Can Deep Learning Blind Docking Methods be used to Predict Allosteric Compounds?

Eric A. Chen[†] and Yingkai Zhang[*,†,‡,¶]

†*Department of Chemistry, New York University, New York, New York 10003, United States*

‡*Simons Center for Computational Physical Chemistry at New York University, New York, New York 10003, United States*

¶*NYU-ECNU Center for Computational Chemistry, Shanghai, Shanghai 200062, China*

E-mail: yingkai.zhang@nyu.edu

Abstract

Allosteric compounds offer an alternative mode of inhibition to orthosteric compounds with opportunities for selectivity and non-competition. Structure-based drug design (SBDD) of allosteric compounds introduce complications compared to their orthosteric counter parts; multiple binding sites of interest are considered and often allosteric binding is only observed in particular protein conformations. Blind docking methods show potential in virtual screening allosteric ligands. Deep learning blind docking methods such as DiffDock achieve state-of-the-art performance on protein–ligand complex prediction benchmarks compared to traditional docking methods such as Vina and Lin_F9. Recently, the first allosteric inhibitors in the Type III binding mode were discovered for Cyclin-Dependent Kinase 2 (CDK2), making it a timely case study for these methods. To this aim, we introduce and explore the utility of a data-driven platform called the Minimum Distance Matrix Representation (MDMR). In contrast to other protein representations, using the minimum residue–residue (or residue–ligand) distance as a feature prioritizes the formation of interactions. Dimension reduction and clustering methods take these inputs and highlight the variety of protein conformations and ligand binding

modes. Here, we identify an intermediate protein conformation that other heuristic-based kinase conformation classification methods do not distinguish. Next, the self- and cross-docking benchmarks assess whether docking methods can predict both orthosteric and allosteric binding modes and if prospective success is conditional on the selection of the protein receptor conformation, respectively. We find that a combined method, DiffDock followed by Lin_F9 Local Re-Docking (DiffDock+LRD) can predict both orthosteric and allosteric binding modes and the intermediate conformation must be selected to predict the allosteric pose. In sum, this work highlights the value of a data-driven method to explore protein conformations and ligand binding modes and outlines the challenges of SBDD of allosteric compounds.

Introduction

Allostery is a perturbation, such as ligand binding, distal to the active site that alters the conformational ensemble to modulate the protein function.[1,2] This allosteric phenomenon has motivated the development of therapeutics that are non-competitive with the native ligand, avoid orthosteric resistance, improve selectivity and limit off-target toxicity.[3] Allosteric inhibitors shift the active/inactive enzymatic equilibrium to the inactive state, thus disabling protein function. This is in contrast to orthosteric inhibitors, whose mechanism of inhibition is to out-compete the native ligand. Recently, there has been considerable interest and promise in the development of allosteric kinase inhibitors.[4]

Deep-learning protein–ligand binding pose prediction methods enable the discovery of allosteric binders for structure-based drug design (SBDD). In this work, we focus on blind docking methods, where the ligand pose search space is unrestricted. Traditional docking methods like Vina and Lin_F9 rely on crafted scoring functions that rates binding poses and optimization algorithms that search for the global maximum of the scoring function.[5–7] Deep learning blind docking methods are principally different in that they rely on trained parameters to build the scoring function. We focus on diffusion-based generative methods, which are likelihood-based models that learn a data distribution by learning the reverse process of the forward diffusion process.[8,9] These

methods vary by how they treat the receptor. Rigid-receptor docking methods such as DiffDock (2023)/-S/-L treat the protein statically.[10,11] DynamicBind, a flexible pocket docking jointly predicts protein side-chain and secondary structures movement and ligand binding pose.[12] These deep learning methods have garnered significant interest and have achieved state-of-the-art performance on benchmarks, so we evaluate their ability to predict allosteric binders.

Improvements in accuracy of blind docking methods assist in scenarios when the target binding site is unknown. In drug discovery campaigns where the binding site is chosen or determined by pocket-searching techniques, the alternative approach to blind docking is known-pocket docking. In this approach, the search-space is restricted to the specific binding pocket and can also have strong performance.[13–15] These diffusion generative models have even been adapted for known-pocket docking.[16] Even if the binding site is known, a blind docking method could still improve hit enrichment of allosteric compounds compared to known-pocket docking. If a blind docking method can robustly discriminate between orthosteric and allosteric compounds, one can filter away the compounds more likely to be orthosteric binders resulting in a set of compounds more likely to be allosteric.

Although deep learning models can generate protein–ligand poses, selection of the receptor structure is non-trivial because proteins are dynamic and show conformational diversity. Kinases are known to fold into one active conformation but multiple inactive conformations. These conformations regulate cellular physiology and are key factors for protein–protein and protein–ligand interactions. There are a few heuristic-based kinase conformation classification methods.[17–20] While these methods can classify structures into multiple categories, they lack applicability to other proteins and require *a priori* knowledge of the structural determinants of kinase activation/inactivation. These classification methods highlight the DFG motif and the αC-helix region. This limits the capability of the classification method to process features outside the catalytic binding domain and classify protein structures based on diverse allosteric sites.

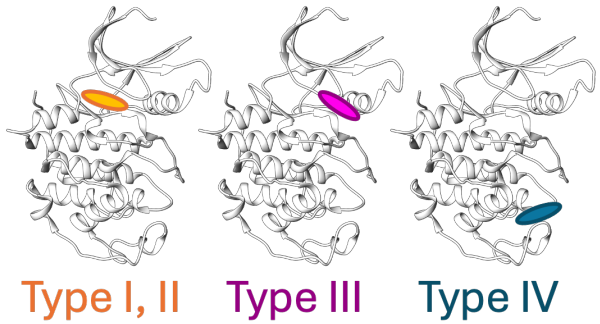


Figure 1: Kinase binding modes and their respective binding pockets. Type I and II ligands bind in the ATP binding site. Type III ligands bind in the site adjacent to the ATP site. Type IV ligands are a more general category allosteric binders where binding occurs at any site proximal to the orthosteric site.

Kinase inhibitor binding modes are also well-categorized. Type I and II inhibitors are orthosteric binders to the active and inactive conformations, respectively. Type III and IV inhibitors are allosteric binders at the site adjacent or proximal to the ATP-site, respectively (**Figure 1**).[21] Some methods use distance based heuristics or rely on superimposition to define the kinase binding modes.[13,14,20] Other methods are principally ligand-based methods and do not take a structural approach.[22–24]

We introduce a data-driven platform called the Minimum Distance Matrix Representation (MDMR) to survey the receptor conformations and ligand binding modes of known Cyclin-Dependent Kinase (CDK) 2 structures (**Figure S1**). A protein–ligand complex structure is represented as a matrix of pairwise residue–residue (or residue–ligand) distances. The distance, when defined as the minimum residue–residue (or residue–ligand) distance, is a proxy for inter-residue (or residue–ligand) interactions. The analysis of these matrices with unsupervised machine learning methods guide the development of benchmarks that assess the capability of docking methods to accurately predict binding poses, also known as docking power.[25] The self-docking test partitions docking success into orthosteric and allosteric binding modes. The cross-docking test relies on a time split and evaluates the permutations of the ligand binding mode and receptor conformation. These benchmarks reveal that the success of blind docking allosteric poses depends on the selection of the receptor conformations and whether methods can discriminate between orthosteric and

allosteric modes.

