



# Combining fragment docking with graph theory to improve ligand docking for homology model structures

Sara Sarfaraz<sup>1</sup> · Iqra Muneer<sup>1</sup> · Haiyan Liu<sup>1</sup>

Received: 30 June 2020 / Accepted: 24 September 2020 / Published online: 9 October 2020  
© Springer Nature Switzerland AG 2020

## Abstract

Computational protein–ligand docking is well-known to be prone to inaccuracies in input receptor structures, and it is challenging to obtain good docking results with computationally predicted receptor structures (e.g. through homology modeling). Here we introduce a fragment-based docking method and test if it reduces requirements on the accuracy of an input receptor structures relative to non-fragment docking approaches. In this method, small rigid fragments are docked first using AutoDock Vina to generate a large number of favorably docked poses spanning the receptor binding pocket. Then a graph theory maximum clique algorithm is applied to find combined sets of docked poses of different fragment types onto which the complete ligand can be properly aligned. On the basis of these alignments, possible binding poses of complete ligand are determined. This docking method is first tested for bound docking on a series of Cytochrome P450 (CYP450) enzyme–substrate complexes, in which experimentally determined receptor structures are used. For all complexes tested, ligand poses of less than 1 Å root mean square deviations (RMSD) from the actual binding positions can be recovered. Then the method is tested for unbound docking with modeled receptor structures for a number of protein–ligand complexes from different families including the very recent severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) protease. For all complexes, poses with RMSD less than 3 Å from actual binding positions can be recovered. Our results suggest that for docking with approximately modeled receptor structures, fragment-based methods can be more effective than common complete ligand docking approaches.

**Keywords** Fragment docking · Homology modeling · Graph theory · Maximum clique · Cytochrome P450-substrate complex

## Introduction

Computational docking is extensively applied to predict protein–ligand binding, both for the prediction and understanding of inter-molecular interactions and for virtual screening

in structure-based drug design [1–3]. Docking calculations require prior knowledge of the three-dimensional structure of the receptor protein. It is not uncommon that a suitable experimental structure of the target receptor is unavailable and one has to manage with a computationally modeled one. Most commonly, the modeled structures are obtained through comparative modeling, i.e., by using existing experimental structures of homologous proteins as templates for model construction [4]. Besides comparative modeling, de novo protein structure prediction is an increasingly promising approach to obtaining modeled receptor structures, given the recent significant progresses making use of deep learning [5]. Despite method advancements, the quality of modeled receptor structures cannot yet match that of high-resolution experimental structures. This is especially true when a homology model has been constructed based on a template protein sharing relatively low sequence identity with the target. In such a situation, the sequence alignment between the

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s10822-020-00345-7>) contains supplementary material, which is available to authorized users.

---

✉ Haiyan Liu  
hyliu@ustc.edu.cn  
Sara Sarfaraz  
sara20@mail.ustc.edu.cn  
Iqra Muneer  
iqra@mail.ustc.edu.cn

<sup>1</sup> School of life sciences, University of Science and Technology of China, Hefei 230026, Anhui, China

target and the template protein contains frequent insertions and deletions, causing the accuracy of the resulting structure model to deteriorate severely [4]. Given the large number of proteins that can potentially serve as drug targets but do not have crystal structures, it is of wide interest to consider computational docking of ligands with modeled receptor structures that are of relatively low accuracy.

Most current docking methods have not been tailored to consider homology models as input receptor structures. Instead, they have usually been optimized by considering benchmarks such as bound docking, in which experimental structures of receptors in complexes with target ligands have been used as input for docking [6–11]. As a result, the performance of docking strongly depends on the input receptor structure. It has been well-known that with usual docking approaches, results obtained by using apo- or modeled receptor structures are significantly worse than results obtained by using high resolution ligand-bound (holo-) experimental receptor structures [12–14]. The high sensitivity of computational docking to input receptor structures is also evidenced by the fact that when docking is applied to screen the same diverse ligand library for the same receptor, different receptor structures yielded different results [15]. This can be explained by the fact that the receptor adopts various conformations in order to establish favorable interactions with different ligands [16], and only one input conformational state (sometimes with local variations such as sidechain rotations) can be considered in one docking run. Because of this strong dependency on receptor structure, computational docking ideally should use input receptor structures coming from high quality experimental studies, preferably determined in complex with a ligand that is similar to the target small molecules that are considered for docking [17]. On the other hand, a docking method that perform well on such inputs does not necessarily lead to the same performance in applications that have to rely on less accurate receptor structures, such as homology models.

One possible way to reduce the above sensitivity to the input structure is to consider the structural flexibility of the receptor. Developments in docking programs have been made in past to allow a certain degree of receptor flexibility by including sidechain variations in e.g. AutoDock [18], GOLD [19], FlexX [11], and RosettaLigand [20]. Although these approaches can help in achieving better results in cases such as docking into the apo-structure of a receptor, they may still not be sufficient for treating modeled receptor structures, which usually contain varied extents of peptide backbone deviations from the actual receptor structure. Ensemble docking is an approach in which backbone variations are included by considering as docking targets multiple receptor conformations generated using MD simulations [21]. However, because of the common insufficiency of using only a small number of conformations to represent a

vast number of possible conformations [22], there is always high risks of missing from the ensemble conformations accommodated to the binding of particular ligands. Besides these limitations, the explicit consideration of receptor flexibility in docking significantly increases the computational costs, which may make the virtual screening of large ligand libraries computationally too expensive. In addition to receptor flexibility, consideration of ligand flexibility is another approach of unbound docking. GalaxyDock3, for example, is a protein–ligand docking approach which considers full ligand conformational flexibility but the protein conformation is fixed at crystal structure [23].

In the current study, we will explore another approach that can potentially reduce the sensitivity to input receptor structure of computational docking. This approach uses fragment docking. We note that fragment docking has emerged over the past decades as a mainstream paradigm in ligand-protein docking and ligand discovery [24–26]. Due to their small sizes, rigid chemical fragments can be used to probe the configurational space inside a binding pocket more efficiently, hence allowing different regions of the pocket to be systematically explored [27, 28]. From the perspective of cheminformatics, the chemical complexity of a fragment is often lower than that of larger compounds, which simplifies the interpretation of structure-activity relationship (SAR) data and may consequently accelerates experimental lead optimization [24, 29]. A number of computational techniques exploiting fragment docking are already available. Examples include software like LUDI [30], GLIDE [10], LigBuilder [31], SEED [32], and web servers like ACFIS [33]. For unbound docking studies, a fragment mapping program, Fsubsite, has been validated on apo-structures of CDK2 [34].

The reasons that fragment docking may better tolerate structure inaccuracies of homology receptor models than complete ligand docking include the following. Consider the docking of a large ligand with a modeled structure of its cognate receptor. For the docking to be successful, every part of the ligand in its correct binding pose need to interact favorably with the modeled receptor structure, at least without any severe steric clashes. If we refer the immediate receptor environment of a part or a fragment of the ligand as a subpocket, any inaccuracy of any subpocket structure that leads to bad ligand-protein interactions (e.g. severe steric clashes) will lead to docking failure. In addition, the covalent connections between different fragments of the ligand translate into stringent constraints on the relative geometries between the subpockets. Structure inaccuracies that lead to errors in these relative geometries also prohibit successful docking. If not the complete ligand but only small fragments comprising the ligand are separately docked, the modeling errors at one subpocket will not affect the proper docking of fragments at other subpockets. In addition, the constraints on

the relative geometries between different subpockets accommodating different fragments may be relaxed, as the docked fragments are not covalently connected. We expect that these properties of fragment docking may be exploited to increase the tolerance of structure inaccuracies when docking a multi-fragments ligand into a modeled receptor structure.

In our proposed approach, after obtaining possible docked poses of fragments, a graph theory algorithm is applied to align a complete ligand with the binding poses of its composing fragments. Graph theory-based methods have been adopted in various biological and bioinformatics studies previously, including metabolic pathway analysis [35, 36], protein flexibility predictions [37], protein side chain predictions [38], secondary structure motif comparisons [39], calcium binding site predictions [40] and more. The application of graph theory in the current work has been inspired by earlier studies using graph theory to find maximum common substructures between two structures [41–44].

Our proposed approach is benchmarked using protein–ligand complexes in which the proteins belong to different families, the cytochrome P450 (CYP450) enzymes, G-protein coupled receptors (GPCR), chaperone, bromodomains and severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2) protease. We have chosen protein families of different sequence length and binding pockets to better understand how our strategy works. The CYP450 enzymes are found in genomes of virtually all organisms. In humans, they are involved in detoxification of various drugs [45]. In plants, they are involved in biosynthesis of various secondary metabolites and thus have numerous applications in industrial biotechnology [46]. The prediction of ligand binding poses using modeled CYP450 proteins is of practical relevance for the prediction of substrate selectivity and reaction site selectivity for CYP450 proteins with unknown structures. The GPCR proteins are important regulatory elements in a wide spectrum of normal and pathological procedures [47]. They are one of the most important families of therapeutic targets for small molecule drugs [48, 49]. Chaperones play a vital role in the stabilization of an unfolded protein [50]. Bromodomain (BRD) is an evolutionary conserved domain with an approximate length of 110 amino acids. Proteins containing BRD regulate the process of gene expression [51]. SARS-CoV-2 is a novel coronavirus that has caused the coronavirus disease 2019 (COVID-19) worldwide [52, 53].

## Materials and methods

Two types of docking were carried out i.e. bound and unbound. In bound docking crystal structures from PDB with bound ligands were used. Whereas in unbound docking, homology models were built and used to dock native

ligands. Scripts and source codes to carry out the calculations can be downloaded from [http://biocomp.ustc.edu.cn/servers/download\\_others.php#scripts-for-small-mol-frag-docking](http://biocomp.ustc.edu.cn/servers/download_others.php#scripts-for-small-mol-frag-docking). Pymol or discovery studio visualizer have been used to carry out visualization [54, 55]. Origin software was used to plot graphs [56].

