ROYAL SOCIETY
OF CHEMISTRY

Check for updates

Assessing the accuracy of binding pose prediction for kinase proteins and 7-azaindole inhibitors: a study with AutoDock4, Vina, DOCK 6, and GNINA 1.0

Abhishek Tripathi, [ID] [a] Kapali Suri,[a] Sriram K.[ab] and N. Arul Murugan [ID] *[a]

Comparative benchmarking of molecular docking tools is vital for assessing the reliability of virtual screening and binding pose prediction in drug discovery. Our study evaluates the performance of four open-source docking programs: AutoDock4, AutoDock Vina, DOCK 6, and GNINA 1.0 in predicting binding poses of 7-azaindole derivative compounds across 70 kinase–ligand complexes from the RCSB Protein Data Bank. These compounds were selected due to the azaindole moiety's ability to mimic adenosine triphosphate (ATP), which is the natural substrate for kinases to form hydrogen bonds in kinase hinge regions. Docking was conducted using both rigid and flexible receptor conditions. Binding pose accuracy was quantified using atom-to-atom root mean square deviation (RMSD) and center-of-mass (CoM) RMSD, with 2 Å set as the success threshold. GNINA 1.0, which incorporates a 3D convolutional neural network (CNN)-based scoring function, achieved the highest atom-to-atom RMSD success rate (85.29%) under rigid docking. DOCK 6 performed second-best by exhibiting 79.71% success in rigid docking, but reduced to 61.19% while using a flexible docking approach. AutoDock Vina demonstrated comparable performance in both docking modes, with a success rate of 62.69% under rigid docking and 60.66% under flexible docking, indicating minimal variation between the two modes compared to the broader differences observed with other docking tools. AutoDock4 demonstrated moderate performance (<50% success) in both rigid and flexible modes. The accuracy of pose prediction varied by inhibitor class. Type 1 and Type 3 inhibitors were predicted with higher fidelity compared to Type 2 inhibitors, due to differences in binding site rigidity, hydrophobic interactions, and DFG-loop dynamics. Redocking analyses also assessed whether docking tools could recover three key interaction features: hinge hydrogen bonding, hydrophobic contacts, and DFG-loop orientations. GNINA 1.0 consistently performed well in recovering these features with lower computational cost. Interestingly, Vina showed better performance in terms of CoM RMSD but lower and consistent for atom-to-atom RMSD. Our study underscores the importance of scoring functions, receptor flexibility, RMSD type, and inhibitor class in determining docking accuracy for kinase-targeted drug design.

1. Introduction

Kinases are an important class of protein enzymes involved in the efficient transfer of a phosphate group from adenosine triphosphate (ATP) to amino acids like tyrosine, serine, and threonine.[1] Kinases regulate key cellular processes such as cell cycle progression and metabolism,[2] and their dysregulation contributes to diseases, making them prime targets for kinase inhibitors. For example, many kinase inhibitors are widely used in cancer treatment,[3] and one such example is the kinase inhibitor that targets CDC7, a serine/threonine kinase that plays a conserved and crucial role in DNA replication and has been recognized as a potential anticancer target.

The protein kinase fold consists of an N-terminal lobe with five beta strands and a C-helix, and a C-terminal lobe with several alpha helices, together forming a central deep pocket that serves as the ATP and ligand-binding active site, as depicted in Fig. 1. Kinase inhibitors are optimized to improve interactions near the hinge region, enhancing potency and selectivity by exploiting novel, target-specific binding cavities. Azaindole derivatives are considered as potential candidates in the development of kinase inhibitors as they share the azaindole moiety with the natural substrate, ATP, for the kinases. Some of the derivatives of azaindole are now being studied in clinical studies and used in the market as kinase inhibitors to

[a]Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Phase 3, New Delhi-110020, India. E-mail: arul.murugan@iiitd.ac.in

[b]Centre for Computational Biology, Indraprastha Institute of Information Technology, Okhla Phase 3, New Delhi-110020, India
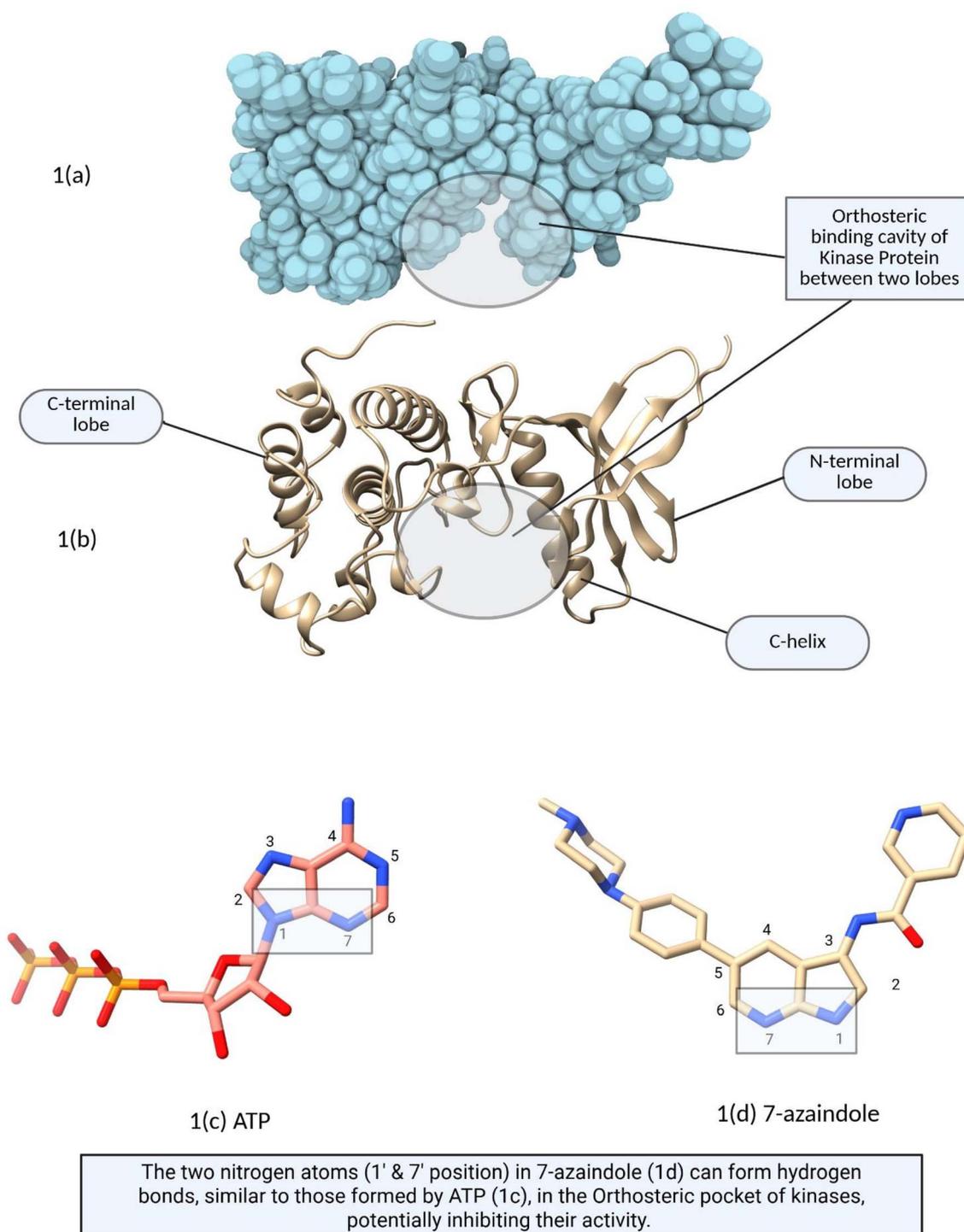
Fig. 1 (a) Orthosteric binding activity of the kinase protein present between two lobes (van der Waals surface representation). (b) Protein kinase contains an N-terminal lobe comprising five beta-sheet strands and one alpha helix known as the C-helix, and a C-terminal lobe consisting of five or six alpha helices. It also shows the orthosteric cavity between both lobes.

treat certain kinase-related conditions. For example, the United States Food and Drug Administration (FDA) has approved vemurafenib, a serine/threonine kinase inhibitor, to treat BRAF-mutated metastatic melanoma.[4]

Azaindole-based kinase inhibitors discovered thus far interact with the adenine binding region, ribose binding region, phosphate binding region, two hydrophobic regions, and many of their cocrystal structures are added to the PDB. Latent back pockets only generated in an inactive conformation are the regions where ATP-competitive inhibitors bind to kinases. In this study, we have performed molecular docking for kinases from a dataset of kinase azaindole complexes. Thirty-seven

Table 1 Different types of 7-azaindole inhibitors specifying their binding pattern and residue orientation

Type	Count	DFG orientation	αC-Helix orientation	7-Azaindole binding mode	Gatekeeper residue involved	Hydrogen bonding pattern
Type 1	43	In	In	Mostly normal	GK+1, GK+3	Normal: GK+1 (acceptor), GK+3 (donor)
Type 2	12	Out	In/out	Mostly flipped	GK+1, GK+3	Flipped: GK+3 (donor & acceptor)
Type 3	15	In (usually)	Out	Non-hinge	Alternate hinge-binding site	H-Bonding outside the hinge region
Type 4	—	Variable	Variable	Allosteric (non-azaindole-based)	Not hinge-dependent	Distant from the ATP site; no hinge H-bonds

distinct kinase types from various kinase families are included in these structural data, which can further be broadly classified under 7 types of human kinases. We took the 70 3D X-ray crystallographic structures of protein kinases complexed with the 7-azaindole derivative compounds from the RCSB Protein Data Bank (PDB), which are available in three types: Type 1, Type 2, and Type 3. The Type 4 kinase inhibitor was not present in the PDB database, as shown in Table 1. The aim of this work is to benchmark the ability of four docking software programs given in Table 2 by predicting the binding poses and distinguishing different types of 7-azaindole kinase inhibitors.