CDK2 is a promising therapeutic target. CDK2 relies on interaction with cyclin proteins for activation. The dysregulation of CDK2 activity is associated with the proliferation of cancer cells and most known inhibitors bind to a highly-conserved orthosteric binding site.[26–28] Recently, CDK2 has also been identified as male non-hormonal contraceptive target.[29] CDK2 has a diverse set of publicly available data leading to the development of strategies such as ensemble docking[30–32] and cross-docking[33–37] to evaluate the role of receptor conformations. Much of the effort thus far focuses on the ATP-binding site but there has yet to be evaluation of computational approaches to the discovery and prediction of allosteric inhibitors to CDK2.[38,39] We retrospectively apply our platform to the discovery of Type III inhibitors to Cyclin-Dependent Kinase (CDK) 2 (**Figure S2**). Researchers are interested in this binding mode because it offers an opportunity for novel mode of inhibition.[29,40] Recently, Faber *et al.* resolved a series of Type III inhibitors with an anthranilic acid scaffold bound to CDK2.[41,42] We survey whether or not these state-of-the-art blind docking methods can predict this allosteric binding mode and if success is conditional on the selection of receptor conformation.

Methods

Residue selection for matrix embedding

The MDMR allows us to classify and assess the conformational diversity of the structural ensemble. There are supplementary functions to reporting the distribution of structure resolution, missing residues, modifications, mutations and binding partners. It also supports visualization and identification of the defining interactions of the conformational clusters.[43–45]

The first stage is the selection and alignment of the residues. Because we only focus on one target, we use the entire length of the UniProt sequence, amounting to 298 residues. Then a n x n pairwise residue–residue distance matrix is calculated for each of structure, where the distance is the minimum residue–residue distance. The data matrix is cleaned by simply removing

the residues from the matrices that are not available in every structure. Contact maps or Cα–Cα distance matrices can also be equally valid or informative representations (**Figure S3**). The results here primarily explore the utility of the minimum distance matrices representation.

Other residues selections can also be useful. Comparison of protein structures with different sequences relies the observation that evolutionary sequence variation directs the folding constraints of protein structure.[46–49] In this instance, a multiple sequence alignment (MSA) of the protein subfamily is generated using Clustal Omega.[50] The MSA provides the basis for comparing structures by identifying the residues that play a role in the protein structural fold and also aligning corresponding residues to one another across different sequences. The identically-, conservatively- and semi-conserved residues of the alignment across CDK1–20 amounts to 93 residues. Other structural alignments, such as the alignment outlined by the Kinase–Ligand Interaction Fingerprints and Structures (KLIFS) database can also be useful.[51,52] This alignment is a manual structural and sequence alignment of 85 residues in the kinase catalytic site. This may allow for more informative comparisons of the kinase catalytic site across the kinome but at the expensive of excluding other regions. (**Figure S3**) Post-processing methods such as those using persistent spectral graph theory could reduce the need for careful selection of residues or handling of missing residues.[53] Combinations of these alignments may also be valid or be more informative dependent on your use case.

Normalized Standardized Mean Difference

Further analysis of these clusters highlights interactions as the defining characteristic that elucidates similarities and dis-similarities between conformations. With the structures classified into multiple clusters, the distance pairs that differentiated the two clusters can be determined using a standardized mean difference (SMD) metric (1). This SMD metric quantifies the effect size of

each *i*th and *j*th residue–residue distance pair between cluster *x* and *y*.

$$\mathrm{SMD}_{ij}^{x|y} = \frac{\bar{d}_{ij}^{x} - \bar{d}_{ij}^{y}}{\sqrt{\sigma_{ij}^{x} \cdot \sigma_{ij}^{y}}} \tag{1}$$

$$\sigma_{ij} = \sqrt{\sum_{i=1}^{N} (d_{ij} - \bar{d}_{ij})^2} \tag{2}$$

Where $\bar{d}_{ij}^{x}$ is the average distance between *i*th and *j*th residues of cluster *x*, and σ_{ij}^{x} is the standard deviation of the *ij* distance pairs from the *N* structures in cluster *y*. To further bias the residue–residue distances that form interactions, the SMD metric is normalized by dividing by the difference between minimum *ij* distance pair and a normalization value α (3).

$$\mathrm{nSMD}_{ij}^{x|y} = \frac{\mathrm{SMD}_{ij}^{x|y}}{\min(d_{ij}) - \alpha} \tag{3}$$

This α value is set to 1.5 to reflect the possibility of hydrogen atoms that are often not observed in X-ray crystal structures or it can be set to 0 when analyzing NMR structures or structures with hydrogen atoms present. A high nSMD value reflects a difference in the distance pair between clusters.

Protein Loop Reconstruction

Prior to performing blind docking, we use a template-based approach with Modeller to model missing residues segments that selects a similar protein conformation as the template.[54] This approach is herein referred to as *conformation-based loop modeling*. The selection of template structures takes advantage of utility of the MDMR to robustly cluster and classify protein structures by their respective conformations. This method is useful for repairing structures that have similar conformations from the same or related protein.

The selection of template structures begins following the dimension reduction and clustering of the receptor-only structures. Firstly, the missing residues for each PDB structure are collected. Any

unresolved terminal residues are excluded. Then, each structure is assigned to template structures based on the closest Principal Component Analysis (PCA)-projected datapoints, the resolution of corresponding structures, and cluster identity.

The reconstruction of the missing loop regions uses an adapted version of the `AutoModel` class in Modeller. A model with the corrected sequence is quickly rebuilt. Any non-canonical amino acids is converted back to the original amino acid as defined by the UniProt sequence. The missing regions are refined based on the assigned template structures. 5 models are generated with thorough molecular dynamics (MD) optimization and addition of hydrogens. The best model by lowest DOPE score is selected for docking.

Blind Docking

For the traditional docking methods Vina and Lin_F9, each receptor conformation is prepared using the pdb2pqr30 v3.5.2 to determine protonation states, add hydrogens, and rebuild any missing side chains.[55] Next, MGLTools 1.5.6 'prepare_receptor4.py' script is used to add Gasteiger charges and AutoDock Vina types, and remove the non-polar hydrogens.[56] The ligand preparation procedure begins with obtaining the SMILES representation of each compound. RDKit version 2020.03.1 is used to read in the compounds in the SMILES format, add hydrogens, generate initial 3D conformers using ETKDG and then optimize with the MMFF94 force field.[57,58] A matrix is then created by pairwise aligning each conformer and calculating RMSD. The matrix is then clustered using the Butina algorithm using a 2 Å threshold.[59] Then each ligand conformer is prepared by using Meeko 0.3.3 to add Gasteiger charges and AutoDock Vina types, and remove the non-polar hydrogens.[60] The cluster defining ligand conformers are then docked using the traditional docking methods with the hyper-parameters listed below.

- Vina 1.2.3:[5,6] Blind docking is performed using exhaustiveness=64 and num_modes=9. The search box is generated around the entire receptor following the autobox procedure outlined in Smina using a +4 Å buffer.

- Lin_F9:[7] Blind docking is performed using Smina docking method with the Lin_F9 scoring function. Exhaustiveness=64 and num_modes=9. The search box is generated around the receptor with a +4 Å buffer.

For the deep learning blind docking methods, the ligand is generated starting from the SMILES string and then docked onto the receptor with the hyper-parameters listed below.

- DiffDock (2023):[10] 10 samples per complex are generated using default parameters
- DiffDock-L:[11] 10 samples per complex are generated using default parameters
- DiffDock-S:[11] 10 samples per complex are generated using default parameters
- DynamicBind:[12] 10 samples per complex are generated using default parameters

For the combination of DiffDock and Lin_F9 local redocking, the preparation begins by following the preparation for the deep learning blind docking methods. Next, the receptor and ligand preparation for traditional docking was performed. The hyper-parameters for this strategy are listed below.