## Structures of the receptor proteins

In bound docking, the receptor structures have been extracted from the experimentally determined PDB structures of the respective complexes (Table 2). In unbound docking, the receptor structures have been either from a structure complexed with a different ligand or, more interestingly, constructed using the homology modeling program Modeller [57]. Homology models of eleven CYP450 proteins, CYP2B4 (PDB id 1suo, 3me6), CYP2A6 (2fdv), CYP3A4 (PDB id 3ua1, 4d7d), CYP1A1 (PDB id 4i8v), CYP105AS1 (PDB id 4oqr), CYP119 (PDB id 4wqj), CYPBM3 (PDB id 4zf8), CYP126A1 (PDB id 5li8), CYP4B1 (PDB id 6c94), four GPCR proteins, 5-HT1B (PDB id 4iar), 5-HT2B (PDB id 4nc3), GRK5 (PDB id 4wnk), GRK2 (PDB id 5he0), two chaperone proteins, Hsc70/Bag1 (PDB id 3fzm), Hsp 90-alpha (PDB id 6f1n), two bromodomains, namely, human BRD2 (PDB id 4a9m), human ATAD2 (PDB id 5lj0) and one SARS-CoV-2 main protease (PDB id 6wtt) have been constructed. Structure templates for comparative modeling have been identified by sequence BLAST against the PDB database. Details of targets and templates are presented in Table 1. Templates have been chosen on the basis of sequence identity. The protein sequence of a target receptor was downloaded from UniProt database (<https://www.uniprot.org/>) and protein blast available at <https://blast.ncbi.nlm.nih.gov> was run for the given sequence against proteins in PDB excluding the target protein. The purpose for building homology models was to assess our methodology in case of poor models or where it was hard to find a good template. Therefore, templates with sequence identity above 85% were ignored. For 14 out of 20 systems, templates with sequence identity of less than 60% were chosen (Table 1). We also included templates with sequence identity as low as 19% and 25%. Besides sequence identity, query coverage was also taken into account and we made sure that it was more than 50%. For each receptor, twenty initial models were constructed based on a selected template. The initial models were optimized and refined using molecular dynamics with simulated annealing. Then one model with the minimum DOPE (Discrete Optimized Protein Energy) score [58] was selected and used for unbound docking. Further details about model quality are summarized in Supplementary Table S1. Analysis was carried out using the PROCHECK program available at <https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/>.

**Table 1** Target and template protein structures. For the target receptor, the original ligands are given. For each template, overall sequence identity and binding pocket identity with respective target are given

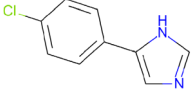
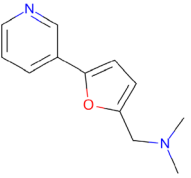
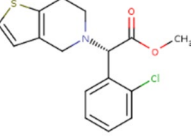
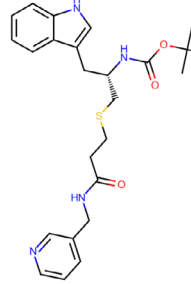
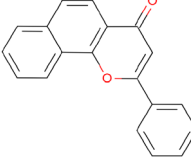
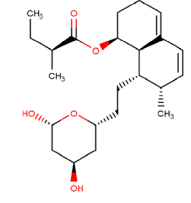
Family	Target Complex PDB ID	Sequence Length	Ligand	Template PDB ID	Sequence Identity	Sequence Identity in pocket
CYP450	1SUO_A	476		3IBD_A	79%	74%
	2FDU_A	476		5E0E_A	57%	62%
	3ME6_A	476		3IBD_A	79 %	74 %
	4D7D_A	487		5X24_A	25%	50%
	4I8V_A	491		3L4D_A	19%	70%
	4OQR_A	457		2Z36_A	45%	73%

Table 1 (continued)

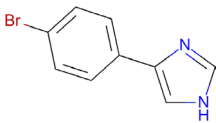
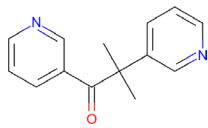
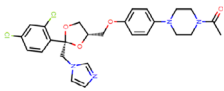
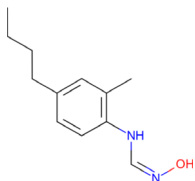
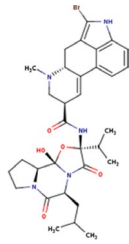
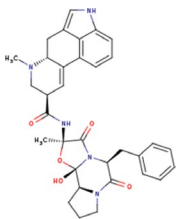
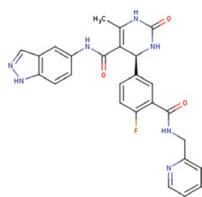
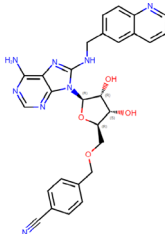
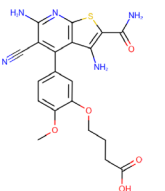
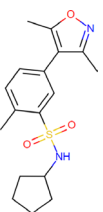
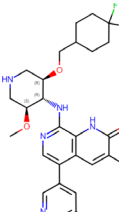
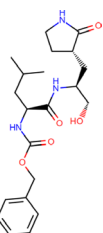
	4WQJ_A	368		1UE8_A	63%	81%
	4ZF8_A	468		6N6Q_A	37 %	63 %
	5LI8_A	414		2WM4_A	37 %	44%
	6C94_A	497		5VEU_A	27%	58%
	3UA1_A	487		5VEU_A	84 %	69 %
GPCR	4IAR_A	401		5WIU_A	52%	45%
	4NC3_A	430		6A93_A	67%	52%
	4WNK_A	598		4PNI_A	47%	55%
	5HE0_A	641		4L9I_A	35%	68%

Table 1 (continued)

Chaperone	3FZM_A	381		3IUC_A	70%	82%
	6F1N_A	236		3PEH_A	52%	58%
Bromodomain	4A9M_A	154		5U2C_A	42 %	100%
	5LJ0_A	130		4YYM_A	32%	85%
SARS-CoV-2	6WTT_A	310		4YLU_A	50%	100%

### Fragment docking

Figure 1 illustrates the overall workflow of the fragment-based docking approach.

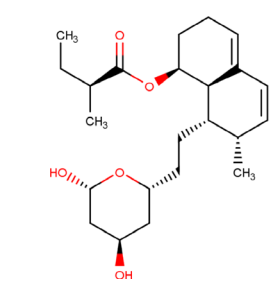
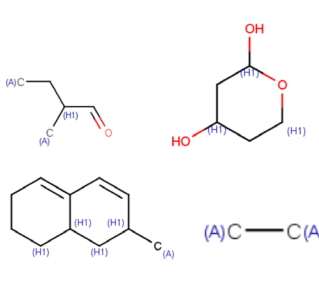
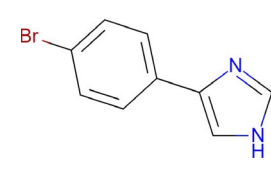
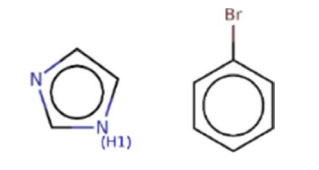
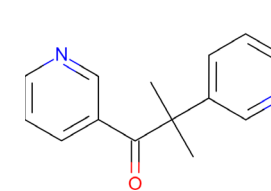
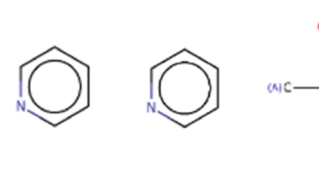
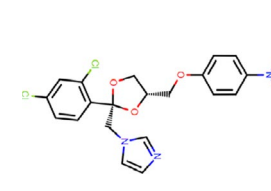
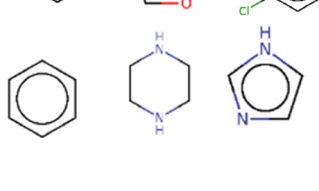
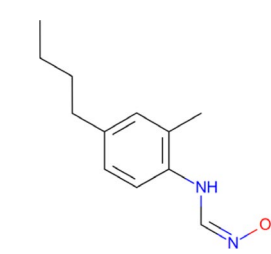
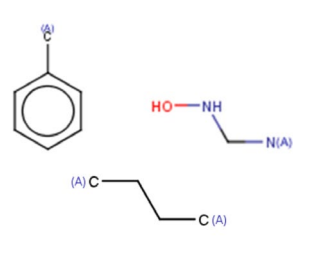
We employed the BRICS (breaking of retrosynthetically interesting chemical substructures) algorithm [59]

as included in the RDKit package [60] to break complete ligands into small, rigid fragments. BRICS shreds the molecules retro-synthetically, avoiding the generation of overlapping fragments. The program also avoids the generation of small terminal fragments such as single hydrogen atoms, hydroxyl, nitro and other small groups. Ligands that were

**Table 2** PDB ids of complexes used for the bound docking tests. The ligands are listed along with corresponding fragments

PDB ID	Ligand ID	Ligand structures	Fragments
1suo (Cyp2B4)	CPZ		
2fdu (Cyp2A6)	D1G		
3me6 (Cyp2B4)	CGE		
4d7d (CYP3A4)	PKT		
4i8v (Cyp1A1)	BHF		

Table 2 (continued)

4oqr (CYP105AS1)	2UO		
4wqj (Cyp119)	36Y		
4zf8 (CypBM3)	MYT		
5li8 (Cyp126A)	KKK		
6c94 (Cyp4B1)	V16		

fragmented are presented in Table 2 (bound docking) and Table 1 (unbound docking).

Fragment and receptor data were preprocessed using respective AutoDock tools [61]. More specifically, complete ligands were extracted from the respective PDB files.