Molecular docking approaches have certain aspects like (i) benchmarking sets that highlight their advantages and caveats, which can affect the generalizability of results, (ii) the advancements in consensus methods that integrate multiple algorithms, (iii) fragment-based approaches that break and reassemble it into the receptor site, and (iv) integration of machine learning algorithms. These additions gradually magnify the accuracy of different studies, such as binding pose prediction and binding affinity prediction. Over time and with the continuous improvement of processing power and hardware capability, it is anticipated that structure-based drug design will ultimately achieve the complete potential of this discipline.[9]

The three-dimensional structure of target biological molecules, such as a protein, DNA, or RNA, is required at the initial stage of a molecular docking computation. The structures of all these macromolecules can be easily obtained from the PDB, whose coordinates are obtained by experimental methods such as X-ray, Cryo-EM, or NMR spectroscopy.[10] The binding site location that should be the focus of the docking calculations is often known. Otherwise, based on a geometric approach, we can find those with several tools like CASTp (Tian *et al.*, 2018)[11] and DoGsitescorer (Volkamer *et al.*, 2012).[12] Nonetheless, two docking types are frequently used. The first is when the binding region knowledge is lacking: either an algorithm to estimate the most probable binding sites, *i.e.*, a "blind docking," or when the binding site is known, *i.e.*, "site-specific docking". The former is more computationally expensive because it searches the entire target structure.[13] The interactions within the target binding site with precalculated potential energies are accomplished by grid representations, a common technique in docking calculations. This method accelerates docking runs and mostly discretizes the binding location. Lennard-Jones and electrostatic potentials are calculated at each grid point where the

interactions occur with the ligand based on the empirical scoring function.[14] This way, molecular docking helps in binding modes prediction and binding affinity prediction for a ligand to a target of a known structure. Based on the desired sampling efficiency, various optimization algorithms such as simulated annealing (SA), genetic algorithm (GA), or Lamarckian genetic algorithm (LGA) are used to perform the docking process.

It is essential to note the flexibility of protein and ligand in both rigid and flexible docking, utilized with all optimization algorithms. As the terms suggest, in rigid docking, the protein is treated as a rigid body, and sampling is done over ligand translational, rotational, and torsional degrees of freedom. In contrast, in flexible docking, the ligand remains fully flexible, and selected side chains of protein residues around the orthosteric binding site are also allowed to sample different conformations, providing a more realistic representation of molecular interactions. The rigid and flexible type of docking follows the principle of the lock-key hypothesis and the induced-fit hypothesis, respectively. The former provides a faster approach to screening drug compounds, and the latter has the potential to mimic the real-world interactions between targets and drugs, but it is computationally more expensive.[15] The flexible docking is implemented in the molecular docking software, such as Vina, AutoDock4, while in DOCK 6 and GNINA 1.0, flexible docking is not implemented. It is worth mentioning that the rigid docking in DOCK 6 treats the ligand as a rigid body, and sampling is done over its translational and rotational degrees of freedom only. The flexible docking in DOCK 6 allows complete conformational flexibility of the ligands (including the sampling over the ligand torsional degrees of freedom) and does not include sampling over the side chain conformations of protein residues (which is different when compared to flexible docking implementation in Autodock 4 and Vina). A general pictorial workflow for binding pose prediction is given in Fig. 2.

AutoDock4 calculates grid maps & generates the map manually by choosing the atom types, whereas AutoDock Vina automatically calculates the grid maps and clusters the results transparently to the user (on the fly during the calculation). Scoring functions are vital in molecular docking and in general in structure-based drug design, and an accurate scoring function is required to rank the ligands appropriately and for the correct binding pose prediction. Recent discoveries in structure-based drug design for evaluating protein–ligand interactions proposed a classification of scoring functions with four

Table 2 Overview of molecular docking programs with technical specifications

Software	Open source/commercial	Latest version	Scoring function	CPU/GPU support	Algorithm	Main reference	Download link
Vina	Open source	1.2.7	Empirical scoring (steric, H-bonding, hydrophobic)	CPU and GPU	Iterated local search + BFGS optimization	Trott & Olson, 2010 (ref. 5)	GitHub (https://github.com/ccsb-scripps/AutoDock-Vina/releases)
AutoDock 4	Open source	4.2.6	Physics-based free energy scoring; van der Waals, electrostatics, torsional entropy, desolvation	CPU and GPU	Lamarckian Genetic Algorithm (LGA)	Morris et al., 2009 (ref. 6)	AutoDoc k4 (https://autodock.scripps.edu/download-autodock4/)
DOCK 6	Open source	6.12	Grid-based energy: van der Waals + electrostatic EGrid = EVDW + EES	CPU only	Anchor-and-grow, rigid-body search, or Grid scoring	Allen et al., 2015 (ref. 7)	DOCK 6 (https://dock.compbio.ucsf.edu/DOCK_6/dock6_manual.htm)
GNINA 1.0	Open source	1.0	Vina-based + CNN deep learning scoring for pose prediction and binding affinity	CPU and GPU	Iterated local search + CNN scoring layer	McNutt et al., 2021 (ref. 8)	GitHub (https://github.com/ccsb-scripps/AutoDock-Vina/releases)

significant types: force-field-based or physics-based,[16] empirical or regression-based,[17] knowledge-based or potential of mean force-based,[18] and descriptor-based, machine learning-based, or deep learning-based.[19] As shown in Table 1 above, the scoring functions employed by the molecular docking software belong to physics-based (Autodock4 and DOCK 6), empirical (VINA), and deep learning-based (GNINA 1.0).

Kinase inhibitors exhibit diverse binding behaviors and are commonly classified into three major types based on their binding site and interaction pattern: Type 1 inhibitors bind to the ATP-competitive orthosteric site, Type 2 inhibitors interact with the DFG-out conformation and occupy an adjacent hydrophobic pocket, and Type 3 inhibitors bind to allosteric non-ATP regions. These distinct binding modes introduce challenges in accurately predicting ligand poses using molecular docking tools. The goal of this manuscript is to systematically analyze the performance of four widely used docking programs—AutoDock Vina, AutoDock4, DOCK 6, and GNINA 1.0—when applied to 7-azaindole derivatives bound to kinase targets. The main objective of this study is to assess the ability of each tool to accurately reproduce experimental binding poses under rigid and flexible receptor conditions. As an additional objective, we have also tried to examine whether the software can correctly identify three key features of 7-azaindole–kinase interactions: hydrogen bonding at conserved hinge regions, the pattern and extent of hydrophobic contacts, and the orientation and involvement of the DFG motif in ligand binding. As a future objective, we have investigated how we can determine whether these docking tools can differentiate between Type 1, Type 2, and Type 3 inhibitor binding modes. This is a challenging task on various fronts. These objectives are addressed through systematic redocking of co-crystallized ligand–kinase complexes, RMSD-based pose comparison, and interaction profiling across inhibitor types, with particular attention to binding site flexibility and type-specific structural features. As we are redocking the kinase inhibitors on ATP sites, it is far easier to determine accurately the binding poses of orthosteric sites than the non-orthosteric sites due to a higher number of Type 1 in nature, and also in the training of the docking scoring function. We believe that accurate binding mode prediction of Type-1, orthosteric kinase inhibitors can be made in comparison to Types 2 and 3 kinase inhibitors. To accomplish this, we have given a correlation between the range of hydrophobic interacting residues and the binding accuracy of each 7-azaindole type.

2. Materials and methodology

2.1. Materials

AutoDock4 is a molecular docking tool that uses a grid-based method to rapidly evaluate the binding energy of possible conformations[6] and it employs a physics-based scoring function that defines docking energy as the sum of van der Waals, electrostatic, hydrogen bonding, and solvation energy, and entropic contribution due to torsional degrees of freedom. Vina is an open-source tool that facilitates the design and execution of simple and complex docking simulations, and it employs an empirical scoring function to rank different binding poses of
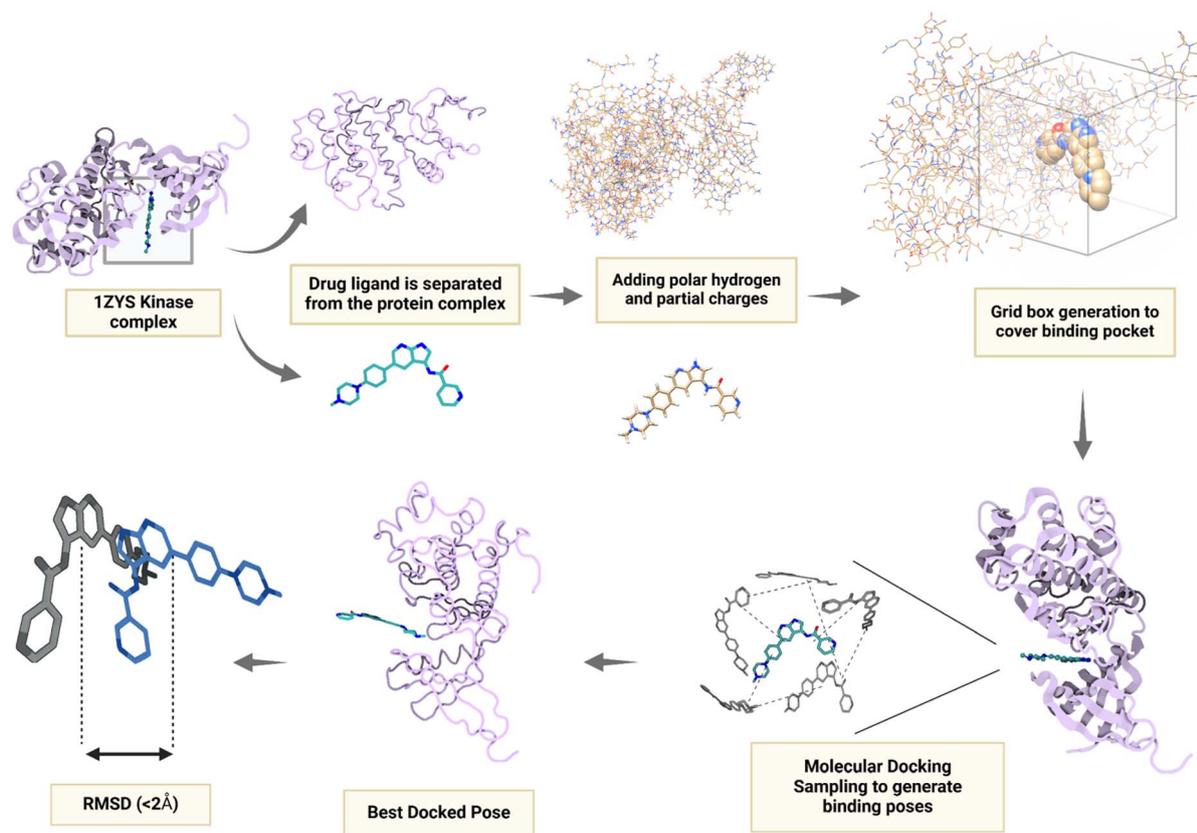
Fig. 2 A general workflow of molecular docking for binding pose prediction.

