and the support from the Research Forward Fund from the University of Wisconsin-Madison Office of the Vice Chancellor for Research with funding from the Wisconsin Alumni Research Foundation.

7. Data Availability

The dataset used in model training can be derived from PDBBind v2020 refined dataset. The source code for data preprocessing, model training, evaluation and postanalysis can be found on the github link: https://github.com/xuhuihuang/featuredock. The parameters of full models and virtual screening libraries used in the results of this manuscript can be found at https://github.com/xuhuihuang/featuredock.

8. References

(1) Wouters, O.; McKee, M.; Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *Jama-Journal of the American Medical Association* **2020**, *323* (9), 844-853, Article. DOI: 10.1001/jama.2020.1166.
(2) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of early drug discovery. *British Journal of Pharmacology* **2011**, *162* (6), 1239-1249, Review. DOI: 10.1111/j.1476-5381.2010.01127.x.
(3) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; et al. Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* **2011**, *10* (3), 188-195. DOI: 10.1038/nrd3368.
(4) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; et al. PubChem Substance and Compound databases. *Nucleic Acids Research* **2016**, *44* (D1), D1202-D1213. DOI: 10.1093/nar/gkv951.
(5) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.; Thiessen, P.; Yu, B.; et al. PubChem 2023 update. *Nucleic Acids Research* **2022**, Article|Early Access. DOI: 10.1093/nar/gkac956.
(6) Irwin, J.; Shoichet, B. ZINC - A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* **2005**, *45* (1), 177-182, Article. DOI: 10.1021/ci049714+.
(7) Irwin, J.; Sterling, T.; Mysinger, M.; Bolstad, E.; Coleman, R. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling* **2012**, *52* (7), 1757-1768, Article. DOI: 10.1021/ci3001277.
(8) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* **2019**, *47* (D1), D930-D940. DOI: 10.1093/nar/gky1075.
(9) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics* **2002**, *47* (4), 409-443. DOI: 10.1002/prot.10115.
(10) Warren, G.; Andrews, C.; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M.; Lindvall, M.; Nevins, N.; Semus, S.; Senger, S.; et al. A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry* **2006**, *49* (20), 5912-5931, Article. DOI: 10.1021/jm050362n.
(11) Kitchen, D.; Decornez, H.; Furr, J.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery* **2004**, *3* (11), 935-949, Review. DOI: 10.1038/nrd1549.
(12) Liu, J.; Wang, R. Classification of Current Scoring Functions. *Journal of Chemical Information and Modeling* **2015**, *55* (3), 475-482, Article. DOI: 10.1021/ci500731a.
(13) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **1982**, *161* (2), 269-288. DOI: 10.1016/0022-2836(82)90153-x.
(14) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **2001**, *15* (5), 411-428. DOI: 10.1023/a:1011115820450.
(15) Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C. DOCK 6: Impact of new features and current docking performance. *J Comput Chem* **2015**, *36* (15), 1132-1156. DOI: 10.1002/jcc.23905.
(16) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J Mol Recognit* **1996**, *9* (1), 1-5. DOI: 10.1002/(sici)1099-1352(199601)9:1<1::aid-jmr241>3.0.co;2-6.
(17) Raha, K.; Merz, K. M. A quantum mechanics-based scoring function: study of zinc ion-mediated ligand binding. *J Am Chem Soc* **2004**, *126* (4), 1020-1021. DOI: 10.1021/ja038496i.