Since AutoDock requires pdbqt format, all the ligand and receptor structures were prepared using graphical user interface program of AutoDock tool [61]. Partial charges were computed using the gasteiger method and resulting molecules were saved as pdbqt format to be used for



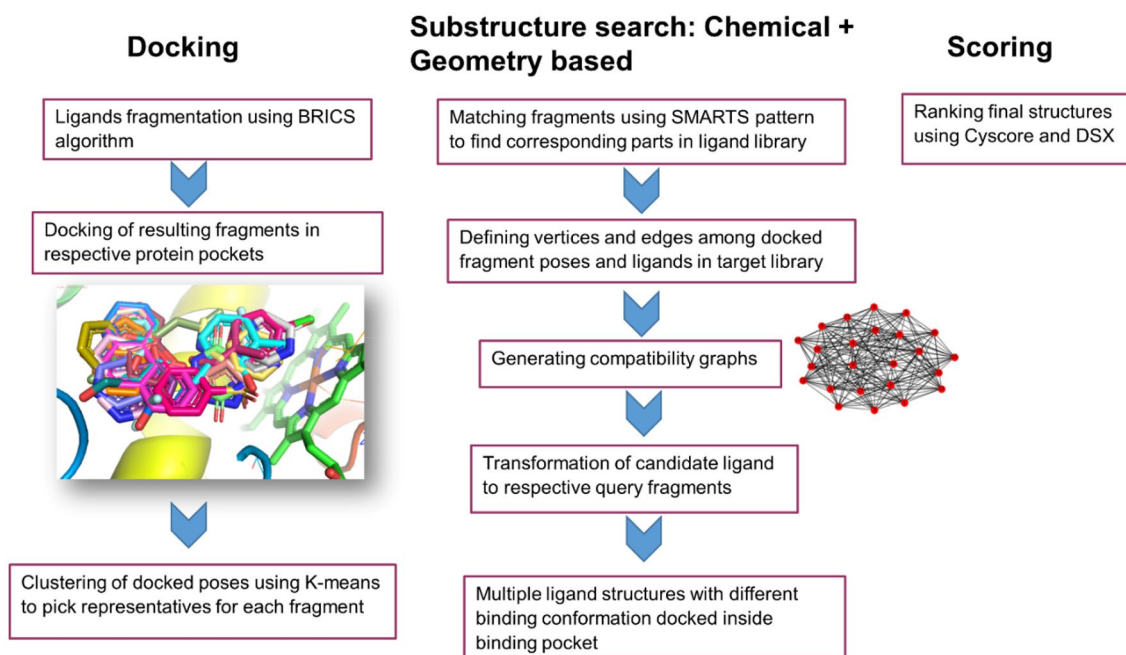


Fig. 1 Workflow of the docking method

docking. Polar hydrogens were added in molecules before charges were computed. In case of fragments, SMARTS for each fragment was saved using MarvinSketch [62] and resulting structures were stored in PDB format. These rigid fragments were then converted to pdbqt format using AutoDock GUI as explained. In case of rigid docking using complete ligands, the receptor molecule was kept rigid whereas ligand was flexible with rotations in rotatable bonds switched on. Cyclic portions of ligands were treated as rigid. The heme cofactor in CYP450 binding pocket was considered as a rigid part of the receptor. For flexible docking, torsions were chosen in the receptor pocket residues which were saved as flexible pdbqt and the rest of the protein was saved as rigid pdbqt. In case of bound docking, the center of the grid for docking was defined based on the information of bound ligand. For unbound docking, pocket residues of the modeled structure were identified and then center coordinates were computed. The same center was used for all fragments of one ligand. The grid size was set to  $15 \times 15 \times 15$  in xyz directions for fragment docking so that fragments can span the entire pocket. All this information was written to a configuration file to be given as an input to AutoDock Vina [18] for carrying out docking. For each configuration 25 docking runs were carried out to obtain up to 500 binding poses.

Poses generated in the previous steps were clustered based on their atomic Cartesian coordinates using k-means clustering [63], with the number of clusters determined using the so-called “elbow” method [64]. After clustering,

the binding pose of the minimum binding score was picked from each cluster to be used further.

### Obtaining docked poses of complete ligands

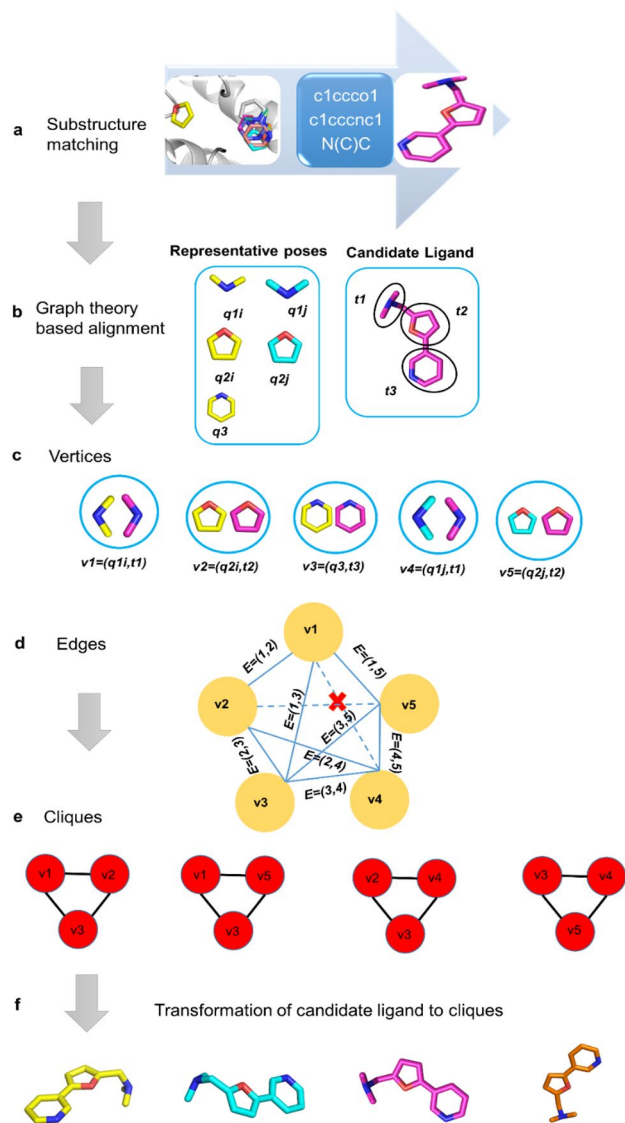
A complete ligand is aligned to combined sets of the docked fragment poses using the following protocol. It makes use of the graph-theory algorithm for determining maximum cliques [65, 66] to identify sets of matches between fragments contained in the ligand and the docked fragment poses (Fig. 2).

To describe the method in graph theory terms, we define two sets, Q and T, respectively,

$$\text{Queryset } Q = \{q_1, q_2, \dots, q_N\} \quad (1)$$

$$\text{Targetset } T = \{t_1, t_2, \dots, t_M\} \quad (2)$$

Each element  $q_i$  of Q corresponds to a docked fragment pose, associated with the corresponding chemical type of the fragment and the atomic positions. For convenience, we refer the chemical type of an element  $q_i$  as  $\Phi(q_i)$ , and the set of atomic positions as  $R(q_i)$ . For each type of fragment, chemically matching substructures in a complete ligand can be identified. We used the RDKit cheminformatics tools to perform the substructure matching [60], by representing the chemical structures of the fragments as SMARTS strings [62] which are matched against a complete ligand. Each element  $t_k$  of the set T in formula (2) corresponds to a part of



**Fig. 2** Determining binding poses of complete ligand through alignments with docked fragment poses using the graph theory maximum clique algorithm. **a** Fragments are given as SMARTS input to find their matching parts in complete ligands. **b** Five fragments of 3 different chemical types are shown. The different 3D positions (docked poses) of the same fragment types are labeled as  $i$  and  $j$  and are colored yellow and blue, respectively. The parts matching the fragment types in a complete ligand are labeled respectively as  $t_1$ ,  $t_2$ ,  $t_3$ . **c** Five vertices of the compatibility graph are shown, each vertex is a pairing of a docked fragment pose and a part of the complete ligand that chemically matches the fragment type. **d** Connectivity between the vertices indicate both chemistry and geometry compatibility of the pairings. Solid lines represent the edges between the vertices, dotted line means there is no edge. **e** Four maximal cliques of size 3 can be found in the compatibility graph, each corresponding to an alignment of the complete ligand to a set of docked fragment poses. **f** Transforming the complete ligand according to each alignment produces a binding pose of the ligand inside the receptor pocket

the ligand that chemically matches a fragment, its chemical type noted as  $\Phi(t_k)$  and the set of atomic positions as  $R(t_k)$ .

An auxiliary graph  $G_c$  representing the compatibility between the elements in the sets  $Q$  and  $T$  is defined as below. Each vertex of  $G_c$ 's corresponds to a combination of two elements associated with the same chemical type, one from  $Q$  and the other from  $T$ , namely,

$$\text{Vertex}(v_{ik}) = (q_i, t_k), q_i \in Q, t_k \in T, \text{ and } \Phi(q_i) = \Phi(t_k) \quad (3)$$

An edge between two vertices  $v_{ik}$  and  $v_{jl}$  is defined in  $G_c$  if and only if the following conditions are met:  $i \neq j, k \neq l$ , and the two sets of atomic positions  $R(q_i) \cup R(q_j)$  and  $R(t_k) \cup R(t_l)$  can be well-aligned against each other. A clique in the graph  $G_c$  is a fully connected (or complete) subgraph in which there is an edge between every vertex in the subgraph to every other vertex in the same subgraph. A maximal clique is a clique which cannot be extended to include new vertices without violating the full connection requirement. The maximal cliques of  $G_c$  can be found with the Bron and Kerbosch algorithm [67]. Each maximal clique corresponds to an alignment between at least two fragments contained in the ligand with the same number of docked fragment poses. The definition of edges in the compatibility graph guarantees that in the alignment, a ligand fragment is only matched to a docked fragment of the same chemical type. In addition, all the inter-fragment relative geometries in the complete ligand are consistent with the relative geometries between the respectively matched docked poses.

In the next step, for the aligned set of fragments defined by each clique, a rigid body geometry transformation is calculated to simultaneously superimpose the different aligned fragments contained in the ligand onto the correspondingly matched docked fragment poses. The resulting transformation is applied to the entire ligand, generating a possible binding pose of the complete ligand inside the binding pocket. To remove potential local steric clashes in the resulting ligand-receptor complex, the complex is subsequently energy minimized by 300 steepest descent steps using UCSF Chimera [68]. All hydrogen atoms were added prior to this minimization step. The minimized structures were scored separately by two scoring functions, namely, Cyscore [69] and DSX [70].