ligands. Python bindings enable easier scripting for virtual screening and other drug discovery tasks in its newest version.[20] DOCK 6 is a software that uses the principles of structure-based design to facilitate the process of molecular docking, and it also employs a physics-based or force-field-based scoring function. Its utility is proven through its ability to accurately reproduce poses, perform cross-docking, and enhance tests on systems targeted for drug development. DOCK 6 utilizes a graph-matching technique to determine the final grid score and binding pose by generating spheres around the binding site.[7] On the other hand, GNINA is a computational docking software that utilizes a deep learning CNN scoring function to score, generate, and rank ligand poses.[8] A general summary of the various docking software used in this work is provided in Table 2. OpenBabel is a chemical file format conversion that also facilitates a programming library to handle chemical data, and it implements a wide range of chemoinformatics algorithms like partial charge assignment, adding hydrogens, aromaticity detection, etc.[21] For visualizing macromolecules, and small molecules and their interaction, which also offers interactive manipulation of the structures, supporting various formats and computational tasks, we have used ChimeraX.[22]

2.2. Methodology

X-ray co-crystal structures of kinases complexed with 7-azain-dole fragments were retrieved from the RCSB Protein Data Bank

(PDB). A curated dataset of seventy kinase–7-azaindole complexes was selected for binding pose analysis.[23] The PDB IDs of all the complexes, along with ligand represented using 3-letter code and in SMILES notation, are provided in the SI. UCSF Chimera was employed in non-GUI mode *via* Python scripting to merge and superimpose all selected protein complexes onto a common reference structure, ensuring consistent alignment. Docking configuration files were prepared using MGLTools for AutoDock4 and AutoDock Vina, enabling parameterization for both rigid and flexible docking protocols. For DOCK 6, standard parameter files—min.in, rigid.in, and flex.in, provided in the official documentation, were adapted and modified suitably to meet the requirements of rigid and flexible docking runs for our systems of interest. Each kinase complex was separated into receptor and ligand components. For AutoDock4 and Vina, these were converted into the '.pdbqt' format, while DOCK 6 utilized the '.mol2' format. GNINA 1.0 employed the '.pdb' format for both receptor and ligand input. The Open Babel toolkit was used for interconversion and manipulation of chemical file formats.[6] Multiple docking calculations, both rigid and flexible, were conducted using AutoDock4, Vina, and DOCK 6. In contrast, GNINA 1.0 was restricted to rigid docking only. This limitation arises because the default convolutional neural network (CNN) scoring models in GNINA 1.0 are currently trained exclusively on rigid receptor conformations and thus are not optimized for flexible receptor docking scenarios. Table 2

provides an overview of open-source molecular docking programs used in this study, along with their key technical specifications, highlighting their search algorithms, scoring functions, and flexibility features. These optimization algorithms are utilized to explore the conformational space of the ligands and their interactions with target receptor proteins. It is reported that the Lamarckian genetic algorithm can outperform the other methods that AutoDock4 generally uses. On the other hand, Autodock Vina uses a combination of Iterative Local search and BFGS, *i.e.*, Broyden–Fletcher–Goldfarb–Shanno,[5] which is optimized for higher speed and accuracy. The computational time for docking is directly proportional to the size of the docked pose conformations generated. LGA docking simulations should be repeated various times to get the potential binding modes from those docked conformations. Thus, multiple conformations may exist based on the clustering by their geometrical proximity, and the clustering criterion is based on geometric similarity by considering the RMSD of the heavier ligand atoms. DOCK 6 uses the 'anchor and grow' search algorithm to find the best binding pose within the conformational space. The CNN scoring function is supported as a crucial component of the docking workflow by the GNINA 1.0 molecular docking tool, which is a fork of smina[24] and AutoDock Vina. It has the capacity to score and rank protein–ligand complex binding poses accurately, as compared to Vina.[8] Flexible residues for each receptor were selected based on a 5 Å distance cutoff from the ligand's mean coordinates (center of mass). To ensure consistency across the dataset, the ligand with the greatest number of atoms among all seventy complexes was used as a reference to identify the maximum set of flexible residues, which was applied uniformly in flexible docking for both AutoDock4 and Vina. Flexible docking was significantly more computationally intensive, requiring approximately 2.5 times longer to complete relative to rigid docking.

Post-docking, generated poses were parsed and analyzed using Python scripts across all methods. Root mean square deviation (RMSD) calculations were performed to quantitatively compare predicted ligand poses against crystallographic references. Both atom-to-atom heavy-atom RMSD and center-of-mass RMSD metrics were employed, following standard definitions (see Fig. 3). RMSD calculation determines the assessment of the docking tools in reproducing the crystallographic pose of the bound ligand. Typically, a 2 Å cutoff threshold was applied to assess docking accuracy, consistent with established benchmarks.[5]

$$\text{(a)} \quad \text{RMSD}(A, B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{a}_i - \mathbf{b}_i\|^2}$$

$$\text{(b)} \quad \text{RMSD}_{\text{COM}}(A,B) = \|\mathbf{R}_{\text{COM}}^A - \mathbf{R}_{\text{COM}}^B\|$$

$$\mathbf{R}_{\text{COM}} = \frac{1}{M}\sum_{i=1}^{N} m_i \mathbf{r}_i$$

Fig. 3 Standard equations used to calculate root mean square displacement in ligands after docking.

In part 3(a), 'A' and 'B' are the two sets of ligand coordinates from the experiment and predicted from the molecular docking software, respectively. 'N' is the number of atoms in each set of ligands. a_i is the ith atom in set A, and b_i is the ith atom in ligand B. In part 3(b), 'm' refers to the mass of a particular atom present in the ligand, whereas he vector $\mathbf{r}_i$ represents the Cartesian coordinates (x, y, z) of atom i in a molecular system. Here, also in the results section, 2 Å was taken as the cut-off criterion for benchmarking the accuracy of these docking tools. RMSD data were binned in 2 Å intervals for distribution analysis; the first bin included poses with RMSD ≤ 2 Å, the second bin ranged between >2 Å and ≤4 Å, and so forth. Approximately twenty low-energy binding poses were generated per docking run. Final analyses included RMSD distribution plots for both the pose with the minimum RMSD and the pose with the best predicted binding affinity or score, aggregated across docking types and tools. A comprehensive master plot summarizes comparative performance for all docking protocols.

3. Results

Docking was automated in Python on a Linux server, generating 20 poses for each tool, which were then split for visualization and RMSD analysis against the original ligand to assess binding mode fluctuations. In two ways, we have calculated the RMSD. The first method is the 'atom-to-atom rmsd', which is a conventional way to calculate the distance between atoms of the X-ray crystallized pose (initial pose) and the generated docked pose. The second method is 'center-of-mass rmsd', where instead of comparing the distance between each atom, the distance between the overall center of mass (CoM) positions of molecular fragments before (which corresponds to the experimental binding pose) and after docking is calculated. RMSD is calculated between two sets of atomic coordinates by comparing the distances between corresponding atoms using a direct mathematical formula (see Fig. 3). In this study, RMSD distribution plots are presented based on two distinct criteria. The first, termed *Minimum RMSD*, represents the lowest RMSD value among all generated docked poses. The second, referred to as *Best Pose RMSD*, corresponds to the RMSD of the pose with the most favorable binding score, defined as the most negative free binding energy (in kcal mol^{-1}) for Vina, AutoDock4, and GNINA 1.0, and the lowest grid score for DOCK 6.

The following section consists of AutoDock4, Vina, DOCK 6, and GNINA 1.0 distribution plots as count *vs.* RMSD. The count of kinase complexes with 7-azaindole fragments *vs.* root mean square deviation, divided into bins of size 2.

3.1. Atom-to-atom RMSD

Fig. 4 represents a comparative analysis of atom-to-atom minimum RMSD distributions across all four docking tools used in this study, including Vina, AutoDock4, DOCK 6, and GNINA 1.0, under both rigid and flexible receptor conditions. The RMSD values are binned, and the proportion of poses within 2.0 Å, a commonly accepted threshold for accurate pose

prediction, is annotated for each method. Notably, in this atom-to-atom method, GNINA 1.0 (86.76%) and DOCK 6 (79.71%) demonstrate the highest success rates, suggesting the best pose