(18) Khoruzhii, O.; Donchev, A. G.; Galkin, N.; Illarionov, A.; Olevanov, M.; Ozrin, V.; Queen, C.; Tarasov, V. Application of a polarizable force field to calculations of relative protein-ligand binding affinities. *Proc Natl Acad Sci U S A* **2008**, *105* (30), 10378-10383. DOI: 10.1073/pnas.0803847105.
(19) Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front Pharmacol* **2018**, *9*, 1089. DOI: 10.3389/fphar.2018.01089.
(20) Pason, L. P.; Sotriffer, C. A. Empirical Scoring Functions for Affinity Prediction of Protein-ligand Complexes. *Mol Inform* **2016**, *35* (11-12), 541-548. DOI: 10.1002/minf.201600048.
(21) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **2010**, *31* (2), 455-461. DOI: 10.1002/jcc.21334.
(22) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model* **2013**, *53* (8), 1893-1904. DOI: 10.1021/ci300604z.
(23) Koebel, M. R.; Schmadeke, G.; Posner, R. G.; Sirimulla, S. AutoDock VinaXB: implementation of XBSF, new empirical halogen bond scoring function, into AutoDock Vina. *J Cheminform* **2016**, *8*, 27. DOI: 10.1186/s13321-016-0139-1.
(24) Nivedha, A. K.; Thieker, D. F.; Makeneni, S.; Hu, H.; Woods, R. J. Vina-Carb: Improving Glycosidic Angles during Carbohydrate Docking. *J Chem Theory Comput* **2016**, *12* (2), 892-901. DOI: 10.1021/acs.jctc.5b00834.
(25) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model* **2021**, *61* (8), 3891-3898. DOI: 10.1021/acs.jcim.1c00203.
(26) Quiroga, R.; Villarreal, M. A. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PLoS One* **2016**, *11* (5), e0155183. DOI: 10.1371/journal.pone.0155183.
(27) McNutt, A. T.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; Koes, D. R. GNINA 1.0: molecular docking with deep learning. *J Cheminform* **2021**, *13* (1), 43. DOI: 10.1186/s13321-021-00522-2.
(28) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776* **2022**. DOI: 10.48550/arXiv:2210.01776.
(29) Lu, W.; Wu, Q.; Zhang, J.; Rao, J.; Li, C.; Zheng, S. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems* **2022**, *35*, 7236-7249. DOI: 10.1101/2022.06.06.495043.
(30) Stärk, H.; Ganea, O.; Pattanaik, L.; Barzilay, R.; Jaakkola, T. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, 2022, 2022; PMLR: pp 20503-20521. DOI: 10.48550/arXiv:2202.05146.
(31) Plewczynski, D.; Łaźniewski, M.; Augustyniak, R.; Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J Comput Chem* **2011**, *32* (4), 742-755. DOI: 10.1002/jcc.21643.
(32) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling* **2019**, *59* (2), 895-913, Article. DOI: 10.1021/acs.jcim.8b00545.
(33) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **2003**, *52* (4), 609-623. DOI: 10.1002/prot.10465.
(34) Vilar, S.; Cozza, G.; Moro, S. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr Top Med Chem* **2008**, *8* (18), 1555-1572. DOI: 10.2174/156802608786786624.
(35) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; 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), 1739-1749. DOI: 10.1021/jm0306430.

(36) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem* **2004**, *47* (7), 1750-1759. DOI: 10.1021/jm030644s.

(37) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip Rev Comput Mol Sci* **2015**, *5* (6), 405-424. DOI: 10.1002/wcms.1225.

(38) Gomes, J.; Ramsundar, B.; Feinberg, E. N.; Pande, V. S. Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv preprint arXiv:1703.10603* **2017**. DOI: 10.48550/arXiv:1703.10603.

(39) Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling* **2018**, *58* (2), 287-296. DOI: 10.1021/acs.jcim.7b00650.

(40) Feinberg, E.; Sur, D.; Wu, Z.; Husic, B.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. PotentialNet for Molecular Property Prediction. *Acs Central Science* **2018**, *4* (11), 1520-1530, Article. DOI: 10.1021/acscentsci.8b00507.

(41) Halperin, I.; Glazer, D. S.; Wu, S.; Altman, R. B. The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. *BMC Genomics* **2008**, *9 Suppl 2* (Suppl 2), S2. DOI: 10.1186/1471-2164-9-S2-S2.

(42) Lam, J.; Li, Y.; Zhu, L.; Umarov, R.; Jiang, H.; Heliou, A.; Sheong, F.; Liu, T.; Long, Y.; Li, Y.; et al. A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nature Communications* **2019**, *10*, Article. DOI: 10.1038/s41467-019-12920-0.

(43) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv preprint arXiv: 1706.03762* **2017**. DOI: 10.48550/ arXiv: 1706.03762.

(44) Wang, R.; Fang, X.; Lu, Y.; Yang, C.; Wang, S. The PDBbind database: Methodologies and updates. *Journal of Medicinal Chemistry* **2005**, *48* (12), 4111-4119, Article. DOI: 10.1021/jm048957q.

(45) 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 Res* **2000**, *28* (1), 235-242. DOI: 10.1093/nar/28.1.235 From NLM.