- DiffDock (2023)[10] + LRD:[7] First run DiffDock with default parameters then perform Lin_F9 Local Re-Docking (LRD). Lin_F9 LRD performs traditional docking using the Lin_F9 scoring function with the box that is auto-generated on the DiffDock pose with +4 Å buffer added, with exhaustiveness=8 and num_modes=1.

Results and Discussion

Minimum-distance matrix approach to embed protein conformation

Key to the SBDD of allosteric inhibitors is the exploration of the conformational states of kinases. The MDMR allows us to cluster and assess the conformational diversity of the structural ensemble. 453 available X-ray crystallography and Cryo-EM experimental entries before 04/13/2023

containing at least one structure of CDK2 (UniProt ID P24941) were downloaded from the RCSB database.[61,62] Parsing only the CDK2 receptor structure from each file amounts to 542 structures of CDK2. Each CDK2 structure is represented as a matrix of pairwise residue–residue distances. The distance is defined as the minimum residue–residue distance and is used as a proxy for inter-residue interactions.

Next, two unsupervised machine learning methods, Principal Component Analysis (PCA) and clustering, are applied to the distance matrices. PCA is an unsupervised machine learning algorithm that reduces the dimensionality of the data while maximizing variance.[63] The algorithm evaluates principal components (PC), uncorrelated linear functions of the dataset along which variance is maximized. The structures are projected onto PCs and the structural similarities and dissimilarities are observed. By projecting each matrix representation onto a few PCs rather than projecting onto the $n(n\text{-}1)/2$ unique features of the matrix, the complexity and dimensionality of the representation are greatly reduced. The PCA-projected data points from the PCs, which cumulatively contribute up to 0.8 of the explained variance ratio, are then clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) **(Figure 2A)**.[64,65] This clustering method relies on the observation that cluster centroids exhibit higher local density compared to their neighbors and are far from other points of higher density. By creating a hierarchy of centroids and its neighbors, the method extracts the set of significant clusters. The combination of these methods has been successfully applied to molecular dynamics (MD) sets in the past.[66] This analysis of the receptor–only representation space classifies similar conformations together and distinguish dissimilar conformations by virtue of their interactions.

The resulting analysis of the CDK2 receptor-only dataset suggests that there are 4 main clusters with clear separation across PC1 (**Figure 2A**). The binding partners details of each cluster indicate green ($c3^R$) and red ($c4^R$) clusters are comprised of structures complexed with cyclin and SPY1 activator proteins (**Figure S4**). Although these clusters are distinguished along PC2, the explained variance ratio of PC2 is low suggesting the variance of the features along this principal component is not as significant as the variance along PC1. Thus, these clusters are grouped

together and labelled the e active state. The purple ($c1^R$) cluster is comprised of structures often bound by inhibitors, and the cyan ($c2^R$) cluster is comprised of structures bound by the allosteric probe, 8-anilino-1-naphthalene sulfonate (ANS; PDB ligand ID: 2AN). These are labelled as the inactive and intermediate states, respectively. The separate clustering of structures bound by the ANS reflect the observations by Betzi *et al.* that the ANS ligand induces or binds to a different conformation that is not observed in cyclin-bound states or typical inactive states.[67] We note that none of the aforementioned kinase conformation classification methods distinguish these intermediate structures (**Table S1–S3**).[17–20] In summary, the receptor-only clustering reveals 4 clusters where $c3^R$ and $c4^R$ are denoted as the active state, $c1^R$ as the inactive state, and $c2^R$ as the intermediate state.

Next, a normalized standardized mean difference metric (nSMD) for each distance pair highlights the variance in the matrices and identifies the significant interactions distinguishing conformation clusters. The nSMD between the active ($c3^R$: green) and an inactive cluster ($c1^R$: purple) is calculated and plotted against the minimum distance of that distance pair across all the structures (**Figure 2B**). All the distance pairs with a nSMD <5 and minimum distance <3.5 Å are categorized as R3, interactions that occur in structures in the active cluster ($c3^R$) but not the inactive cluster ($c1^R$) **(Figure 2B)**. The highlighted interactions reflect important structural and catalytic roles. The top ranking nSMD interaction is the important LYS33–GLU51 salt bridge, which is a prerequisite for the formation of active conformation in protein kinases and is the defining feature of the αC helix-in (αC-in) conformation.[27] This interaction is not observed in the inactive cluster and the GLU51 side chain is turned away from the active site, consistent with the αC helix-out (αC-out) conformation. Other notable interactions include the electrostatic interaction between GLU57 and ARG122, and regulatory spine residue LEU55 (RS3) hydrophobic interactions with the shell residues VAL64 (Sh1) and PHE80 (Sh2) as defined by Kornev *et al.*[27,68] Conversely, distance pairs with a nSMD $<$-5 and minimum distance <3.5 Å are categorized as R1, interactions that occur in the structures in the inactive cluster but not the active cluster (**Figure 2B**). These R1 distances mostly reflect interactions between residues in the αC-helix and β5 sheet. These interactions appear

to stabilize the αC-out conformation and force GLU51 to be solvent exposed. The histograms of the distance distributions also confirm these observations (**Figure S5**). This analysis underscores the interactions that distinguish or qualify these clusters of structures and that the common αC-in vs αC-out nomenclature that is necessary for distinguishing the active CDK2 conformation from the inactive conformation.[27]

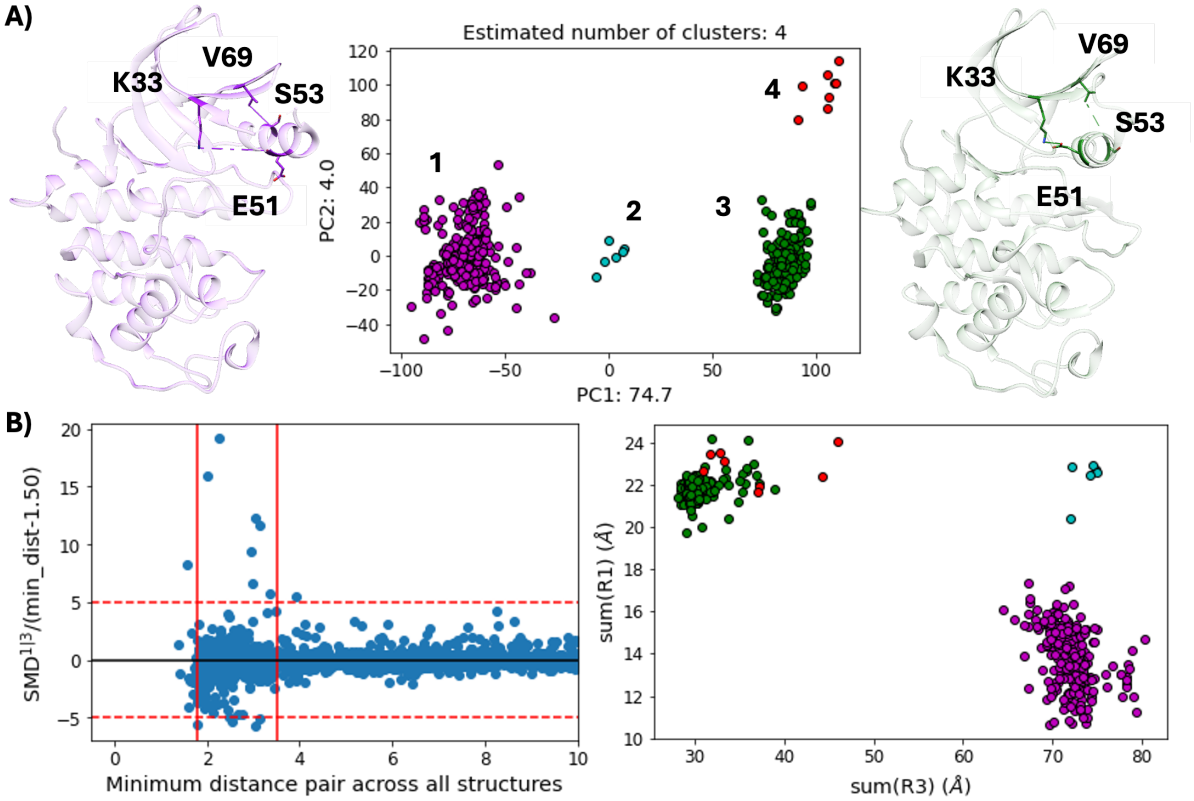
The sum of the R3 distances versus sum of the R1 distances plot is an informative plot about the interactions **(Figure 2B)**. Here, the active structures are all clustered distinctly from the inactive structures clustered. The ANS-bound, intermediate structures cluster to the top right of the plot indicating that both R1 and R3 distances are broken and highlighting the uniqueness of the conformation. In the crystal structure of an intermediate conformation, the CDK2–ANS ternary complex (PDB ID: 3PXF), two ANS ligands bind to the allosteric pocket adjacent to the ATP site between αC-helix and the β4 and β5 sheets. The primary ANS ligand interrupts important R1 and R3 distances by interacting with LYS33, PHE80, LYS56 and VAL64, while the secondary ANS ligand interrupts R3 interactions by interacting with LYS56 and VAL69 (**Figure S6**).