accuracy. These insights underscore the impact of receptor flexibility and scoring functions on docking accuracy, like the empirical scoring function, which calculates the fitness of
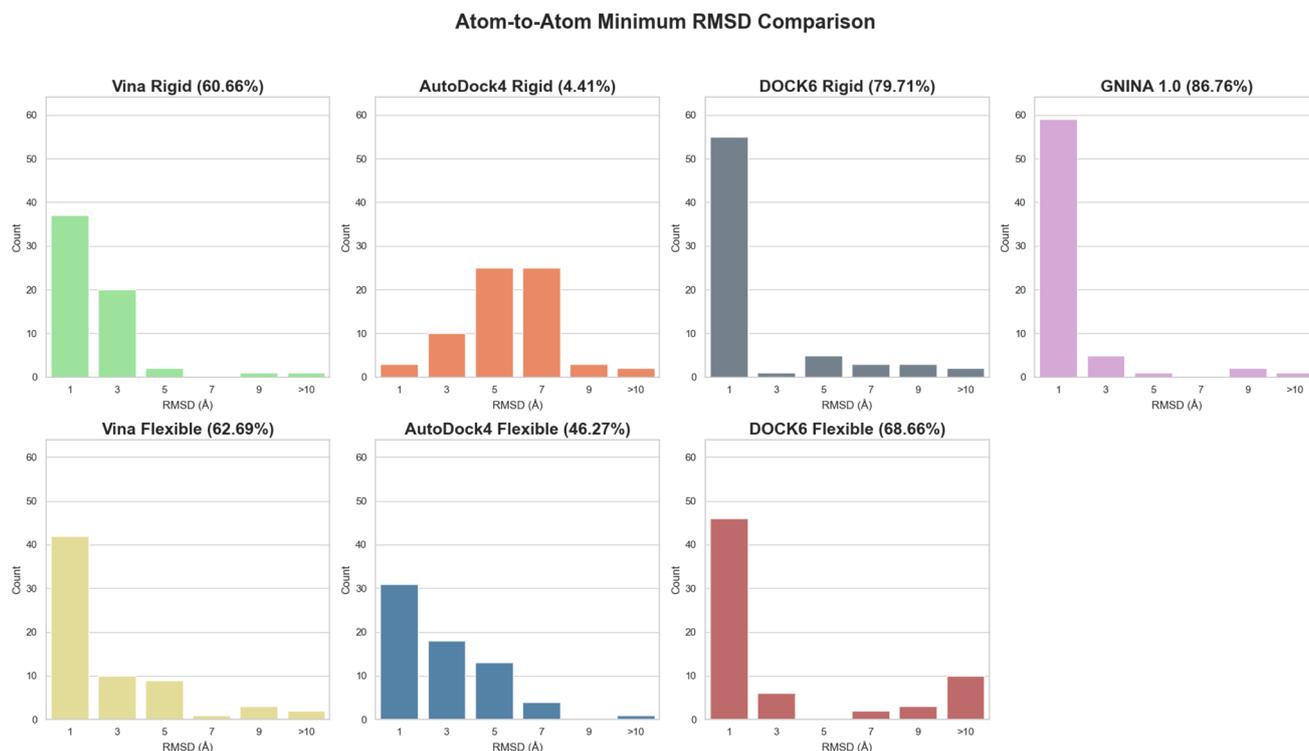

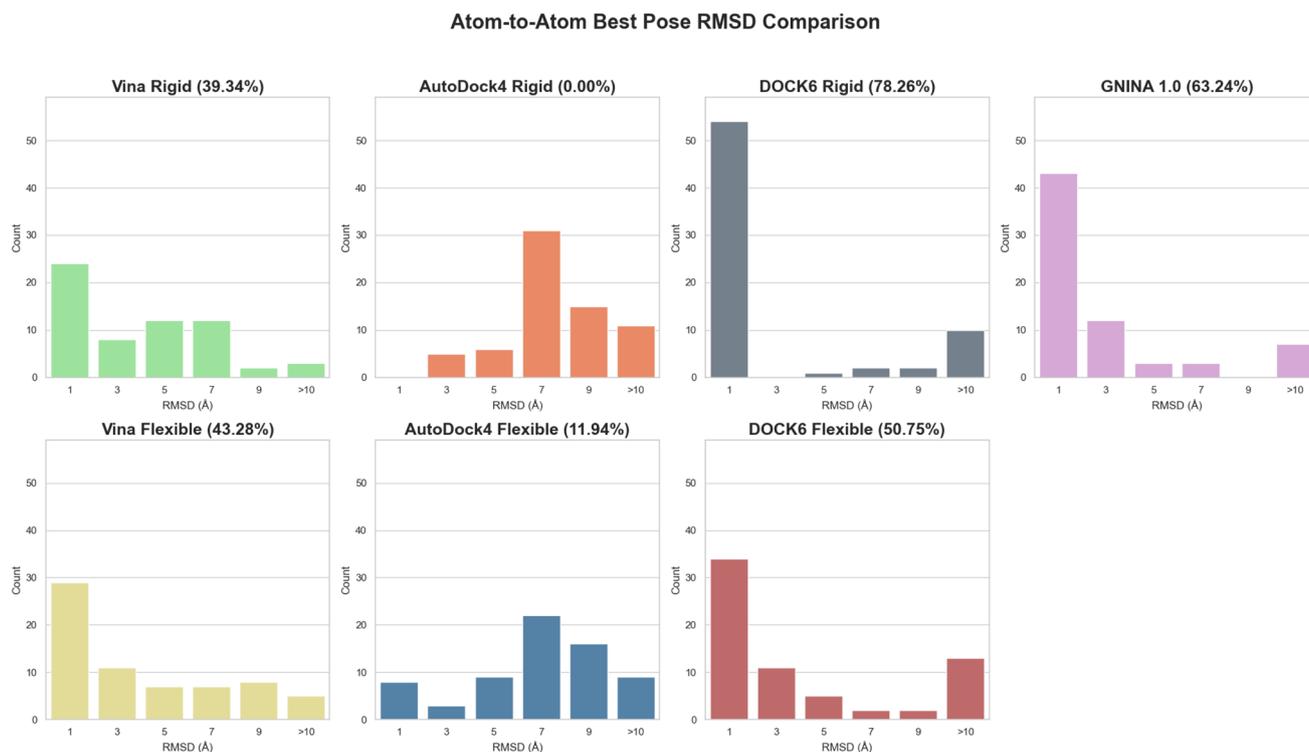**Fig. 4** Comparison of atom-to-atom minimum RMSD among all docking tools in rigid and flexible modes.

Fig. 5 Comparison of atom-to-atom best pose RMSD among all docking tools in rigid and flexible modes.

binding between protein and ligand by adding the contribution from each energetic factor of the protein–ligand binding. It is giving better results for the Minimum RMSD. GNINA 1.0 (63.24%) and DOCK 6 Rigid (78.26%) demonstrate the highest success rates. In contrast, AutoDock4 rigid fails to generate any poses within this threshold, highlighting potential limitations when extracting best pose conditions (Fig. 5).

A master plot is provided below as (Fig. 6), in which one can notice the distinction between the outcomes of all tools and types of docking for 7-azaindole fragments containing Kinase complexes. Among all the cases, GNINA 1.0 outperforms all other cases, having almost ~87% accuracy. In most cases, flexible docking in AutoDock4, Vina, and DOCK 6 shows better or comparable accuracy than rigid docking. One may prefer flexible docking over rigid approaches when higher accuracy is desired and computational expense is not a limiting factor.

3.2. Centre-of-mass RMSD

In this section, all the plots for the center-of-mass RMSD calculation for the four docking tools are given below in Fig. 7 and 8.

It is also mentioned in the discussion in detail that the center-of-mass RMSD has better accuracy than the conventional atom-to-atom type. CoM RMSD concentrates on the translational movement of the ligand as a whole, whereas atom-to-atom RMSD records intricate structural aberrations such as torsions and bond angles. This aids in determining whether, despite maintaining its internal conformation, the ligand has

undergone a considerable change within the binding pocket. Fig. 9 compiles the outcomes of all docking tools into two types: rigid and flexible. In this CoM RMSD analysis, Vina (96.72%) is giving better results than all other tools in both types of docking in minimum RMSD, but in best pose RMSD, GNINA 1.0 (83.82%) is giving better accuracy than all other tools.

3.3. Factors affecting docking accuracy for different types of kinase inhibitors

Generally, we can classify the kinase inhibitors based on their position in the kinase protein and how they interact with different residues like DFG loops and αC-helix (Fig. 1b). Our dataset ligands consistently bind at the conserved hinge region, supported by mixed binding modes (Fig. 10) & LigPlot+ interactions (Fig. 11), which in detail will be discussed in the later section. An analysis of accurate docking poses (RMSD < 2 Å) for all docking types involving rigid protein residues and fully flexible ligands, free to undergo torsional and rotational movements, is presented in Fig. 12. It reveals distinct trends in the performance of scoring functions across kinase inhibitor classes. GNINA demonstrates a marked preference for Type 1 inhibitors, which bind at the ATP site. This trend is likely due to the model's training on datasets such as PDBbind, which are often enriched with orthosteric kinase–ligand complexes. As a result, GNINA's convolutional neural network may be inherently biased toward recognizing canonical ATP-binding poses. GNINA performs better than AutoDock Vina or physics or empirical-based methods because it uses a 3D CNN that
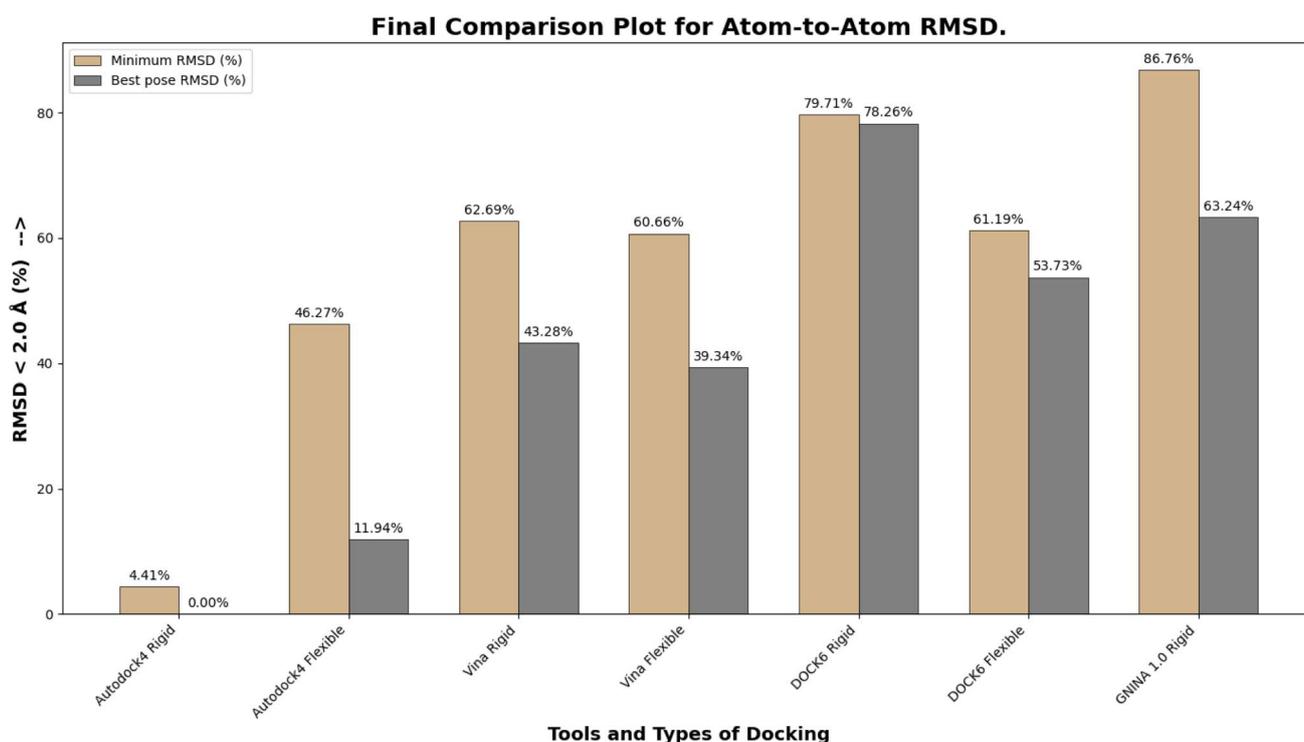