## Results and discussion

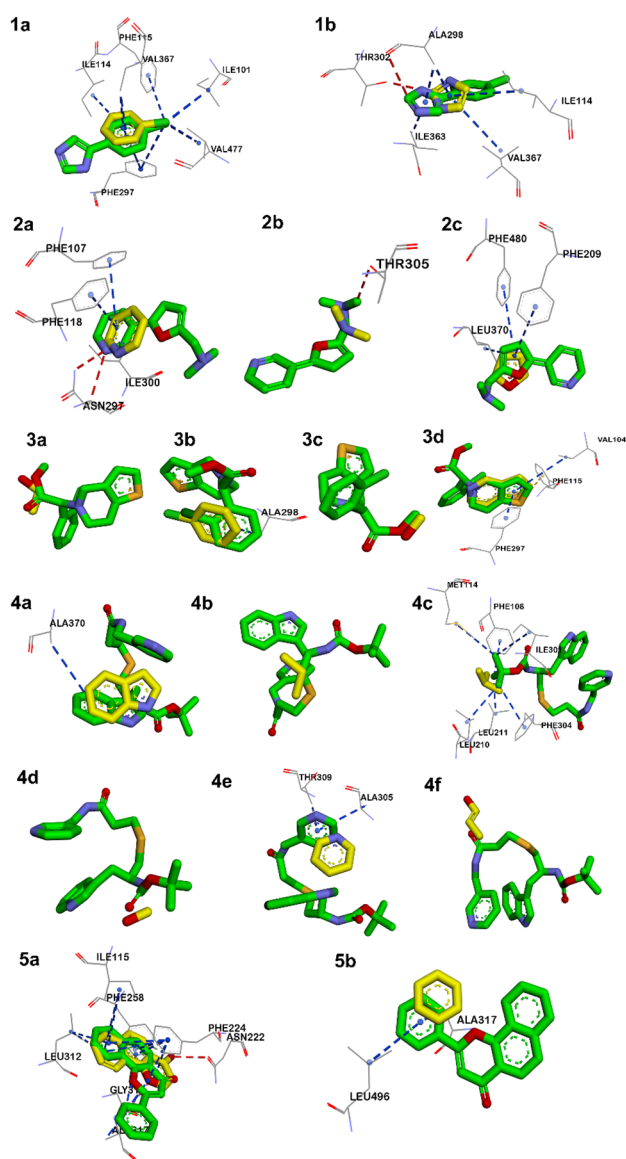
### Binding poses of fragments generated by fragment docking

The PDB ids and ligand structures of the cytochrome P450 protein–ligand complexes used as test dataset for bound docking are given in Table 2. The ligands in these

complexes are of different sizes ranging from 12 to 36 heavy atoms. Fragments generated by applying the BRICS method to each ligand are given in the same table.

Docking the ligand fragments into the binding pocket of the respective receptor structure led to a large number of binding poses for each fragment type. Inspecting the poses suggested that by considering 25 docking runs for each fragment, the positioning of the fragments in the actual ligand-receptor complexes can usually be closely reproduced by one or more docked poses. On the other hand, a large number of docked poses are redundant, i.e., of similar structures. Therefore, following fragment docking, the docked poses for a given fragment were clustered and one representative pose from each cluster was chosen to be used for the subsequent steps. RMSDs of the representative poses from the respective ligand parts in the actual complex structures were computed. For every fragment, the set of representative poses included at least one experimental binding position-like pose. This indicates that because of the small sizes and conformational rigidity of the fragments, it is relatively easy for standard docking protocols to thoroughly explore and identify their favorable binding positions inside a receptor pocket. In Fig. 3, the experimental-like representative poses of fragments are shown together with the containing ligands in experimental binding poses inside the corresponding binding pockets.

We observed that the experimental-like docked poses of fragments do not always correspond to the docked poses associated with the lowest energy scores. Thus if we only retain a small number of top-ranking poses from the docking results without the clustering step to remove redundant poses, we could have missed the near native binding poses from the retained poses. Thus the clustering step minimized the chance of losing native like poses while retaining only a small number of docked poses for subsequent processing. We note that as the fragments are small and can form only reduced number of specific interactions within the binding pockets, a fragment acting as a probe can sample an entire binding pocket more efficiently compared with a complete ligand. Despite this, for a large ligand comprising many different fragments, there can still be chances that for some fragments, the respective near native poses are missed because of the lack of specific interactions for strong binding (for example, in Fig. 3, result 9f, docking of the phenyl fragment generated poses distant from the native binding position of the fragment in the complete ligand). For this reason, in the subsequent graph-theory-based fragment-ligand alignment step, an accepted alignment is required to cover not all but as many as possible the fragments that comprise the ligand. In this way, near native poses of the complete ligand can still be produced making use of the subset of fragments whose near native poses have been generated and retained.



**Fig. 3** Experimental position-like representative fragment poses (yellow) docked inside binding pockets. (1) 1suo, (2) 2fdv, (3) 3me6, (4) 4d7d, (5) 4i8v, (6) 4oqr, (7) 4wqj, (8) 4zf8, (9) 5li8 (10) 6c94. Experimental ligand positions (green) are also shown. Binding pocket residues are labelled and interactions are shown as dotted lines. Red lines represent hydrogen bonds and blue represent hydrophobic interactions

## Binding poses of complete ligands derived from docked fragment poses

### Bound docking

For each of the 10 ligands listed in Table 2, the number of poses generated by applying the graph theory algorithm to align the complete ligands with docked fragment poses are given in Table 3, together with the minimum RMSDs from the respective actual bound conformations in crystal

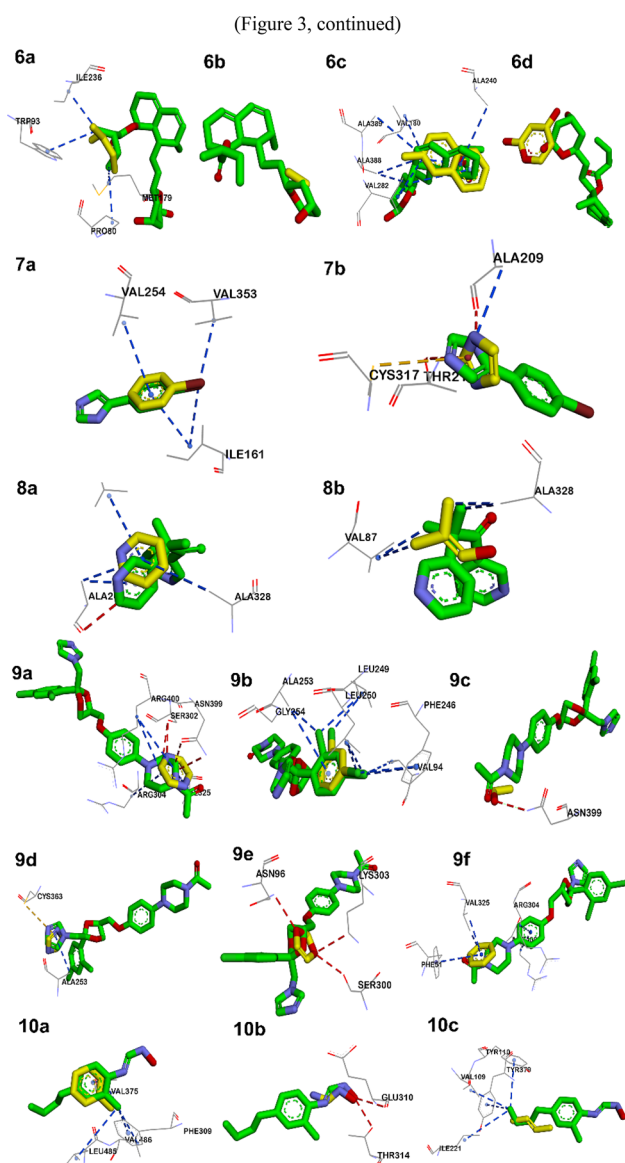


Fig. 3 (continued)

**Table 3** Minimum RMSD values of the bound docking results

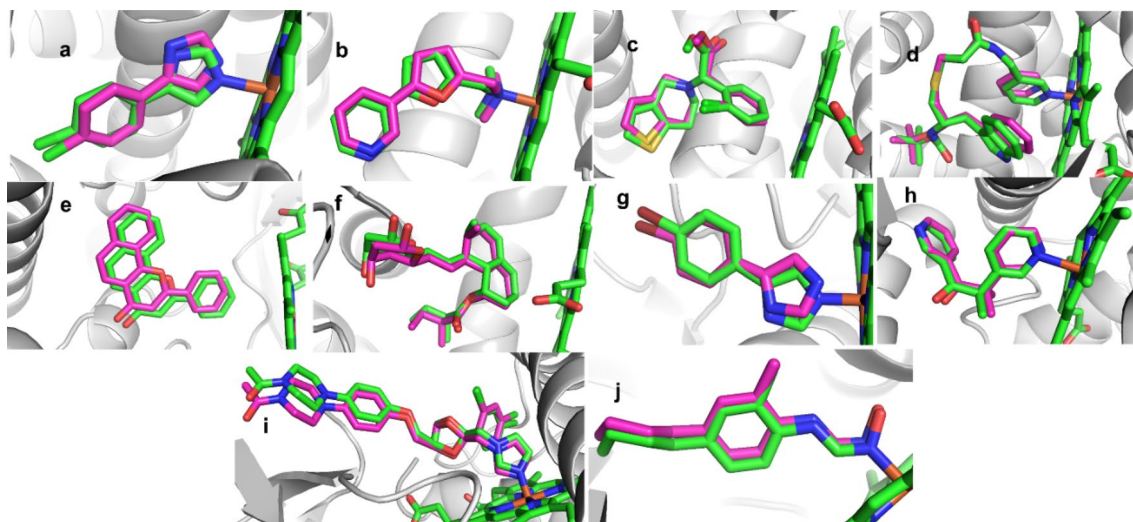
Protein id (ligand id)	Total number of docked poses generated from cliques	Minimum RMSD (Å)
1suo (cpz)	443	0.54
2fdu (d1g)	4971	0.72
3me6 (cge)	254,524	0.46
4d7d (pkt)	2147	0.81
4i8v (bhf)	80	0.8
4oqr (2uo)	125,860	0.79
4wqj (36y)	551	0.59
4zf8 (myt)	18,827	0.55
5li8 (kkk)	1703	0.65
6c94 (v16)	4313	0.42

structures. For some complexes, the total number of poses generated from cliques is quite high due to the large numbers of representative fragment poses chosen in the previous step. The RMSD for every pose from the experimental ligand position was calculated, and low RMSD poses have been found for every complex. Figure 4 shows the superimposed structures of the minimum RMSD docked ligands superimposed with the actual bound ligands inside the binding pockets. These results confirm that the fragment-based docking approach thoroughly searches for possible binding positions of complete ligands inside the binding pockets.