We perform a triplicate of 500 ns MD simulations initialized from the active, intermediate and inactive conformations. We produce density histograms of the sum(R1) vs sum(R3) to track the conformational state of CDK2 during the simulations (**Figure S7**). The results indicate that all three, active, intermediate and inactive conformations are meta-stable in the *holo* conformation because they remain in a similar state to their respective crystal structures **(Figure S7A,B,C)**. *Apo* simulations starting the intermediate conformation populate the inactive regime. This suggests that the intermediate conformation may not be observed natively and the ANS-probe is required to be bound for the conformation to retain the meta-stable state where both the R1/R3 distances broken **(Figure S7C,D)**.

This minimum distance matrix is one of many possible representations of a protein conformation. Some methods first require superimposition or RMSD,[69–71] or use differential geometry,[39,72–74] or take combinations of input features.[75,76] The minimum residue–residue distance structural representation presented here is both superposition independent and incorporates the

combined effect of changing backbones and side chains. The resulting PCA analyses of the minimum distance matrices show clear separation of conformations and can provide novel insights of interactions that define these conformations.

The MDMR has been applied to other systems as well. It has been used to distinguish the active and auto-inhibited conformations of the FGFR kinase isoforms and elucidate how mutations of the gatekeeper residue alter the conformations defining interactions.[77] It has also assisted the identification of the binding site of a hit compound to the SARS-CoV-2 nsp13 by classifying the receptor conformation and performing blind ensemble docking.[78] In both of these cases, this approach details downstream analyses in the context of the conformational ensemble.



R1: Interactions stabilizing inactive state				R3: Interactions stabilizing active states			
Res 1	Res 2	Min dist (Å)	Norm(SMD)	Res 1	Res 2	Min dist (Å)	Norm(SMD)
SER 53	VAL 69	3.034	-5.678	LYS 33	GLU 51	2.264	19.185
LYS 56	ASP 65	2.763	-5.269	GLU 57	ARG 122	2.204	15.951
LEU 55	VAL 123	2.546	-5.012	GLU 51	LEU 78	3.043	12.247
				GLU 51	PHE 80	3.147	11.666
				LYS 34	GLU 51	2.938	9.462
				LEU 55	PHE 80	2.993	6.614
				LEU 55	VAL 64	3.364	5.695

Figure 2: **A)** Receptor-only Principal Component Analysis (PCA) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) clustering. Each available CDK2 protein conformation is represented by pairwise residue–residue minimum distance matrix and then input into PCA and HDBSCAN. Example structures from the inactive cluster ($c1^R$; purple) and the active cluster ($c3^R$; green) with the most distinguishing distance pair shown in sticks **B)** *Upper left:* Selected residue–residue distance pairs that have 1.8 Å $<\min(d_{ij}^{1|3})$ <3.5 Å and $\text{nSMD}^{1|3}$ >5 (R3) or $\text{nSMD}^{1|3}$ <5 (R1) reflect interactions that distinguish cluster 1 from 3. The solid and dashed red lines on the plot indicate the thresholds for selecting the R1 and R3 interactions of the minimum distance and nSMD, respectively. *Upper right:* These distances were summed up to separate the CDK2 structures along two axes, R1 and R3. *Below:* Table of selected distance pairs and the minimum distance and nSMD of the corresponding cluster. The color of the text corresponds to the region of the protein as depicted in **Figure S2**.

Self- and cross-docking benchmarking strategy to assess deep-learning blind-docking methods

The minimum distance metric also naturally extends to embed the ligand binding location. After removing structures that are *apo* or have multiple of the same ligand bound, we calculate the receptor–ligand minimum distance vector and input this new feature set into PCA and HDBSCAN (**Figure 3**). This results into two main clusters, orthosteric binders in the ATP binding site ($c1^L$; yellow) and allosteric binders in the Type III binding site ($c2^L$; pink). Ligand binding locations that are considered noise are represented with an $\times$ and are excluded from further analysis. Cross-referencing between the receptor-only and ligand-only clustering then reveals the variety of binding modes within the data. In principle, similar binding mode analyses can be performed with other methods when extended to include ligand or pocket-based features.[79,80]

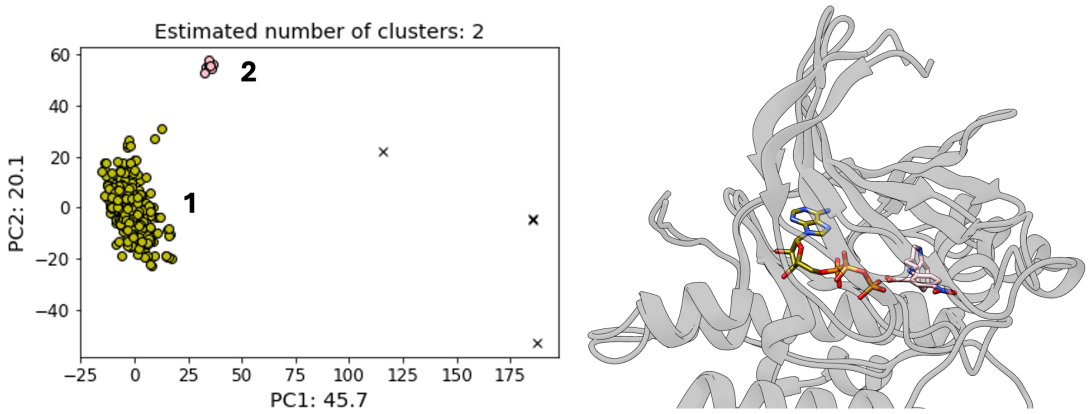


Figure 3: Ligand-only PCA and clustering. The ligand binding location is represented by the residue–ligand minimum distance vector and then input into PCA and HDBSCAN. The docking power of blind docking methods is benchmarked to predict orthosteric ($c1^L$; yellow) and allosteric ligands ($c2^L$; pink). Noisy data points are represented with an $\times$ are excluded. Example structures of orthosteric- and allosteric-bound complexes are colored by the ligand-only clustering and represented in sticks.