Fig. 6 A comprehensive visual representation of rigid and flexible docking 'atom-to-atom RMSD' calculation by AutoDock4, Vina, DOCK 6, and GNINA 1.0. Among all the tools and types of docking, GNINA 1.0 remarkably showed the highest accuracy, with 89.71% for minimum rmsd and 82.35% for the best pose rmsd.
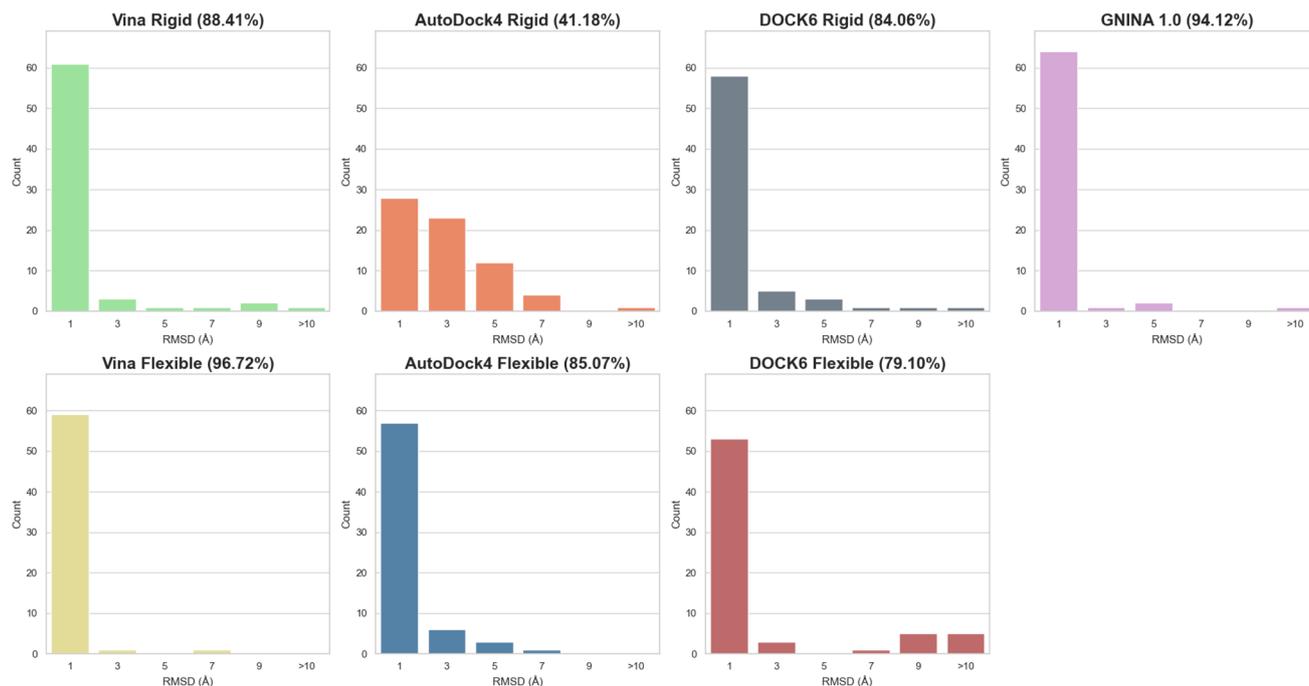
Centre-of-mass Minimum RMSD Comparison

Fig. 7 Comparison of centre-of-mass minimum RMSD among all docking tools in rigid and flexible mode.

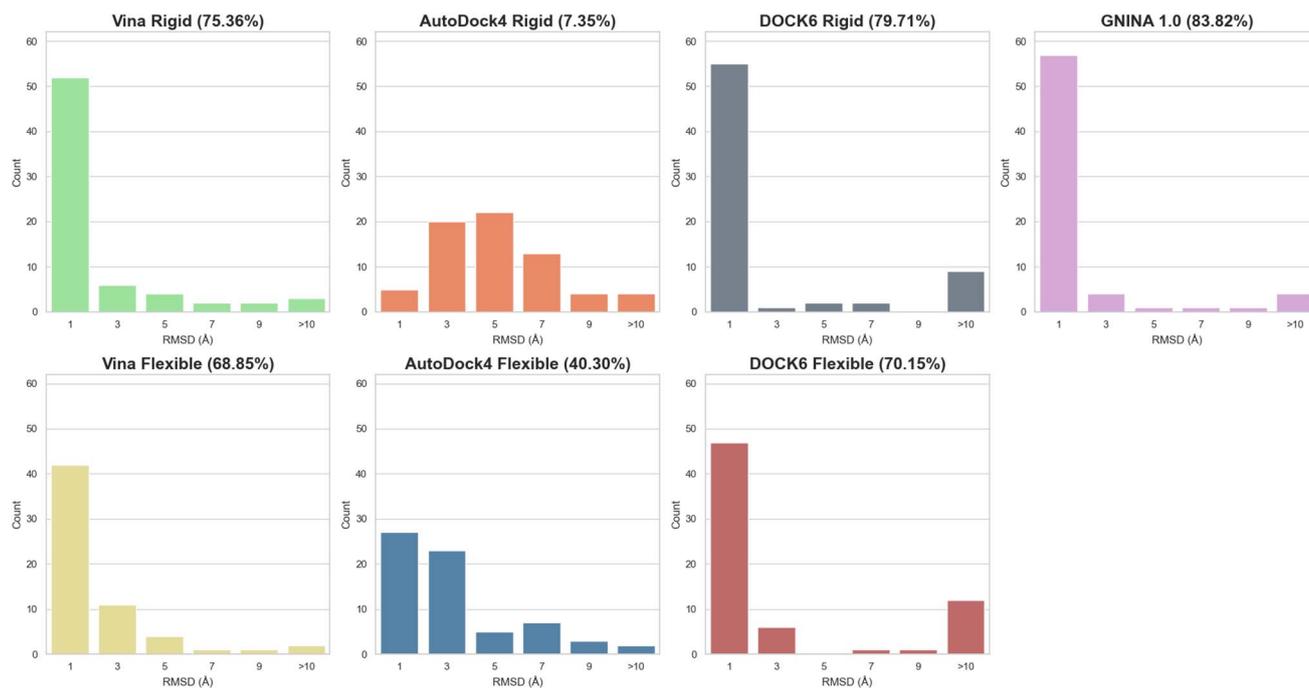## Centre-of-mass Best Pose RMSD Comparison

Fig. 8 Comparison of center-of-mass best pose RMSD among all docking tools in rigid and flexible mode.

learns complex, non-linear features from experimental protein–ligand complexes. GNINA evaluates the candidate binding poses inside a user-defined 3D region (typically a binding pocket), and the CNN-based scoring functions are able to distinguish correct binding poses from incorrect ones, even in flexible or ambiguous binding sites, by capturing
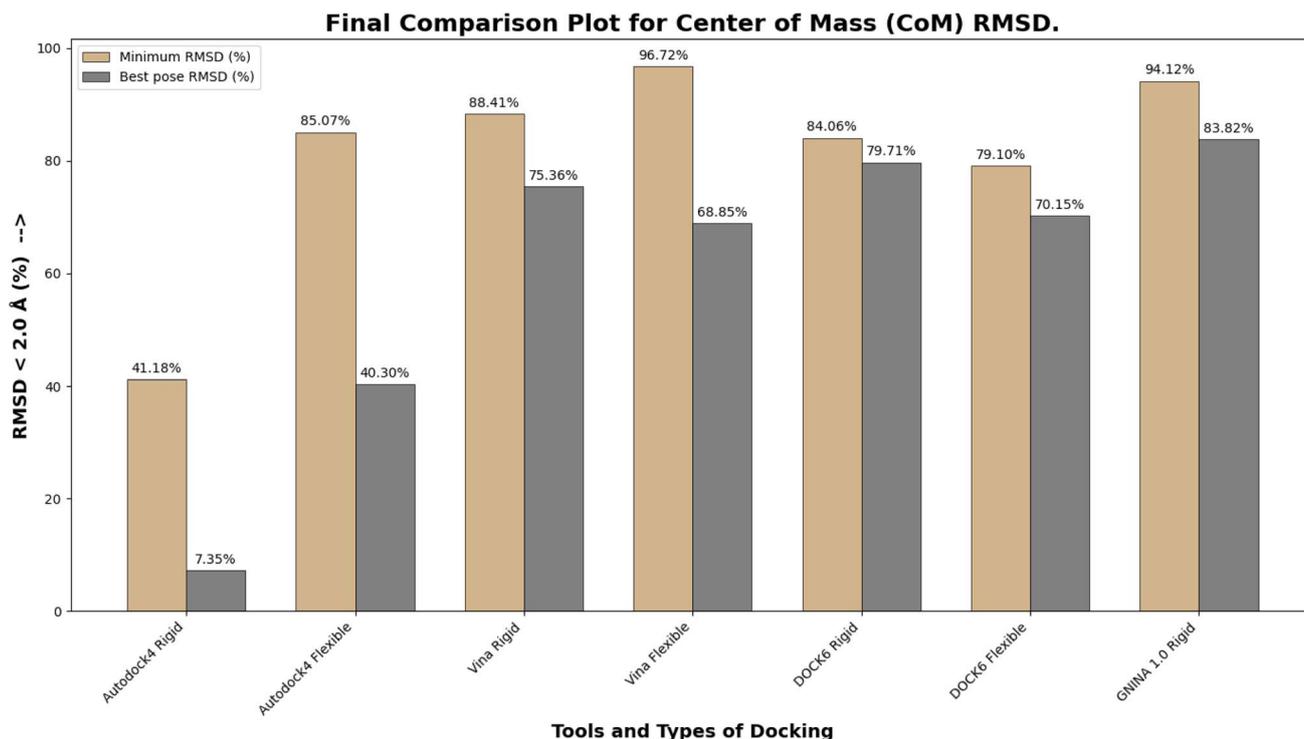
Fig. 9 A comprehensive visual representation of rigid and flexible docking 'center-of-mass RMSD' calculation by AutoDock4, Vina, DOCK 6, and GNINA 1.0. Among all the tools and types of docking, Vina remarkably showed the highest accuracy, with 96.72% for minimum rmsd, and GNINA 1.0 shows 83.82% for the best pose RMSD.