### Unbound docking

For unbound docking, the fragment-based docking approach have been applied to dock the ligands contained in 20 PDB complexes (eleven CYP450-substrate, four GPCR, two chaperons, two bromodomains and one SARS-Cov-2 complex, see Table 1) with receptor structures generated by comparative modeling. The PDB IDs of the template structures as well their sequence identities with respect to the actual receptor proteins are also given in Table 1. Possible binding poses of complete ligands have been generated with the fragment-based approach described above. For simplicity, we have only aligned and transformed each ligand in its actual bound conformation, as present in the respective PDB structures. In real unbound docking, it is usually not difficult to first systematically explore the conformational space of a small molecule ligand to identify its low energy conformational states, and then separately align each conformation with the docked fragment poses by the graph theory algorithm.

The fragment-based docking results are compared with those obtained by direct docking of complete ligands with AutoDock Vina. The latter calculations included rigid receptor docking, in which the receptor structures were not allowed to change, as well as flexible docking, in which the side chains of pocket residues were allowed to rotate. Table 4 shows the minimum RMSDs of docked poses from the experimental bound structures for different docking approaches. The results suggest that for several complexes, while direct docking of the complete ligands was not able to yield any experimental result-like binding poses, the fragment-based approach was able to produce poses that were of relatively small RMSDs from experimental results. Figures 5 and 6 show the docked minimum RMSD binding poses in comparison with respective actual ligand positions in respective experimental complexes. The results obtained by the fragment-based approach and the direct docking approach are compared side by side. Results clearly indicate that the proposed fragment-based strategy outperforms complete ligand docking with AutoDock Vina, with all complexes having RMSD within 3 Å



**Fig. 4** Minimum RMSD poses from bound docking. Green sticks represent experimental ligands and pink sticks represent docked poses

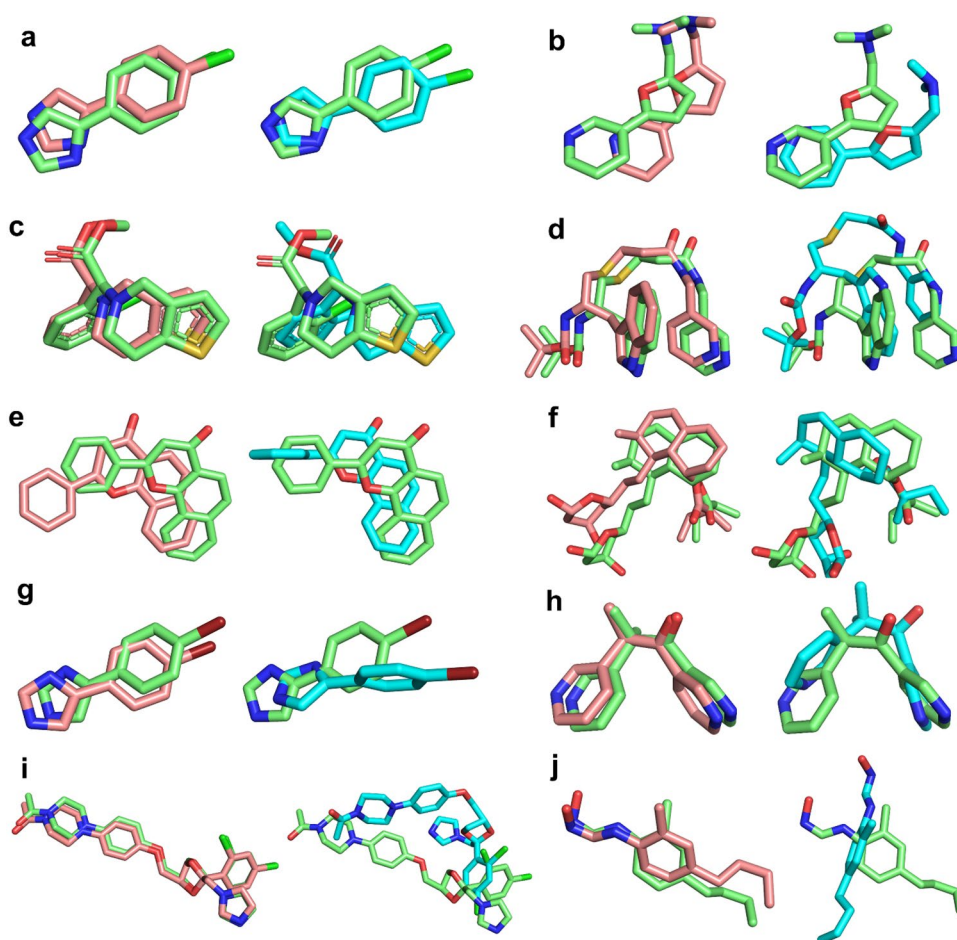
**Table 4** Minimum RMSD values generated by our methodology and AutoDock (rigid, flexible) with modeled structures

Family	Model protein	Minimum RMSD generated by our methodology (Å)	Minimum RMSD generated by AutoDock rigid docking (Å)	Minimum RMSD generated by AutoDock flexible docking(Å)
CYP450	1suo_model	0.7	0.85	1.57
	2fdu_model	2.62	3.97	2.86
	3me6_model	1.33	3.19	2.25
	4d7d_model	0.96	3.17	4.53
	4i8v_model	2.64	2.19	2.5
	4oqr_model	1.75	2.34	2.46
	4wqj_model	1.19	2.69	2.38
	4zf8_model	0.55	1.83	1.7
	5li8_model	0.84	5.68	7.28
	6c94_model	2.38	6.65	4.48
GPCR	3ua1_model	2.25	2.92	3.24
	4iar_model	1.35	7.08	6.23
	4nc3_model	2.22	3.27	3.49
	4wnk_model	1.18	5.4	3.82
Chaperone	5he0_model	0.44	3.1	1.71
	3fzm_model	1.82	6.1	3.33
Bromodomain	6f1n_model	2.13	1.14	3.58
	4a9m_model	2.60	2.64	1.88
SARS-CoV-2	5lj0_model	2.70	1.03	1.91
	6wtt_model	0.25	3.04	4.7

where 12 complexes present RMSD of  $< 2$  Å and 6 complexes having the RMSD of less than 1 Å. In case of direct docking of complete ligands with AutoDock Vina, 12 out of 20 complexes yielded the RMSD of less than 3 Å, 6 complexes within RMSD of 2 Å and only 1 complex with RMSD of less than 1 Å (Table 4). We have found that

complexes for which Autodock gave good results either have small protein sequence length e.g. bromodomains or they are bound to small ligands e.g. in case of 1suo\_A and 4zf8\_A (Table 1). Inaccuracies in the modeled structures prevent appropriate binding of large sized ligands therefore usual docking methods often fail to generate native

**Fig. 5** Docked ligand positions obtained with modeled CYP450 receptors compared with actual ligand positions. **a** 1suo, **b** 2fdu, **c** 3me6, **d** 4d7d, **e** 4i8v, **f** 4oqr, **g** 4wqj, **h** 4zf8, **i** 5li8, **j** 6c94. Ligand binding modes generated by our fragment-based docking approach are colored light pink. Results of direct docking of complete ligands by AutoDock Vina are colored cyan. Actual ligand positions are shown as green sticks



ligand conformations in such cases. Our methodology, on the other hand, performed quite well under circumstances where the protein size is big and they are bound to large ligands e.g. CYP450, GPCR and SARS-CoV-2.

Different modeled receptor structures (4zf8\_model, 3fzm\_model, 1suo\_model) were considered for which results were compared between two different homology models generated via the same respective templates (for the quality of these models see supplementary Table S1). Only slight variations in RMSD ( $<0.5$  Å) were observed (Table 5).

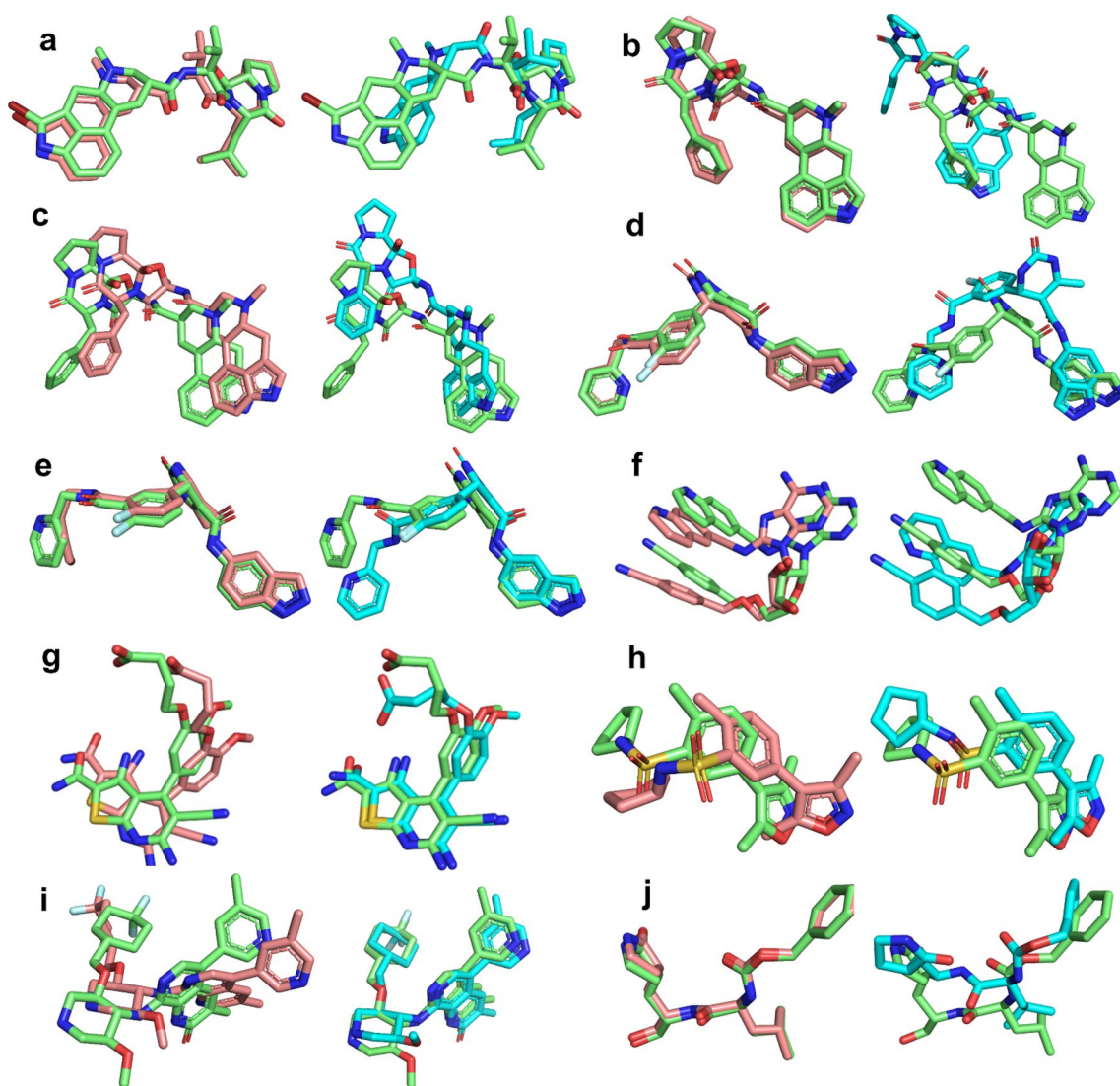
In addition to docking with modeled structures, cross docking analysis was also performed. For this purpose, three protein–ligand complexes were considered. CYP2B4 is present in various conformations in protein data bank bound with different ligands. We took 1suo bound with cpz ligand, 3me6 bound with cge ligand and 3r1a co-crystallized with tb2 ligand. Every ligand was docked with every receptor and RMSD from the native ligand conformation was calculated (Table 6). All the ligands were well aligned with the native conformations with RMSD less than 2 Å except for one ligand tb2 which yielded RMSD of 2.4 Å when cross docked with 1suo receptor. These results clearly indicate that