Table 1: Self-docking benchmark. Fraction of complexes successfully predicted by blind docking models (RMSD <2 Å). The binding mode is defined by the clustering observed in **Fig 3** and receptor conformation is defined by the clustering in **Fig 2A.** LRD: Local Re-Docking

Binding Mode	Type III bound to inactive state ($c2^L$–$c1^R$; n=7)	Orthosteric ligands to inactive state ($c1^L$–$c1^R$; n=8)	Orthosteric ligands to active state ($c1^L$–$c3^R$; n=8)
Vina[5]	1.000	0.375	0.250
Lin_F9[7]	1.000	0.125	0.500
DiffDock[10]	0.000	1.000	1.000
DiffDock-L[11]	0.000	1.000	1.000
DiffDock-S[11]	0.000	1.000	1.000
DynamicBind[12]	0.000	0.875	1.000
DiffDock+LRD[7,10]	0.429	0.375	0.500

The exploration of the receptor-only and ligand-only representation spaces are then leveraged to a couple of docking related tasks downstream as well. First, we reconstruct missing loop regions using *conformation-based loop modeling*. This method uses nearby conformations in receptor-only space as template structures for remodelling. Next, we develop target-specific benchmarks that evaluate blind docking methods to predict ligands binding poses in allosteric sites and to discriminate between orthosteric and allosteric ligands.

We create the self-docking benchmark, which assesses the capability of docking methods to accurately predict the pose when given the original receptor conformation and groups results depending on the binding mode (**Figure 2,3**). We select 7 ligands from each cluster plus the native orthosteric ligand, ATP, and ensure that there is no leakage of these test complexes in the training set (**Table S4**). We report the fraction of complexes with at least one successfully predicted pose with RMSD <2 Å to the reference crystal structure (**Table 1**). The deep-learning blind-docking methods are able to predict the orthosteric ligands but are all unable to predict the Type III binding mode.[10–12] Interestingly, we observe that traditional docking methods perform best when predicting the Type III binding pose, but have limited success predicting the orthosteric binding pose. Similarly, a combination of deep learning and traditional methods, DiffDock followed by Lin_F9[7] Local Re-Docking (DiffDock+LRD), is able to predict the Type III binding pose but loses accuracy

with orthosteric binders. These results on Type III binders are consistent with the observations by Yu *et al.*, where the best overall performance is observed when these deep learning docking methods performs the role of pocket searching and traditional docking as pocket refinement.[15]

Table 2: Cross-docking benchmark. Fraction of complexes successfully predicted by deep learning models (all-Cα alignment centroid distance <5 Å). Each Type III ligand is docked a set of receptor conformations where I=inactive (purple; $c1^R$; n=8), Mid=intermediate (cyan; $c2^R$; n=3), active (green; $c3^R$; n=8)). If any pose generated from a specific receptor–ligand pair meets the criteria, then the prediction is marked as successful. The binding mode is defined by the clustering observed in **Fig 3** and receptor conformation is defined by the clustering in **Fig 2A.**

Binding Mode	Type III bound to inactive state ($c2^L$–$c1^R$; n=7)		
Receptor conformation	I	Mid	A
Vina	0.000	0.571	0.000
Lin_F9	0.054	0.048	0.000
DiffDock	0.000	0.143	0.000
DiffDock-L	0.000	0.000	0.000
DiffDock-S	0.000	0.000	0.000
DynamicBind	0.000	0.000	0.000
DiffDock+LRD	0.000	0.952	0.000

Next, we create a scenario that is akin to real-world screening by setting a time split. This limits receptor structures to those known prior to the determination of the allosteric complexes (08/02/2021; **Table S5, Figure S3** boxed). We then create the cross-docking benchmark by docking compounds which are labelled by their binding mode to the set of receptor conformations (**Table S4**, **Figure S8**). This assesses the capability of blind docking methods to accurately predict and discriminate between the binding modes depending on the given receptor conformation in a prospective-like manner.

Here we use a different metric, the ligand–centroid distance from crystal structure following all-Cα structural alignment <5 Å. This quantifies whether the ligand pose is spatially similar to the reference crystal structure (**Table 2**). We then report the fraction of complexes that meet these criteria. We observe that any conformation can be used to recreate the orthosteric ligand pose and that deep learning docking methods more accurately predict the orthosteric ligand pose than

traditional docking methods (**Table S6**). The cross-docking results indicate the converse for the prediction of the Type III ligand pose (**Table 2**). Here, the ligands must be docked to the intermediate receptor conformation and the results follow methodology-dependent trends that are similar to the self-docking results. We observe that methods involving traditional docking are able to sample this allosteric mode at a higher rate compared to purely deep learning methods. A combination of approaches, DiffDock+LRD, samples this mode best when given the intermediate ANS–bound receptor structure (**Table 2**). These results are consistent with the broadly accepted notion that traditional docking performs best when given a structure that had been resolved with a ligand bound in the desired binding site.[81–83] These results are in consistent with to the presence of an orthosteric binding site in all conformations but the absence of a highly ligandable Type III-binding site in the active and inactive conformation (**Figure S9B,C**). A flexible pocket docking method, DynamicBind, can be promising in predicting ligand bound poses using unbound/holo protein structures as a receptor.[12] This method can improve docking performance on structures where the Type III site is at least partially ligandable, as in the inactive conformation, but we are unable to observe this in our case (**Figure S9C**).

Conclusion

An exploration of the protein conformational space and the known ligand binding modes is an essential first step to any drug discovery campaign. This case study with CDK2 highlights the utility of MDMR to providing context and visualization for the structural determinants of the conformational ensemble. This data-driven platform surveys ligand binding modes and protein conformations and identifies a unique intermediate receptor conformation. The survey guides benchmarking deep learning blind docking methods contingent on the ligand binding mode and protein conformation through self- and cross-docking. In the self-docking benchmark, we observe traditional docking methods predict the allosteric binding mode the best, deep learning methods predict the orthosteric binding mode the best, and a combination of traditional and deep learning methods,

DiffDock+LRD, has balanced performance across binding modes. In the prospective-like scenario from the cross-docking benchmark, we observe the best allosteric binding pose prediction occurs when using DiffDock+LRD on the intermediate conformation. As the field develops and the number of structures grow for this therapeutic mode of inhibition, this platform can be used to test and compare future targets and methods.

Our case study reveals the complications of using deep learning blind docking methods for SBDD of allosteric compounds. First, allosteric pose prediction is only observed when given a conformation where the receptor is ligand-bound at this site. In contrast, orthosteric compounds are able to be docked regardless of conformation or method. This observation highlights the need to identify or generate diverse receptor conformations for docking.[84,85] Benchmarking and development of methods that can produce diverse conformations would greatly aid the discovery of allosteric compounds. On one hand, there exists strategies that do not rely on deep learning like homology modelling[54,86] and enhanced sampling simulations.[87] On the other hand, there exists deep learning sampling strategies such as AlphaFold2-based methods that sub-sample the multiple sequence alignment[88–91] and those that produce distributions of conformations[92,93] or reveal cryptic pockets.[94] Second, there are fewer allosterically-bound kinase structures and binding data points compared to their orthosteric counterparts (**Figure S10**). Thus, diffusion-based generative models trained on this data would be more likely to sample these high-density regions. This imbalanced data regime begets the need to adapt these generative blind docking models to sample high fidelity poses from the low-density regions.[95,96]

It is important to acknowledge this retrospective docking study is biased by past results and does not preclude that these methods or selection of other conformations could be successful in future campaigns for allosteric binders.[97,98] These allosteric binding poses are influenced by the receptor structure, and the receptor conformation is biased to the ligands and methods with which it was determined. A prospective study, where a multitude of ligands with diverse chemical properties are screened and are resolved in a multitude of conformations, can assess blind docking methods and reveal the utility of selecting the correct conformation in an unbiased manner.[98] Fur-

thermore, increasing the number of sampled poses or conformations can also improve the results.