spatial and chemical patterns. This is where the empirical scoring functions may miss. It is about training datasets and also the sampling bias because more Type 1 structures have been used to optimize the scoring function, whereas when success rates are normalized within each type by calculating the percentage of successful cases relative to the total number of cases per type, the differences in accuracy among the types do not show a higher difference, as shown in Fig. 12. There are a lot more Type 1 cases in our dataset, which is about three times more than Type 2 and Type 3. Because of this, the overall accuracy looks higher for Type 1 when percentages are calculated using the total number of successful cases. However, when computed with numbers successfully identified divided by the total number of that particular type, these findings highlight that the performance of scoring functions is not

uniform across ligand classes and is influenced by algorithmic design, training data biases, and sampling strategies. This underscores the importance of docking method selection, or consensus scoring based on the specific binding mode characteristics of the ligand under study.

From Fig. 12, we can observe the lower docking accuracy for Type 2 as compared to other Type 1 and 3 kinase inhibitors, which can be structurally explained by the broader interaction profile of their binding pockets. Specifically, the hydrophobic interaction residues in Type 2 (*e.g.*, 4HVS) span a wider residue range (254 residues: Trp557 to Phe811) compared to Type 1 (133 residues: Leu15 to Asp148) and Type 3 (146 residues: Ile1084 to Tyr1230), as shown in Table 3. This broader span suggests a larger and potentially more flexible binding site, which reduces the likelihood of conserved hydrophobic contacts.

Table 3 Non-covalent interactions like hydrogen bonding and hydrophobic interaction residues in each type of kinase inhibitors

Type	H-Bonding residues	H-Bond distance (Å)	Hydrophobic interaction residues	Hinge/non-hinge
Type 1: 1ZYS	Lys38	3.08 Å	Leu15, Tyr20, Val23, Ala36, Leu84, Tyr86, Gly90, Glu91, Asp94, Leu137, Asp148	Hinge (Glu85 & Cys87)
	Glu85	2.71 Å		
	Cys87	2.95 Å		
Type 2: 4HVS	Glu671	2.93 Å	Trp557, Leu595, Val603, Ala621, Lys623, Glu640, Val654, Thr670, Tyr672, Leu799, Ile808, Cys809, Phe811	Hinge (Glu671 & Cys673)
	Cys673	2.88 Å		
	Asp810	3.25 Å		
Type 3: 4AOI	Pro1158	2.82 Å	Ile1084, Ala1108, Val1092, Leu1140, Leu1157, Asp1164, Asn1167, Arg1208, Asn1209, Met1211, Ala1221, Ala1226, Tyr1230	Hinge (Pro1158 & Met1160)
	Met1160	2.98 Å		
	Asp1222	2.96 Å		

The lack of tightly conserved hydrophobic interactions diminishes the spatial constraints required for accurate ligand placement, making it difficult for both classical and ML-based docking tools to predict correct poses. The rigid receptor approximation used in all four tools further limits accurate pose prediction for Type 2 inhibitors, which require side-chain flexibility (*e.g.*, DFG-out conformations) for correct binding. This induced fit behavior is not accounted for during docking, leading to missed interactions and suboptimal scores.

Additionally, such flexibility is not well accounted for in rigid-receptor docking protocols and may not be sufficiently represented in the training data of machine learning models, both of which contribute to the reduced predictive performance observed for Type 2 inhibitors. These outcomes can also be confirmed by checking the length of H-bonding, which is shortest for Type 1 (2.71 Å) as compared to Type 3 (2.82 Å) and Type 2 (2.88 Å). Although PDB ID 4AOI is classified as a Type 3 inhibitor, it binds at the hinge region as mentioned in Table 3
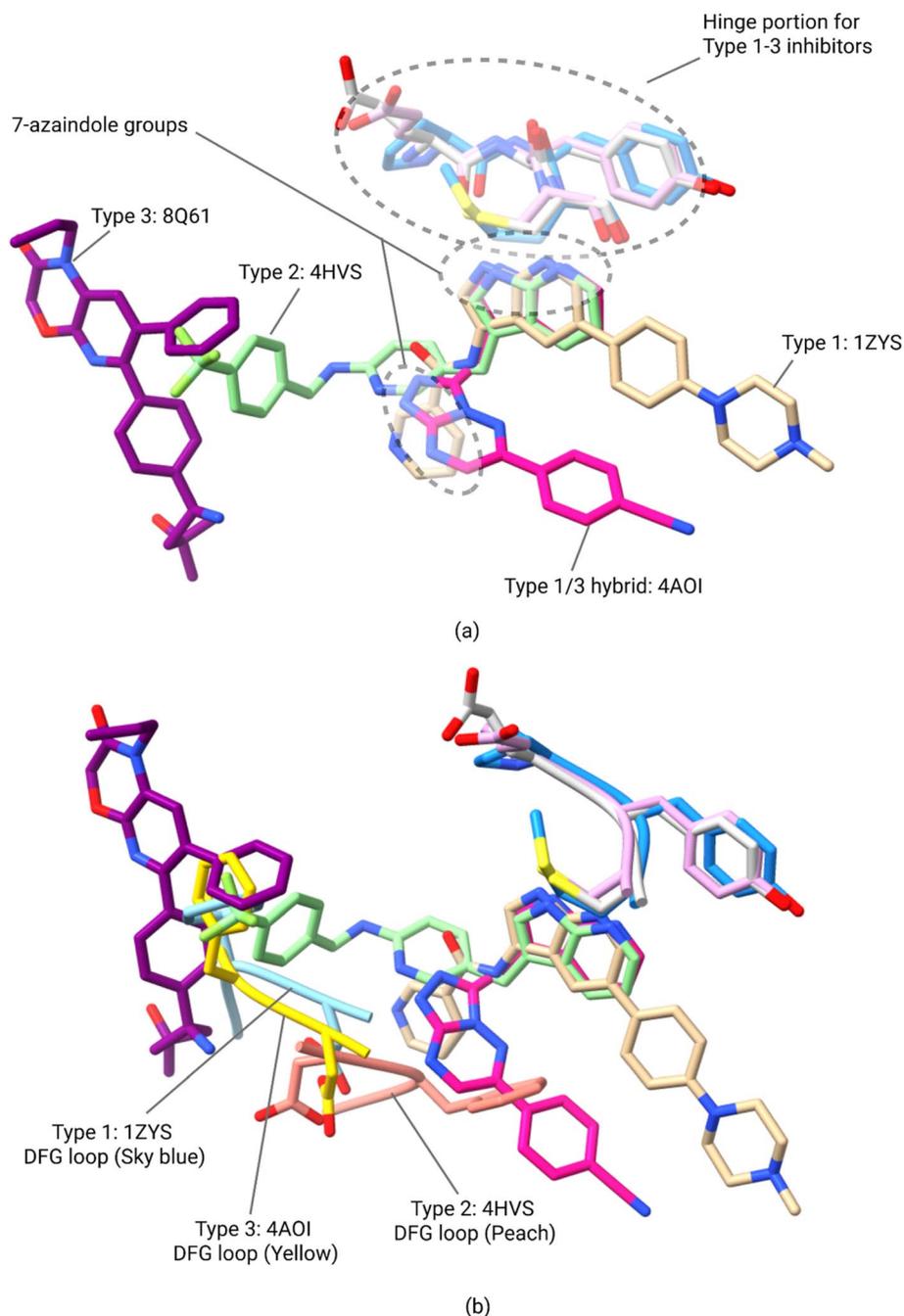