the proposed methodology works efficiently for unbound docking.

### Scoring the docked poses of complete ligands

Before subjected to scoring, each ligand–receptor complex was first energy-minimized using Chimera [68]. As the transformed ligands may be involved in unacceptable steric clashes with the protein environment, this local optimization step is necessary. Two scoring functions have been considered: Cyscore [69] and DSX [70].

Figure 7 shows the RMSDs from respective experimental ligand positions of the top 10% scored conformations for bound docking. Evaluated with the Cyscore scoring function, for 9 of the 10 complexes, the 10% top scoring binding poses included one or more experimental result-like ones. Evaluated with the DSX scoring function, experimental result-like binding poses were included in the 10% top scoring binding poses for all complexes (Fig. 7b). Because of inaccuracies in both the structural models and the scoring functions, some false binding poses were unavoidably ranked top along with true binding poses. Visual inspection of several such false poses inside binding pockets reveal multiple favorable



**Fig. 6** Docking of modeled proteins. **a** 3ua1, **b** 4iar, **c** 4nc3, **d** 4wnk, **e** 5he0 **f** 3fzm, **g** 6f1n, **h** 4a9m, **i** 5lj0, **j** 6wt. Ligand binding modes generated by our fragment-based docking approach are colored light

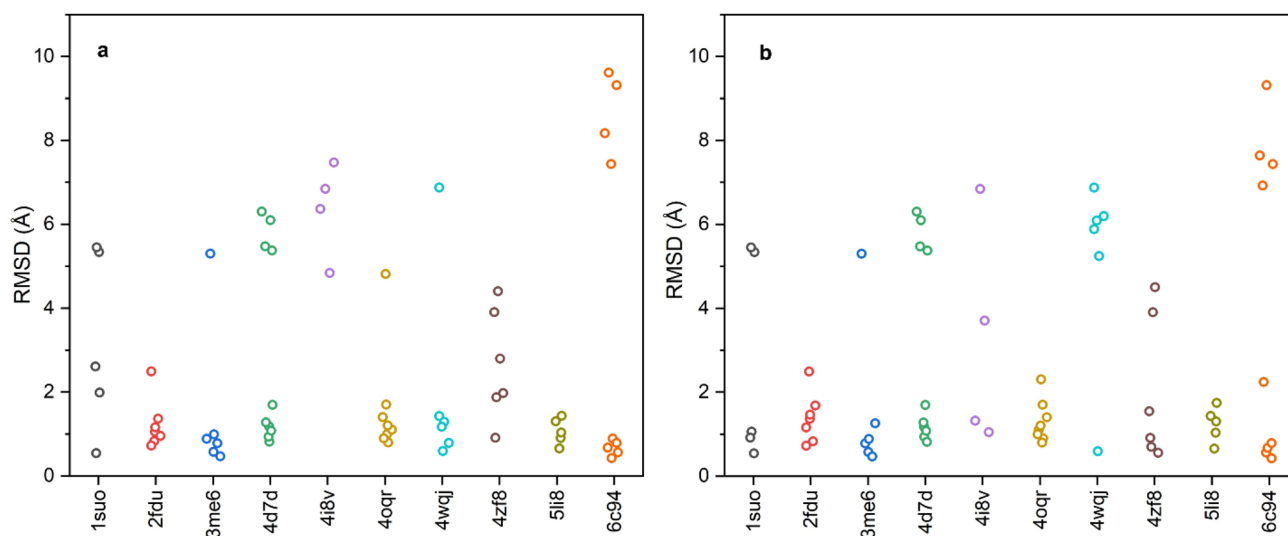
pink. Results of direct docking of complete ligands by AutoDock Vina are colored cyan. Actual ligand positions are shown as green sticks

**Table 5** Results indicating RMSD values using different models for three receptors

Protein	Model 1 RMSD (Å)	Model 2 RMSD (Å)	All atoms RMSD (Å) between models	RMSD (Å) between models (with outliers rejection)
4zf8	0.5	0.8	1.44	0.12
3fzm	1.8	2.2	1.33	0.14
1suo	0.7	1.0	1.18	0.07

**Table 6** Cross docking RMSD results for three complexes

Protein PDB (co- crystallized ligand ID)	CPZ ligand RMSD (Å)	CGE ligand RMSD (Å)	TB2 ligand RMSD (Å)
1suo (cpz)	0.5	1.8	2.4
3me6 (cge)	1.6	0.4	0.7
3r1a (tb2)	1.9	0.2	0.6



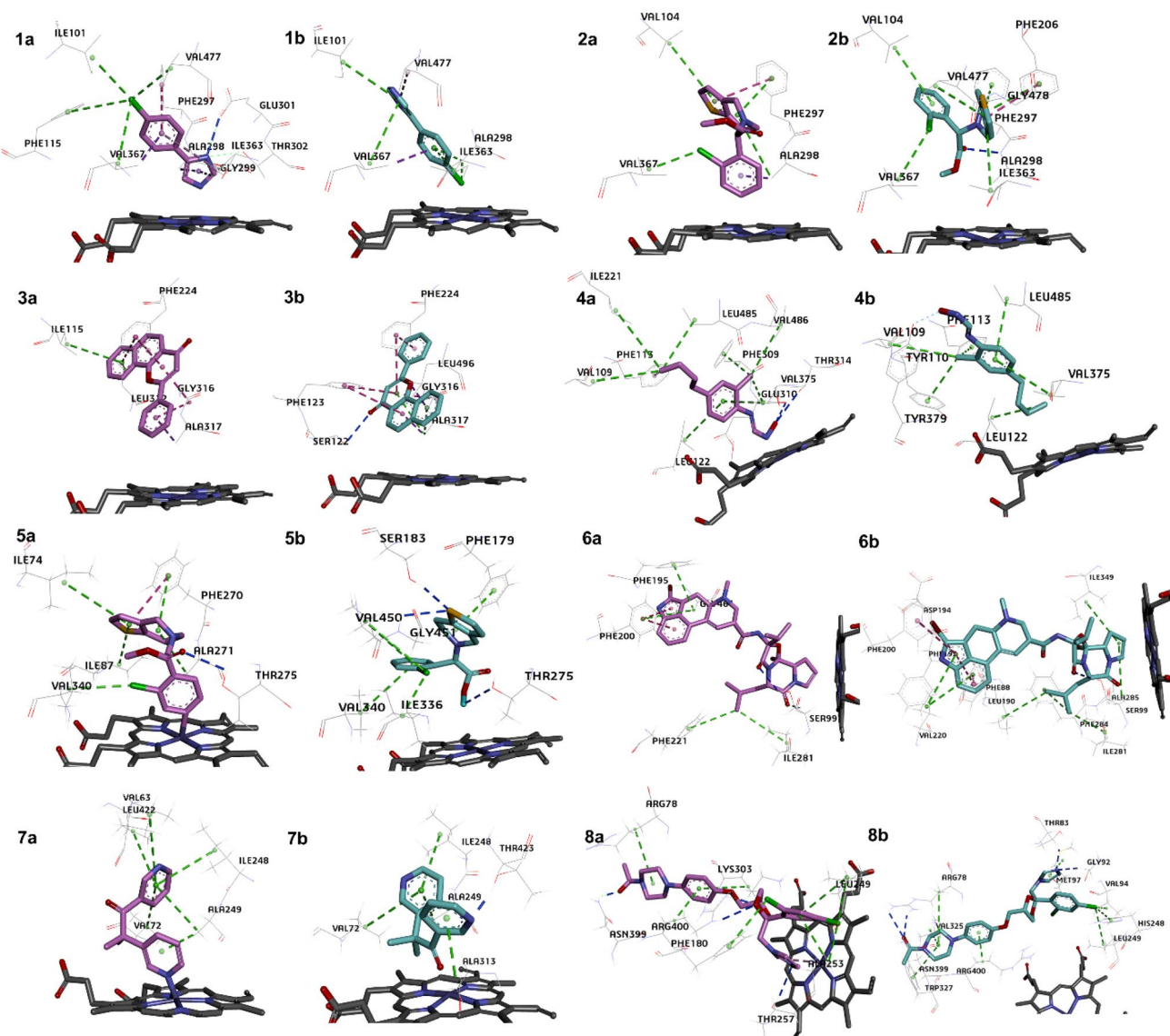
**Fig. 7** RMSD distributions of the bound docking results in top 10% hits according to **a** Cyscore **b** DSX ranking

interactions with the pocket residues (Fig. 8). In some cases, high RMSD can be attributed to flipped conformations. The molecule is completely flipped along its axis such that the reaction center is pointed in opposite direction (Fig. 8). In native pose 1a, imidazole is pointing towards heme group but in pose 1b the molecule has flipped by 180 degrees orienting chlorophenyl towards heme group. In pose 3b, the ligand molecule has flipped along two axis, moving benzene ring away from the heme group, and flipping the carbonyl at the same time. For pose 4b, we can see the flip along both axis. The good ligand-receptor shape-complementarity as well as the favorable local interactions of the false poses cause difficulty the scoring functions to tell them from the true poses. In some cases, there could be the possibility that not all possible binding poses have been captured in the corresponding crystallographic structures, as for some CYP450 enzymes, a ligand may bind in different conformations and leading to more than one reaction sites on the ligand.