Future work aims to explore other use cases for this representation and platform. For example, this representation can be used as training data for a pose or conformation classifier. Instead of highlighting the distinguishing distances, we can determine structurally conserved residues by selecting low variance cliques (**Figure S11**). This method can be extended to study families of related proteins with an appropriate selection of residues through a multiple sequence alignment or structural alignment. Other work aims to expand this framework beyond a target-specific task and show whether trends observed in this work are consistent among other allosteric binders to other kinases or proteins. Improvement of the generalization capabilities of deep learning methods to sample therapeutically relevant protein conformations and to predict allosteric binding poses is essential to the SBDD of allosteric compounds. A diverse database of allosteric and peptide binders to the kinome is required to quantitatively assess the preference of allosteric binding modes for certain receptor conformation and establish pocket-specific fine-tuning strategy for generative blind docking models.[99,100] Lastly, it will also be necessary to assess heterogeneous complex prediction methods that combine receptor fold generation and ligand pose prediction, such as NeuralPlexer,[101] RoseTTAFold-AllAtom/RoseTTAFold Diffusion[102] or AlphaFold3.[103] These methods are poised to circumvent the need to select the correct receptor conformation to dock to by generating the complex in full.

Data and Software Availability

The dataset and analyses underlying this study are available at `https://github.com/echen1214/dist_analy` or at `https://doi.org/10.5281/zenodo.13964938`. The Jupyter notebooks that underlie the generation of the figures and tables from the MDMR platform and for self- and cross-docking dataset and results can be found in the ‘tutorial‘ directory. The notebooks contain data visualization tools to assist the analysis of the plots. The raw receptor poses and ligand binding poses are in the ‘tutorial/datafiles‘ directory. The list of structures and ligands used for self- and cross-docking can be found in the supporting information. The docking poses, input topologies,

coordinates, and simulation control files are provided in the can be found in the ‘Zenodo‘ database.

Acknowledgement

The authors thank Qi Ouyang, Justin Green, John Arnell, Alida Besch, Song Xia, Zixuan Feng, Alan Robledo and Thomas Kelly for their discussions on the paper. Y.Z. acknowledges support from the U.S. National Institutes of Health (NIH) (R35-GM127040). E.C. is partially supported by a graduate fellowship from the Simons Center of Computational Physical Chemistry (SCCPC) at NYU.

Supporting Information Available

Computational methods of the file processing, MD simulations, lDDT pocket matching, and low variance cliques and supporting tables and figures.

References

(1) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. The ensemble nature of allostery. *508*, 331–339.

(2) Wodak, S. J. et al. Allostery in Its Many Disguises: From Theory to Applications. *27*, 566–578.

(3) Lu, S.; He, X.; Ni, D.; Zhang, J. Allosteric Modulator Discovery: From Serendipity to Structure-Based Design. *62*, 6405–6421.

(4) Lu, X.; Smaill, J. B.; Ding, K. New Promise and Opportunities for Allosteric Kinase Inhibitors. *59*, 13764–13776, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201914525.

(5) Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *31*, 455–461.

(6) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *61*, 3891–3898, Publisher: American Chemical Society.

(7) Yang, C.; Zhang, Y. Lin_F9: A Linear Empirical Scoring Function for Protein–Ligand Docking. *61*, 4630–4644, Type: Journal Article.

(8) Yim, J.; Stärk, H.; Corso, G.; Jing, B.; Barzilay, R.; Jaakkola, T. S. Diffusion models in protein structure and docking. *14*, e1711, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1711.

(9) Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations.

(10) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. arXiv:2210.01776, Type: Journal Article.

(11) Corso, G.; Deng, A.; Fry, B.; Polizzi, N.; Barzilay, R.; Jaakkola, T. Deep Confident Steps to New Pockets: Strategies for Docking Generalization. `http://arxiv.org/abs/2402.18396`.

(12) Lu, W.; Zhang, J.; Huang, W.; Zhang, Z.; Jia, X.; Wang, Z.; Shi, L.; Li, C.; Wolynes, P. G.; Zheng, S. DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *15*, 1071, Number: 1 Publisher: Nature Publishing Group.

(13) Sturm, N.; Tinivella, A.; Rastelli, G. Exploration and Comparison of the Geometrical and Physicochemical Properties of an αC Allosteric Pocket in the Structural Kinome. *58*, 1094–1103, Publisher: American Chemical Society.

(14) Yueh, C.; Rettenmaier, J.; Xia, B.; Hall, D. R.; Alekseenko, A.; Porter, K. A.; Barkovich, K.; Keseru, G.; Whitty, A.; Wells, J. A.; Vajda, S.; Kozakov, D. Kinase Atlas: Druggability Analysis of Potential Allosteric Sites in Kinases. *62*, 6512–6524, Publisher: American Chemical Society.

(15) Yu, Y.; Lu, S.; Gao, Z.; Zheng, H.; Ke, G. Do Deep Learning Models Really Outperform Traditional Approaches in Molecular Docking? `http://arxiv.org/abs/2302.07134`.

(16) Plainer, M.; Toth, M.; Dobers, S.; Stärk, H.; Corso, G.; Marquet, C.; Barzilay, R. DiffDock-Pocket: Diffusion for Pocket-Level Docking with Sidechain Flexibility.

(17) Möbitz, H. The ABC of protein kinase conformations. *1854*, 1555–1566.

(18) Ung, P. M.-U.; Rahman, R.; Schlessinger, A. Redefining the Protein Kinase Conformational Space with Machine Learning. *25*, 916–924.e2, Publisher: Elsevier.

(19) Modi, V.; Dunbrack, R. L. Defining a new nomenclature for the structures of active and inactive kinases. *116*, 6818–6827.

(20) Modi, V.; Dunbrack, R. L., Jr Kincore: a web resource for structural classification of protein kinases and their inhibitors. *50*, D654–D664.

(21) Pan, Y.; Mader, M. M. Principles of Kinase Allosteric Inhibition and Pocket Validation. *65*, 5288–5299, Publisher: American Chemical Society.

(22) Zhou, Y.; Al-Jarf, R.; Alavi, A.; Nguyen, T. B.; Rodrigues, C. H. M.; Pires, D. E. V.; Ascher, D. B. kinCSM: Using graph-based signatures to predict small molecule CDK2 inhibitors. *31*, e4453, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4453.

(23) Abdelbaky, I.; Tayara, H.; Chong, K. T. Prediction of kinase inhibitors binding modes with machine learning and reduced descriptor sets. *11*, 706, Publisher: Nature Publishing Group.

(24) Miljković, F.; Rodríguez-Pérez, R.; Bajorath, J. Machine Learning Models for Accurate Prediction of Kinase Inhibitors with Different Binding Modes. *63*, 8738–8748, Publisher: American Chemical Society.

(25) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *59*, 895–913, Type: Journal Article.

(26) Lapenna, S.; Giordano, A. Cell cycle kinases as therapeutic targets for cancer. *8*, 547–566, Number: 7 Publisher: Nature Publishing Group.