Fig. 10 (a) A detailed analysis of Type 1–3 kinase inhibitors reveals distinctions between hinge-binding and non-hinge-binding modes. Notably, some Type 3 inhibitors exhibit hybrid characteristics by partially occupying the hinge region, as highlighted by the small circled areas, indicating overlap with Type 1 binding features. (b) Types 1 and 3 have DFG-in orientation (upward direction), whereas Type 2 has DFG-out orientation (downward direction).
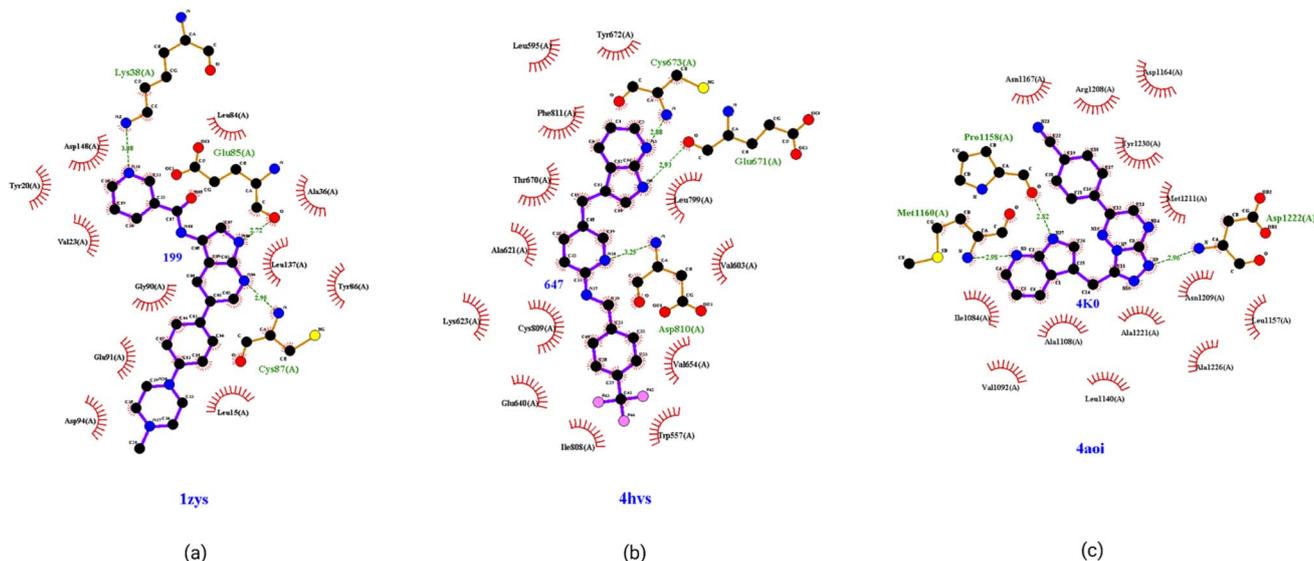
Fig. 11 2D plot of non-covalent interactions showing hydrogen bonding at the hinge region, consistent with interaction patterns reported in the KLIFS database. (a) Type 1 has Glu85 & Cys87 in the hinge region. (b) Type 2 has Glu671 & Cys673 in the hinge region. (c) Type 3 has Pro1158 & Met1160 in the hinge region.

because it represents a hybrid of Type 1 and Type 3 inhibition; in fact, all Type 3 entries in our dataset exhibit such mixed binding modes, which explains their consistent hinge-binding behavior. A pictorial representation of this is given in Fig. 10. Fig. 11 shows the 2D interaction plot generated *via* Ligplot+,[25] which illustrates key non-covalent interactions between the ligand and kinase, notably hydrogen bonds formed at the hinge region. All hydrogen-bonding residues are located within a one-residue gap, confirming their position within the conserved hinge region. This observation aligns with interaction profiles retrieved from the KLIFS database,[26] further confirming the ligand's binding orientation and key contact residues.

4. Discussion

Molecular docking tools can be evaluated on two key parameters: (i) binding affinity predictions, which relate to experimental constants like IC_{50}, K_i, or K_d, and (ii) binding pose predictions, typically assessed *via* RMSD against experimentally solved complex structures. Identifying the accurate binding pose of a drug ligand is also essential for evaluating its binding affinity, and it allows for its use in lead optimization by building a quantitative structure–property relationship.[27] Accurate RMSD has to be calculated for the docked poses to evaluate the proper binding pose and further binding affinity. Several tools and web servers are available to calculate the RMSD between the original pose of the ligand and the generated pose after docking. The simplest way is to calculate atom-by-atom distance using the formula mentioned in the methodology, considering only heavy atoms. Optionally, other tools like DockRMSD can also be used,[28] which claims to calculate RMSD by correcting the ligand symmetry. Prior to the study, we were assuming that Type 1 would give the best accuracy among all the types, as it is the highest in number. But, from Fig. 12, it is clear that even Type 3

is coming with good accuracy; the probable reason behind this could be the hybrid nature of the Type 3 inhibitors in our dataset. Type 2 is showing comparatively less accuracy in both ways: when accuracy is calculated with the total number of successful cases, or the total number of successful cases in a particular type. Binding pose prediction accuracy varied by docking tool and inhibitor class. GNINA 1.0 achieved the best atom-to-atom RMSD accuracy at 89.71% under rigid conditions, outperforming others due to its deep learning-based scoring. AutoDock Vina, with an empirical scoring function, attained 62.69% atom-to-atom RMSD success and 97% CoM RMSD success under flexible docking, though it required higher exhaustiveness for optimal performance. In contrast, Auto-Dock4 showed less than 50% success despite physics-based scoring, while DOCK 6 performed 79.71% in rigid docking mode but lowered to 61.19% in flexible docking. Docking accuracy also depended on inhibitor type. This trend can be structurally rationalized by examining the range and conservation of hydrophobic interaction residues within the binding pockets. Type 1 inhibitors interact with a more compact and conserved residue range (*e.g.*, Leu15–Asp148, 133 residues), whereas Type 2 inhibitors span a broader and more flexible region (Trp557–Phe811, 254 residues), and Type 3 a moderately wider region (Ile1084–Tyr1230, 146 residues). In contrast, Type 1 and Type 3 inhibitors show comparable performance, especially with GNINA and DOCK 6, indicating that well-defined orthosteric and structured allosteric sites support more accurate pose prediction. An interesting point to notice is that 4AOI PDB ID is formally classified as a Type 3 inhibitor, but our analysis (Fig. 10) reveals that it engages residues within the hinge region (Table 3), suggesting a hybrid binding mode incorporating features of both Type 1 and Type 3 inhibition as mentioned above. This behavior is consistently observed across
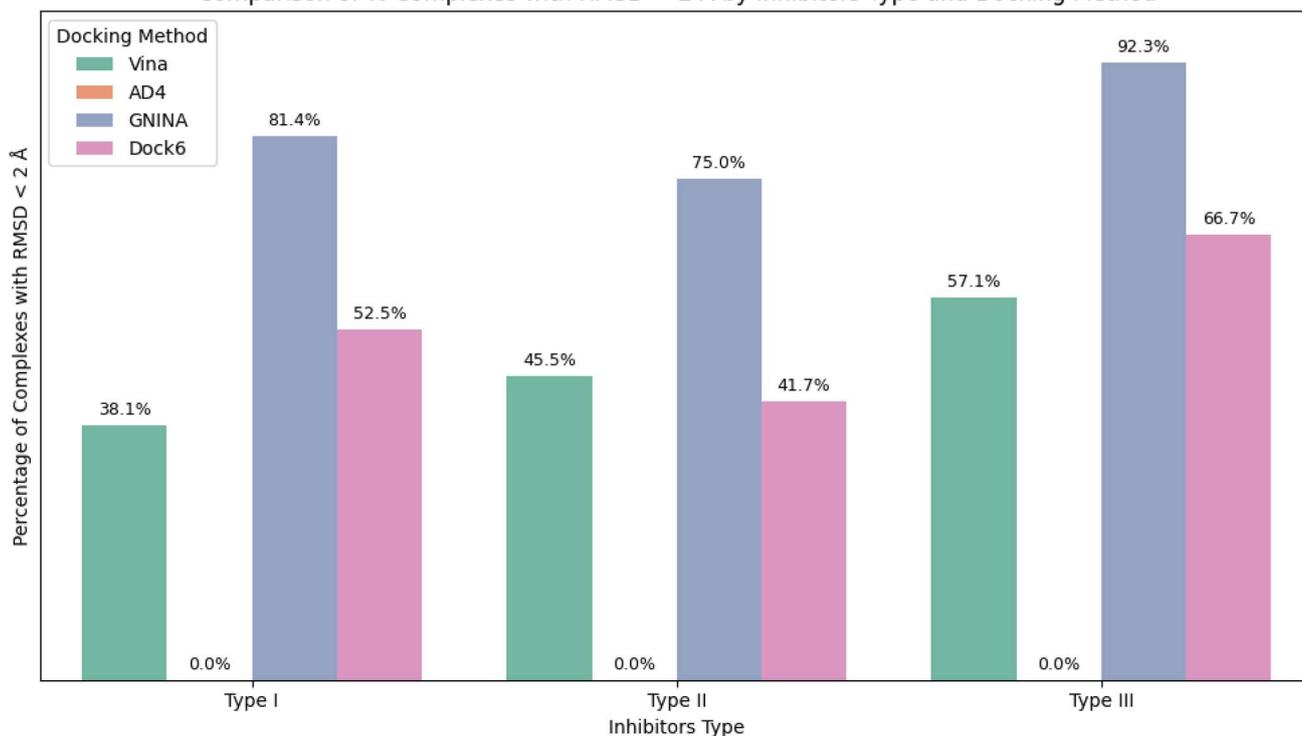
Fig. 12 Correlation between docking accuracy and tool performance across 7-azaindole inhibitor types. Docking accuracy among all types is calculated by: (the number successfully identified/total number of that particular type) × 100.

all Type 3 inhibitors in our dataset, all of which demonstrate hinge-binding characteristics. The 2D interaction plot generated using LigPlot+ (Fig. 11) highlights key non-covalent interactions, notably hydrogen bonds localized within a one-residue span. Such proximity strongly supports binding at the conserved hinge region. This structural observation is further corroborated by interaction data from the KLIFS database, emphasizing the overlapping nature of binding modes in kinase–ligand interactions. While assessing these docking studies, scoring functions come into play significantly to reach any debatable outcome.[29] Vina uses an empirical scoring function or a regression-based class of scoring functions, a weighted sum of energy terms like hydrogen bonding, steric interactions, and hydrophobic interactions.[5] AutoDock4 and DOCK6 employ physics-based scoring functions, incorporating van der Waals, hydrogen bonding, electrostatics, and desolvation, while in Autodock Vina, the weights of these energy interactions were determined by a non-linear fit to experimental binding affinity data.[8] These scoring functions help rank drug ligands, identify potential drugs likely to bind to a target protein, and attain the same pose in the X-ray crystallographic structure. In a study, Wang *et al.* examined the 100 external decoy sets for the assessment of binding pose prediction, which indicates a very high success rate of 87% with a cut-off criterion of RMSD of 2 Å by using the AutoDock4 scoring function.[30]

Vina is built to operate at far higher speeds, and its authors have demonstrated that it surpasses AutoDock4 in terms of accuracy when it comes to re-docking protein–ligand

complexes. Vina could accurately reproduce the reported binding mode within an RMSD of 2 Å for 78% of the 190 protein–ligand complexes.[25] In contrast, AutoDock4 achieved this for just 49% of the complexes. At lower exhaustiveness, the AutoDock4 scoring function performs better, whereas, on the other hand, Vina requires higher exhaustiveness to perform better.[20] Exhaustiveness impacts pose accuracy: GNINA performs well at exhaustiveness 8, balancing accuracy and computation. Increasing this parameter beyond 16 doubles the time with marginal gains. Surprisingly, the rigid docking mode of DOCK 6 outperforms its flexible counterpart, achieving a notable 79.71% success rate in meeting the conventional RMSD cutoff criterion, thereby demonstrating promising predictive accuracy. This can likely be attributed to the fact that flexible-ligand docking often shows lower accuracy than rigid docking because it requires extensive sampling, introduces internal strain, and fails to fully account for conformational entropy loss upon binding. These challenges, combined with higher computational costs, can lead to suboptimal pose prediction compared to rigid docking based on experimentally observed conformers.