Figure 7 shows the RMSDs from respective experimental ligand positions of the top 10% scored conformations for bound docking. Evaluated with the Cyscore scoring function, for 9 of the 10 complexes, the 10% top scoring binding poses included one or more experimental result-like ones. Evaluated with the DSX scoring function, experimental result-like

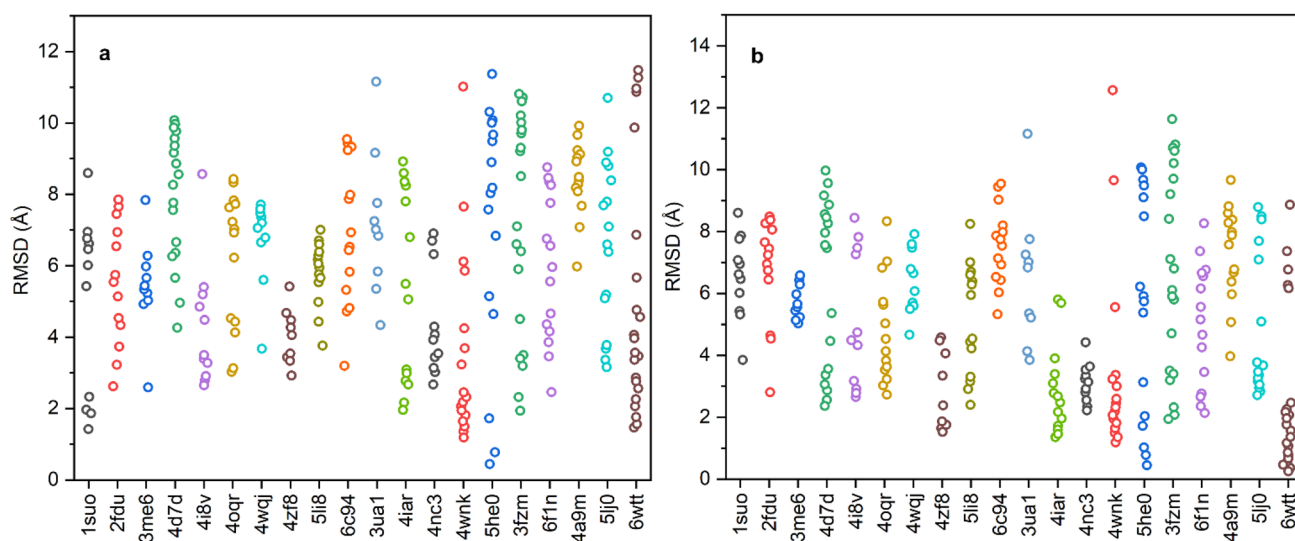
binding poses were included in the 10% top scoring binding poses for all complexes (Fig. 7b). Because of inaccuracies in both the structural models and the scoring functions, some false binding poses were unavoidably ranked top along with true binding poses. Visual inspection of several such false poses inside binding pockets reveal multiple favorable interactions with the pocket residues (Fig. 8). In some cases, high RMSD can be attributed to flipped conformations. The molecule is completely flipped along its axis such that the reaction center is pointed in opposite direction (Fig. 8). In native pose 1a, imidazole is pointing towards heme group but in pose 1b the molecule has flipped by 180 degrees orienting chlorophenyl towards heme group. In pose 3b, the ligand molecule has flipped along two axis, moving benzene ring away from the heme group, and flipping the carbonyl at the same time. For pose 4b, we can see the flip along both axis. The good ligand-receptor shape-complementarity as well as the favorable local interactions of the false poses cause difficulty the scoring functions to tell them from the true poses. In some cases, there could be the possibility that not all possible binding poses have been captured in the corresponding crystallographic structures, as for some CYP450 enzymes, a ligand may bind in different conformations and leading to more than one reaction sites on the ligand





**Fig. 8** Comparison of experimental-like poses with non-experimental-like poses that have been ranked top by the scoring function. The experimental-like poses are colored pink (a) and non-experimental-like poses are colored cyan (b). 3D representation of 8 docked complexes from bound and unbound docking are presented: (1) 1suo-cpz, (2) 3me6-cge, (3) 4i8v-bhf, (4) 6c94-v16, (5) 3me6-model, (6) 3ua1-

model, (7) 4zf8-model, (8) 5li8-model. The heme group is shown in gray sticks. Favorable contacts with binding pocket residues are indicated. Green lines are hydrophobic interactions of alkyl or pi-alkyl groups. Blue lines are hydrogen bonds. Pink lines are pi-pi interactions



**Fig. 9** RMSD distributions of unbound docking results for top 20% hits according to **a** Cyscore **b** DSX ranking

Figure 9 displays the RMSDs from respective experimental ligand positions of the top 20% scored conformations for unbound docking. In total 20 complexes have been considered for unbound docking. The Cyscore evaluation reveals that for 12 of 20 complexes, the top 20% scoring binding poses included one or more experimental result-like poses (within 3 Å). The DSX score evaluation yielded one or more experimental result-like poses for 14 of 20 complexes in top 20% results. In general, the accuracy of scoring functions was lower for ranking the unbound docking results than the bound docking results. This can be attributed to the fact that the unbound docked complexes originated from approximately modeled structures.

### Considering multiple ligand conformations

To test the fragment-based docking approach in a more realistic scenario, we further considered different ligand conformations. Two complexes (PDB IDs 5li8 and 4wnk) have been considered as examples. The respective ligands are named as KKK and 453, as in the corresponding PDB records. For each ligand, we generated 15 different

conformers (Figs. 10 and 11) using the ETKDG (Experimental-Torsion basic knowledge distance geometry) algorithm [71] contained in RDKit. Each of the resulting conformers were separately aligned to the docked fragment poses using the graph theory algorithm. Each conformer was also used for direct docking with AutoDock Vina. Complete ligand docking results for the GPCR complex (ligand 453) are as good as the fragment-based docking results. However, for the CYP450 complex (ligand KKK), fragment-based docking yielded much better RMSD results (Table 7). The binding pocket of the GPCR receptor is comparatively exposed, which makes it easier for complete ligand docking to find correct ligand orientation. On the other hand, for the docking of a relatively large ligand (ligand KKK) in to the buried pocket of the CYP450 receptor, direct docking of complete ligands seems to be rather challenging, especially with modeled receptor structures.

The resulting docked poses were also energy minimized and then scored using the DSX scoring function [70]. Figure 12 shows that for 10 of 15 KKK and 453 conformations, the top 20% scoring poses included one or more binding poses close to native (<3 Å).

**Table 7** Minimum RMSD values for different ligand (KKK, 453) conformations

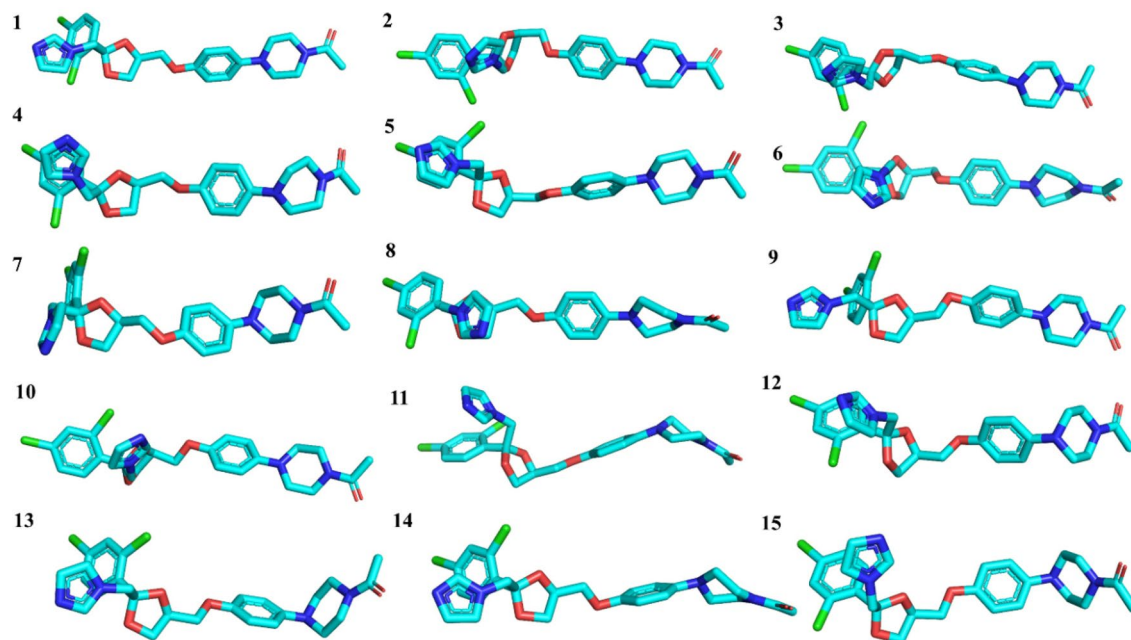
Ligand conformations	Minimum RMSD (Å) with our method	AutoDock Vina minimum RMSD (Å)
KKK-1	1.71	4.74
KKK-2	1.69	4.66
KKK-3	1.37	4.75
KKK-4	1.56	5.55
KKK-5	2.11	4.52
KKK-6	2.28	4.88
KKK-7	2.67	6.32
KKK-8	2.19	4.7
KKK-9	2.32	4.53
KKK-10	2.28	4.8
KKK-11	2.02	4.97
KKK-12	2.07	5.5
KKK-13	2.29	6.5
KKK-14	1.58	4.67
KKK-15	2.14	5.98
453-1	2.26	2.40
453-2	2.21	2.61
453-3	2.61	2.25
453-4	2.80	3.0
453-5	2.21	2.70
453-6	2.69	1.88
453-7	3.38	3.13
453-8	2.28	2.40
453-9	1.98	2.29
453-10	2.18	3.12
453-11	2.70	2.62
453-12	3.15	3.67
453-13	2.66	2.80
453-14	3.20	2.83
453-15	3.11	2.50