(27) Roskoski, R. Cyclin-dependent protein serine/threonine kinase inhibitors as anticancer drugs. *139*, 471–488.

(28) Tadesse, S.; Anshabo, A. T.; Portman, N.; Lim, E.; Tilley, W.; Caldon, C. E.; Wang, S. Targeting CDK2 in cancer: challenges and opportunities for therapy. *25*, 406–413.

(29) Faber, E. B.; Wang, N.; Georg, G. I. Review of rationale and progress toward targeting cyclin-dependent kinase 2 (CDK2) for male contraception†. *103*, 357–367.

(30) Campbell, A. J.; Lamb, M. L.; Joseph-McCarthy, D. Ensemble-Based Docking Using Biased Molecular Dynamics. *54*, 2127–2138, Publisher: American Chemical Society.

(31) Ricci-Lopez, J.; Aguila, S. A.; Gilson, M. K.; Brizuela, C. A. Improving Structure-Based Virtual Screening with Ensemble Docking and Machine Learning. *61*, 5362–5376, Publisher: American Chemical Society.

(32) Mohammadi, S.; Narimani, Z.; Ashouri, M.; Firouzi, R.; Karimi-Jafari, M. H. Ensemble learning from ensemble docking: revisiting the optimum ensemble size problem. *12*, 410, Publisher: Nature Publishing Group.

(33) Duca, J. S.; Madison, V. S.; Voigt, J. H. Cross-Docking of Inhibitors into CDK2 Structures. 1. *48*, 659–668, Publisher: American Chemical Society.

(34) Broccatelli, F.; Brown, N. Best of Both Worlds: On the Complementarity of Ligand-Based and Structure-Based Virtual Screening. *54*, 1634–1641, Publisher: American Chemical Society.

(35) Ravindranath, P. A.; Forli, S.; Goodsell, D. S.; Olson, A. J.; Sanner, M. F. AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. *11*, e1004586, Publisher: Public Library of Science.

(36) Shamsara, J. CrossDocker: a tool for performing cross-docking using Autodock Vina. *5*, 344.

(37) Shen, C.; Hu, X.; Gao, J.; Zhang, X.; Zhong, H.; Wang, Z.; Xu, L.; Kang, Y.; Cao, D.; Hou, T. The impact of cross-docked poses on performance of machine learning classifier for protein–ligand binding pose prediction. *13*, 81.

(38) Tutone, M.; Almerico, A. M. Recent advances on CDK inhibitors: An insight by means of in silico methods. *142*, 300–315.

(39) Pitt, W. R.; Montalvão, R. W.; Blundell, T. L. Polyphony: superposition independent methods for ensemble-based drug discovery. *15*.

(40) Rastelli, G.; Anighoro, A.; Chripkova, M.; Carrassa, L.; Broggini, M. Structure-based discovery of the first allosteric inhibitors of cyclin-dependent kinase 2. *13*, 2296–2305, Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/cc.29295.

(41) Faber, E. B. et al. Screening through Lead Optimization of High Affinity, Allosteric Cyclin-Dependent Kinase 2 (CDK2) Inhibitors as Male Contraceptives That Reduce Sperm Counts in Mice. *66*, 1928–1940, Publisher: American Chemical Society.

(42) Faber, E. B. et al. Development of allosteric and selective CDK2 inhibitors for contraception with negative cooperativity to cyclin binding. *14*, 3213, Publisher: Nature Publishing Group.

(43) Meng, E. C.; Goddard, T. D.; Pettersen, E. F.; Couch, G. S.; Pearson, Z. J.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Tools for structure building and analysis. *32*, e4792, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4792.

(44) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera–a visualization system for exploratory research and analysis. *25*, 1605–1612.

(45) VanderPlas, J.; Granger, B. E.; Heer, J.; Moritz, D.; Wongsuphasawat, K.; Satyanarayan, A.; Lees, E.; Timofeev, I.; Welsh, B.; Sievert, S. Altair: Interactive Statistical Visualizations for Python. *3*, 1057.

(46) Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated mutations and residue contacts in proteins. *18*, 309–317, Type: Journal Article.

(47) Lapedes, A. S.; Bertrand, G. G.; LonChang, L.; Stormo, G. D. Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. *33*, 236–256, Type: Journal Article.

(48) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *106*, 67–72, Type: Journal Article.

(49) Rao, R. M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; Rives, A. MSA Transformer. `https://proceedings.mlr.press/v139/rao21a.html`, Pages: 8844–8856 Type: Conference Paper Volume: 139.

(50) Sievers, F.; Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *27*, 135–145, Type: Journal Article.

(51) van Linden, O. P. J.; Kooistra, A. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A

Knowledge-Based Structural Database To Navigate Kinase–Ligand Interaction Space. *57*, 249–277, Type: Journal Article.

(52) Kanev, G. K.; de Graaf, C.; Westerman, B. A.; de Esch, I. J. P.; Kooistra, A. J. KLIFS: an overhaul after the first 5 years of supporting kinase research. *49*, D562–D569, Type: Journal Article.

(53) Wang, R.; Zhao, R.; Ribando-Gros, E.; Chen, J.; Tong, Y.; Wei, G.-W. HERMES: Persistent spectral graph software. *3*, 67–97, Publisher: Foundations of Data Science.

(54) Šali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *234*, 779–815, Type: Journal Article.

(55) Jurrus, E. et al. Improvements to the APBS biomolecular solvation software suite. *27*, 112–128, Type: Journal Article.

(56) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *30*, 2785–2791.

(57) Landrum, G. et al. rdkit/rdkit: 2020_03_1 (Q1 2020) Release. `https://zenodo.org/record/3732262`.

(58) Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *60*, 2044–2058, Publisher: American Chemical Society.

(59) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *39*, 747–750, Publisher: American Chemical Society.

(60) Meeko: preparation of small molecules for AutoDock. `https://github.com/forlilab/Meeko/tree/v0.3.3`.

(61) Burley, S. K. et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *49*, D437–D451.

(62) The UniProt Consortium UniProt: the Universal Protein Knowledgebase in 2023. *51*, D523–D531.

(63) Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments. *374*, 20150202, Type: Journal Article.

(64) Campello, R. J. G. B.; Moulavi, D.; Sander, J. In *Advances in Knowledge Discovery and Data Mining*; Pei, J., Tseng, V. S., Cao, L., Motoda, H., Xu, G., Eds.; Springer Berlin Heidelberg, Vol. 7819; pp 160–172, Series Title: Lecture Notes in Computer Science.

(65) Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *344*, 1492–1496, Publisher: American Association for the Advancement of Science.

(66) Slough, D. P.; McHugh, S. M.; Cummings, A. E.; Dai, P.; Pentelute, B. L.; Kritzer, J. A.; Lin, Y.-S. Designing Well-Structured Cyclic Pentapeptides Based on Sequence–Structure Relationships. *122*, 3908–3919, Publisher: American Chemical Society.

(67) 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. *6*, 492–501, Publisher: American Chemical Society.

(68) Kornev, A. P.; Taylor, S. S.; Eyck, L. F. T. A helix scaffold for the assembly of active protein kinases. *105*, 14377–14382, Publisher: National Academy of Sciences Section: Biological Sciences.

(69) Ye, Y.; Godzik, A. FATCAT: a web server for flexible structure comparison and structure similarity searching. *32*, W582–W585.

(70) Hayward, S.; Lee, R. A. Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *21*, 181–183.