In another study by Sunseri & Koes, the default scoring in GNINA 1.0 outperforms the empirical AutoDock Vina scoring function on 89 of the 117 targets of the DUD-E and LIT-PCBA virtual screening benchmarks.[8] Our study aligns with this, suggesting GNINA's best pose prediction accuracy and Vina's high success under flexible sampling. Wang *et al.* performed seven docking protocols with varying features, in which rDock,

DOCK 6-energy, and DOCK 6-contact and AutoDock-Vinardo were rejected due to low accuracy predictions or lack of robustness across different hosts/receptors. AutoDock-Vina and PLANTS were considered usable options for docking. They performed reasonably well in ranking different host–guest pairs with a given receptor/host, with more than 60% host–guest pairs having structural deviations <2 Å (Wang *et al.*, 2024).[31] But in our study, we have extensively covered 4 docking tools with rigid and flexible modes, and also with two types of RMSD analyses (atom-to-atom & centre-of-mass RMSD). In summary, GNINA 1.0 is optimal for accurate, efficient docking of kinase inhibitors, especially for Type 1/3 ligands. Vina performs competitively in CoM RMSD under flexible docking, but it is more computationally demanding. DOCK 6 benefits from flexible docking, while AutoDock4 shows consistent but moderate accuracy. Future docking workflows should consider ligand class, binding site rigidity, and scoring method to optimize predictions in structure-based drug design.

5. Conclusion

Benchmarking docking tools against kinase targets reveals that GNINA 1.0, utilizing a convolutional neural network (CNN)-based deep learning scoring function, achieves superior performance in reproducing experimental binding modes within a 2 Å atom-to-atom RMSD threshold. AutoDock Vina, employing an empirical scoring function, demonstrates optimal results when evaluated with center-of-mass (CoM) RMSD metrics. Both GNINA and Vina thus represent reliable docking options for kinase ligand pose prediction. Scoring function accuracy, whether physics-based or machine learning-derived, critically influences docking outcomes, as insufficient scoring fidelity can lead to deviations from native poses despite high sampling exhaustiveness. The complexity of the ligand, quantified by rotatable bond count and known binding site topology, underscores the importance of exhaustiveness as a parameter for conformational sampling, since increased ligand flexibility expands the conformational search space and computational cost. This study further confirms that no single scoring function universally outperforms others across all protein classes; instead, docking tool selection or consensus scoring strategies should be informed by the specific binding mode characteristics of the ligand. Tailoring docking methodologies to the interaction profile of each ligand enhances the reliability and interpretability of predicted binding poses. This benchmarking study of seventy 7-azaindole kinase complexes yielded approximately 87% accuracy for GNINA, 80% for Dock6, and 63% for Vina using atom-to-atom RMSD criteria. In CoM RMSD evaluations, GNINA and Vina showed the best stability (∼94% and 97%, respectively), corroborating atom-level findings. Consequently, CNN-based GNINA 1.0 has demonstrated high effectiveness as a structure-based docking within kinase protein complexes. Our analysis demonstrates that binding pose prediction accuracy varies significantly across kinase inhibitor types, with Type 1 and 3 inhibitors consistently outperforming Type 2. This disparity correlates with the structural features of their respective binding pockets: Type 1 and Type 3 inhibitors engage more rigid and spatially constrained pockets, characterized by shorter hydrogen bond distances and more conserved hydrophobic interactions, facilitating more predictable binding modes. Conversely, Type 2 inhibitors interact with larger, more flexible binding sites spanning a broader residue range, exhibiting less conserved interactions, which likely contributes to their reduced pose prediction accuracy. A limitation of this study is that the dataset could have been more balanced, as it contained a higher proportion of Type 1 inhibitors compared to Types 2 and 3.

Author contributions

Abhishek Tripathi: data curation, formal analysis, investigation, methodology, visualization, validation, writing – original draft, writing – review and editing. Kapali Suri: formal analysis, investigation, writing – review and editing. Sriram K.: investigation, project administration, supervision, writing – review and editing. Arul Murugan: conceptualization, investigation, project administration, supervision, writing – review and editing.

Conflicts of interest

The authors declare no conflict of interest. The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data availability

All molecular docking input files, receptor and ligand structures, pose prediction outputs, and custom scripts used for pose analysis are available in the supplementary information (SI). The dataset includes kinase–ligand complexes in PDBQT format, docking logs for AutoDock4, Vina, DOCK 6, and GNINA 1.0, and RMSD calculation, extracting flexible residues scripts. No proprietary or confidential data was used. All computational analyses were conducted *in silico* and complied with institutional and international guidelines for ethical research conduct. Supplementary information is available. See DOI: https://doi.org/10.1039/d5ra05526a.

Acknowledgements

References

1 G. Manning, D. B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam, *Science*, 2002, **298**, 1912–1934.
2 K. S. Bhullar, N. O. Lagarón, E. M. McGowan, I. Parmar, A. Jha, B. P. Hubbard and H. P. V. Rupasinghe, *Mol. Cancer*, 2018, **17**, 48.

3 S. Anwar, A. Ahmed, V. Sarli and I. Hassan, *Front. Cell Dev. Biol.*, 2024, **12**, 1413293.

4 A. Kim and M. S. Cohen, *Expert Opin. Drug Discovery*, 2016, **11**, 907–916.

5 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**, 455–461.

6 G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, *J. Comput. Chem.*, 2009, **30**, 2785.

7 W. J. Allen, T. E. Balius, S. Mukherjee, S. R. Brozell, D. T. Moustakas, P. T. Lang, D. A. Case, I. D. Kuntz and R. C. Rizzo, *J. Comput. Chem.*, 2015, **36**, 1132.

8 A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri and D. R. Koes, *J. Cheminform.*, 2021, **13**, 43.

9 P. H. M. Torres, A. C. R. Sodero, P. Jofily and F. P. Silva-Jr, *Int. J. Mol. Sci.*, 2019, **20**, 4574.

10 H. M. Berman, *Nucleic Acids Res.*, 2000, **28**, 235–242.

11 T. Wei, *Nucleic Acids Res.*, 2018, **46**, W363–W367.

12 A. Volkamer, D. Kuhn, F. Rippmann and M. Rarey, *Bioinformatics*, 2012, **28**(15), 2074–2075.

13 C. Hetényi and D. Van Der Spoel, *FEBS Lett.*, 2006, **580**, 1447–1450.

14 B. J. Bender, S. Gahbauer, A. Luttens, J. Lyu, C. M. Webb, R. M. Stein, E. A. Fink, T. E. Balius, J. Carlsson, J. J. Irwin and B. K. Shoichet, *Nat. Protoc.*, 2021, **16**, 4799–4832.

15 M. Mohanty and P. S. Mohanty, *Monatsh. Chem.*, 2023, **154**, 683–707.

16 N. Huang, C. Kalyanaraman, J. J. Irwin and M. P. Jacobson, *J. Chem. Inf. Model.*, 2006, **46**, 243–253.

17 M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee, *J. Comput.-Aided Mol. Des.*, 1997, **11**, 425–445.

18 H. Gohlke, M. Hendlich and G. Klebe, *J. Mol. Biol.*, 2000, **295**, 337–356.

19 J. Liu and R. Wang, *J. Chem. Inf. Model.*, 2015, **55**, 475–482.

20 J. Eberhardt, D. Santos-Martins, A. F. Tillack and S. Forli, *J. Chem. Inf. Model.*, 2021, **61**, 3891–3898.

21 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminform.*, 2011, **3**, 33.

22 E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *J. Comput. Chem.*, 2004, **25**, 1605–1612.

23 T. Irie and M. Sawa, *Chem. Pharm. Bull.*, 2018, **66**, 29–36.

24 D. R. Koes, M. P. Baumgartner and C. J. Camacho, *J. Chem. Inf. Model.*, 2013, **53**, 1893–1904.

25 M. W. Chang, C. Ayeni, S. Breuer and B. E. Torbett, *PLoS One*, 2010, **5**, e11955.

26 A. J. Kooistra, *Nucleic Acids Res.*, 2016, **44**(D1), D365–D371, DOI: 10.1093/nar/gkv1082.

27 K. Wanat, *Mol. Biol. Rep.*, 2020, **47**, 3221–3231.

28 E. W. Bell, *J. Cheminf.*, 2019, **11**(1), 40–40.

29 R. Meli, G. M. Morris and P. C. Biggin, *Front. Bioinform.*, 2022, **2**, 885983.

30 J.-C. Wang, J.-H. Lin, C.-M. Chen, A. L. Perryman and A. J. Olson, *J. Chem. Inf. Model.*, 2011, **51**, 2528–2537.

31 X. Wang, *Liquids*, 2024, **4**(3), 485–504.