(46) Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **2017**, *35* (11), 1026-1028. DOI: 10.1038/nbt.3988.

(47) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation* **1997**, *9* (8), 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.

(48) Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**. DOI: 10.48550/arXiv:2010.11929.

(49) Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10-17 Oct. 2021, 2021; pp 9992-10002. DOI: 10.1109/ICCV48922.2021.00986.

(50) Huang, K.; Xiao, C.; Glass, L. M.; Sun, J. MolTrans: Molecular Interaction Transformer for drug-target interaction prediction. *Bioinformatics* **2021**, *37* (6), 830-836. DOI: 10.1093/bioinformatics/btaa880.

(51) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics* **2018**. DOI: 10.48550/ arXiv:1810.04805.

(52) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **2017**, *60* (6), 84-90. DOI: 10.1145/3065386.

(53) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **2015**. DOI: 10.48550/ARXIV.1512.03385.
(54) Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *International Conference on Learning Representations* **2017**. DOI: 10.48550/arXiv:1711.05101.
(55) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *Journal of cheminformatics* **2014**, *6*, 1-4. DOI: 10.1186/s13321-014-0037-3.
(56) Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; Xu, X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* **2017**, *42* (3), 1-21. DOI: 10.1145/3068335.
(57) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. 1996, Vol. 96, pp 226-231. DOI: 10.5555/3001460.3001507.
(58) Vieira, T.; Sousa, S. Comparing AutoDock and Vina in Ligand/Decoy Discrimination for Virtual Screening. *Applied Sciences-Basel* **2019**, *9* (21), Article. DOI: 10.3390/app9214538.
(59) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J Cheminform* **2011**, *3*, 33. DOI: 10.1186/1758-2946-3-33.
(60) Betzi, S.; Alam, R.; Martin, M.; Lubbers, D. J.; Han, H.; Jakkaraj, S. R.; Georg, G. I.; Schönbrunn, E. Discovery of a potential allosteric ligand binding site in CDK2. *ACS Chem Biol* **2011**, *6* (5), 492-501. DOI: 10.1021/cb100410m.
(61) Brown, N. R.; Noble, M. E.; Lawrie, A. M.; Morris, M. C.; Tunnah, P.; Divita, G.; Johnson, L. N.; Endicott, J. A. Effects of phosphorylation of threonine 160 on cyclin-dependent kinase 2 structure and activity. *J Biol Chem* **1999**, *274* (13), 8746-8756. DOI: 10.1074/jbc.274.13.8746.
(62) Zou, J.; Xie, H. Z.; Yang, S. Y.; Chen, J. J.; Ren, J. X.; Wei, Y. Q. Towards more accurate pharmacophore modeling: Multicomplex-based comprehensive pharmacophore map and most-frequent-feature pharmacophore model of CDK2. *J Mol Graph Model* **2008**, *27* (4), 430-438. DOI: 10.1016/j.jmgm.2008.07.004.
(63) Chohan, T. A.; Chen, J. J.; Qian, H. Y.; Pan, Y. L.; Chen, J. Z. Molecular modeling studies to characterize N-phenylpyrimidin-2-amine selectivity for CDK2 and CDK4 through 3D-QSAR and molecular dynamics simulations. *Mol Biosyst* **2016**, *12* (4), 1250-1268. DOI: 10.1039/c5mb00860c.
(64) Chohan, T. A.; Qian, H. Y.; Pan, Y. L.; Chen, J. Z. Molecular simulation studies on the binding selectivity of 2-anilino-4-(thiazol-5-yl)-pyrimidines in complexes with CDK2 and CDK7. *Mol Biosyst* **2016**, *12* (1), 145-161. DOI: 10.1039/c5mb00630a.
(65) Natesh, R.; Schwager, S. L. U.; Sturrock, E. D.; Acharya, K. R. Crystal structure of the human angiotensin-converting enzyme–lisinopril complex. *Nature* **2003**, *421* (6922), 551-554. DOI: 10.1038/nature01370.
(66) Izzo Jr, J. L.; Weir, M. R. Angiotensin-converting enzyme inhibitors. *The Journal of Clinical Hypertension* **2011**, *13* (9), 667. DOI: 10.1111/j.1751-7176.2011.00508.x.
(67) Buttenschoen, M.; Morris, G. M.; Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem Sci* **2024**, *15* (9), 3130-3139. DOI: 10.1039/d3sc04185a.