(71) Echols, N.; Milburn, D.; Gerstein, M. MolMovDB: analysis and visualization of conformational change and structural flexibility. *31*, 478–482.

(72) Rackovsky, S.; Scheraga, H. A. Differential Geometry and Polymer Conformation. 1. Comparison of Protein Conformations1a,b. *11*, 1168–1174, Publisher: American Chemical Society.

(73) Leung, H. T. A.; Montaño, B. O.; Blundell, T.; Vendruscolo, M.; Montalvão, R. W. ARABESQUE: A Tool For Protein Structural Comparison Using Differential Geometry And Knot Theory. *1*, 8.

(74) Neto, A. M. d. S.; Silva, S. R.; Vendruscolo, M.; Camilloni, C.; Montalvão, R. W. A superposition free method for protein conformational ensemble analyses and local clustering based on a differential geometry representation of backbone. *87*, 302–312, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25652.

(75) Maschietto, F.; Allen, B.; Kyro, G. W.; Batista, V. S. MDiGest: A Python package for describing allostery from molecular dynamics simulations. *158*, 215103.

(76) Vögele, M.; Thomson, N. J.; Truong, S. T.; McAvity, J.; Zachariae, U.; Dror, R. O. Systematic Analysis of Biomolecular Conformational Ensembles with PENSA. `http://arxiv.org/abs/2212.02714`.

(77) Besch, A.; Marsiglia, W. M.; Mohammadi, M.; Zhang, Y.; Traaseth, N. J. Gatekeeper mutations activate FGF receptor tyrosine kinases by destabilizing the autoinhibited state. *120*, e2213090120, Publisher: Proceedings of the National Academy of Sciences.

(78) Soper, N.; Yardumian, I.; Chen, E.; Yang, C.; Ciervo, S.; Oom, A. L.; Desvignes, L.; Mulligan, M. J.; Zhang, Y.; Lupoli, T. J. A Repurposed Drug Interferes with Nucleic Acid to

Inhibit the Dual Activities of Coronavirus Nsp13. *19*, 1593–1603, Publisher: American Chemical Society.

(79) Chen, S.; Gao, J.; Chen, J.; Xie, Y.; Shen, Z.; Xu, L.; Che, J.; Wu, J.; Dong, X. ClusterX: a novel representation learning-based deep clustering framework for accurate visual inspection in virtual screening. *24*, bbad126.

(80) Al-Masri, C.; Trozzi, F.; Lin, S.-H.; Tran, O.; Sahni, N.; Patek, M.; Cichonska, A.; Ravikumar, B.; Rahman, R. Investigating the conformational landscape of AlphaFold2-predicted protein kinase structures. *3*, vbad129.

(81) McGovern, S. L.; Shoichet, B. K. Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes. *46*, 2895–2907, Publisher: American Chemical Society.

(82) An, X.; Lu, S.; Song, K.; Shen, Q.; Huang, M.; Yao, X.; Liu, H.; Zhang, J. Are the Apo Proteins Suitable for the Rational Discovery of Allosteric Drugs? *59*, 597–604, Publisher: American Chemical Society.

(83) Paggi, J. M.; Pandit, A.; Dror, R. O. The Art and Science of Molecular Docking.

(84) Nussinov, R.; Zhang, M.; Liu, Y.; Jang, H. AlphaFold, allosteric, and orthosteric drug discovery: Ways forward. *28*, 103551.

(85) Lane, T. J. Protein structure prediction has reached the single-structure frontier. *20*, 170–173, Publisher: Nature Publishing Group.

(86) Monzon, A. M.; Fornasari, M. S.; Zea, D. J.; Parisi, G. In *Computational Methods in Protein Evolution*; Sikosek, T., Ed.; Springer, pp 353–365.

(87) Bernardi, R. C.; Melo, M. C.; Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *1850*, 872–877.

(88) Stein, R. A.; Mchaourab, H. S. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *18*, e1010483, Publisher: Public Library of Science.

(89) del Alamo, D.; Sala, D.; Mchaourab, H. S.; Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *11*, e75751, Publisher: eLife Sciences Publications, Ltd.

(90) Monteiro da Silva, G.; Cui, J. Y.; Dalgarno, D. C.; Lisi, G. P.; Rubenstein, B. M. High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *15*, 2464, Publisher: Nature Publishing Group.

(91) Sala, D.; Hildebrand, P. W.; Meiler, J. Biasing AlphaFold2 to predict GPCRs and kinases with user-defined functional or structural properties. *10*, Publisher: Frontiers.

(92) Zheng, S. et al. Predicting equilibrium distributions for molecular systems with deep learning. *6*, 558–567, Publisher: Nature Publishing Group.

(93) Jing, B.; Berger, B.; Jaakkola, T. AlphaFold Meets Flow Matching for Generating Protein Ensembles. `http://arxiv.org/abs/2402.04845`.

(94) Meller, A.; Ward, M.; Borowsky, J.; Kshirsagar, M.; Lotthammer, J. M.; Oviedo, F.; Ferres, J. L.; Bowman, G. R. Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *14*, 1177, Number: 1 Publisher: Nature Publishing Group.

(95) Sehwag, V.; Hazirbas, C.; Gordo, A.; Ozgenel, F.; Ferrer, C. C. Generating High Fidelity Data from Low-density Regions using Diffusion Models. `http://arxiv.org/abs/2203.17260`.

(96) Corso, G.; Xu, Y.; de Bortoli, V.; Barzilay, R.; Jaakkola, T. Particle Guidance: non-I.I.D. Diverse Sampling with Diffusion Models. `http://arxiv.org/abs/2310.13102`.

(97) Kearnes, S. Pursuing a Prospective Perspective. *3*, 77–79.

(98) Lyu, J. et al. AlphaFold2 structures guide prospective ligand discovery. *384*, eadn6354, Publisher: American Association for the Advancement of Science.

(99) Laufkötter, O.; Hu, H.; Miljković, F.; Bajorath, J. Structure- and Similarity-Based Survey of Allosteric Kinase Inhibitors, Activators, and Closely Related Compounds. *65*, 922–934, Publisher: American Chemical Society.

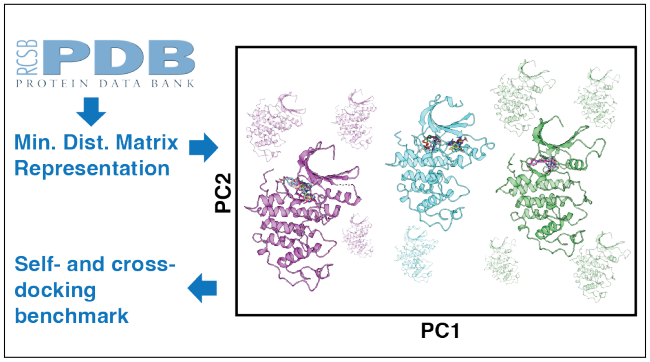
(100) Xerxa, E.; Laufkötter, O.; Bajorath, J. Systematic Analysis of Covalent and Allosteric Protein Kinase Inhibitors. *28*, 5805, Number: 15 Publisher: Multidisciplinary Digital Publishing Institute.

(101) Qiao, Z.; Nie, W.; Vahdat, A.; Miller, T. F.; Anandkumar, A. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *6*, 195–208, Publisher: Nature Publishing Group.

(102) Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *0*, eadl2528, Publisher: American Association for the Advancement of Science.

(103) Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. 1–3, Publisher: Nature Publishing Group.

TOC Graphic


RCSB PDB
PROTEIN DATA BANK
Min. Dist. Matrix Representation
Self- and cross-docking benchmark
PC2
PC1