## Conclusions

In this article, we introduce a fragment-based docking method in which binding poses of complete ligands are determined based on alignments with docked fragment

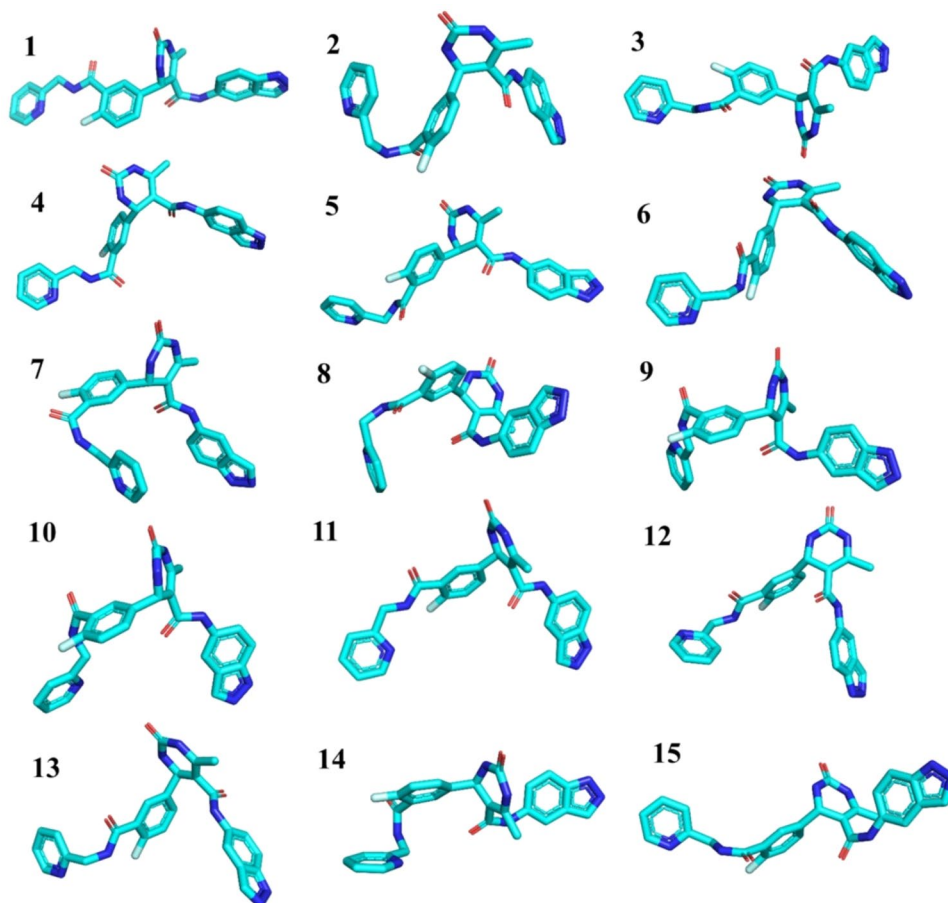
poses. A graph theory maximum clique algorithm is adopted to identify proper alignments between a multi-fragment ligand and a set of docked poses of various types of fragments. Bound docking of 10 CYP450 complexes verified that this fragment-based approach yields valid docking results as other established docking methods: for all the tested complexes, experimental-like binding positions of complete ligands can be successfully recovered. Then the fragment-based method was challenged with unbound docking, in which the input receptor structures were not from experimental complexes, but constructed through homology modeling. For all the complexes tested, binding poses with RMSD less than 3 Å from actual ligand binding positions could be recovered by the fragment-based approach, with 12 complexes having RMSD within 2 Å. For comparison, direct docking of complete ligands into the same modeled receptor pockets generated the binding positions in 12 of 20 complexes with RMSD < 3 Å.

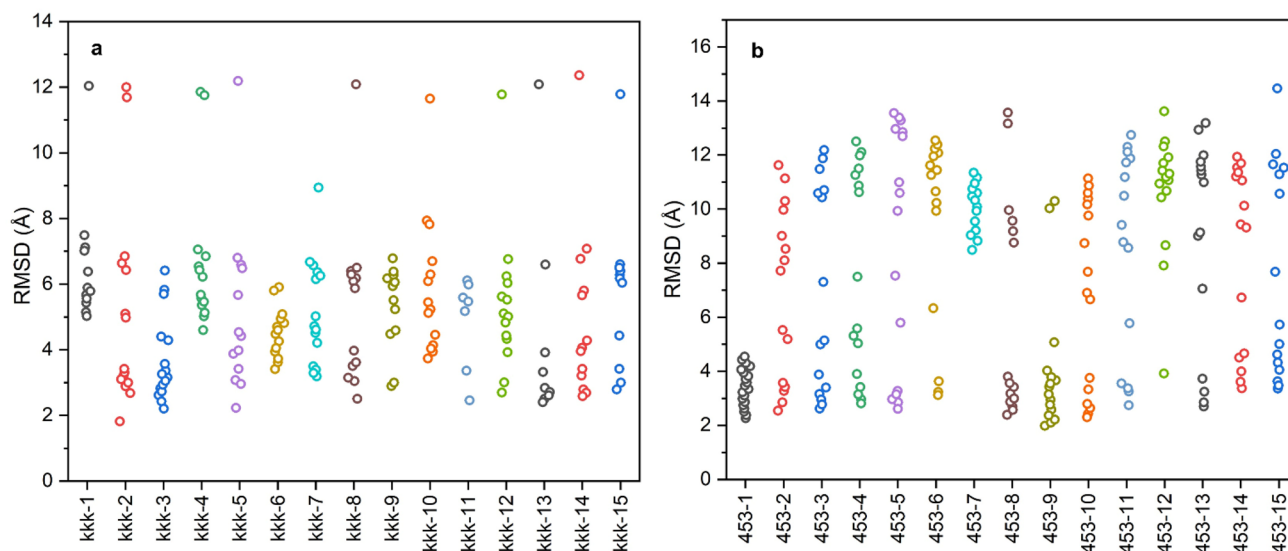
The relatively poor performance of unbound docking of complete ligands are consistent with the well-known fact that conventional docking approaches are highly sensitive to receptor structures, making unbound docking, especially docking with modeled receptor structures, a challenging task. The graph theory based fragment docking approach proposed here is intended to be less sensitive to receptor structures. In this new approach, different regions of a binding pocket are probed more systematically and thoroughly by small, rigid fragments. By retrieving possible binding positions of complete ligands through alignments with the docked fragments, one avoids difficulties associated with moving a relatively large ligand within the limited volume of a binding pocket according to an energy function. Because the different sub-pockets (the favorable binding sites for different fragments) are searched separately and independently, the requirements on the modeled receptor pocket to be accurate as a whole is efficiently reduced. Additionally, the relative geometries between the docked fragment poses are far less restricted than the relative geometries between the covalently linked fragments in a complete ligand. This may lead to better tolerance of inaccuracies in the relative geometries between different regions of a modeled receptor pocket. Compared with existed flexible docking approaches, the graph-based approach does not require extensive sampling of receptor



**Fig. 10** 15 conformations of ligand KKK used in unbound docking

**Fig. 11** 15 conformations of ligand 453 used for unbound docking





**Fig. 12** RMSD distributions for docking different conformations of **a** ligand KKK and **b** ligand 453 in top 20% hits according to DSX ranking

structures, which is time consuming and convergence or thorough coverage of different states important for ligand binding is not easily achieved. Our graph based docking strategy is less sensitive to receptor inaccuracies and hence gave good results even with poor quality structure models. However, with this method, one will end up in having multiple similar poses that need to be filtered out. In addition, this method requires prior knowledge of possible binding conformations to pick the best binding pose. In practice, this should not cause serious problems as the conformational space of a ligand are usually not enormous and a set of candidate binding conformations may be generated from systematic conformational analysis. These theoretical considerations and the actual docking results taken together, the fragment-based docking approach proposed here provides a useful new method for computational ligand docking with approximate, modeled receptor structure.

**Acknowledgements** This work has been supported by the Natural Science Foundation of China (Grant number 21773220).

**Author contributions** Haiyan Liu conceived the idea and supervised the project. Sara Sarfaraz and Iqra Muneer designed the algorithm and carried out the experiments. Sara Sarfaraz performed analysis and wrote the manuscript. Haiyan Liu revised the manuscript.

### Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interests.

### References

- Meng XY, Zhang HX, Mezei M, Cui M (2011) Molecular docking: a powerful approach for structure-based drug discovery. *Curr Comput Aid Drug Des* 7:146–157
- Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD (2015) Molecular docking and structure-based drug design strategies. *Molecules* 20:13384–13421
- López-Vallejo F, Caulfield T, Martínez-Mayorga K, Giulianotti MA, Nefzi A, Houghten RA, Medina-Franco JL (2011) Integrating virtual screening and combinatorial chemistry for accelerated drug discovery. *Comb Chem High T Scr* 14:475–487
- Xiang Z (2006) Advances in homology protein structure modeling. *Curr Protein Pept Sci* 7:217–227
- Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AW, Bridgland A, Penadones H (2020) Improved protein structure prediction using potentials from deep learning. *Nature* 577:706–710
- Ewing TJ, Makino S, Skillman AG, Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* 15:411–428
- Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* 46:499–511
- Thomsen R, Christensen MH (2006) MolDock: a new technique for high-accuracy molecular docking. *J Med Chem* 49:3315–3321
- Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R (2002) A new test set for validating predictions of protein–ligand interaction. *Proteins* 49:457–471
- Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem* 47:1750–1759
- Kramer B, Rarey M, Lengauer T (1999) Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. *Proteins* 37:228–241

12. McGovern SL, Shoichet BK (2003) Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J Med Chem* 46:2895–2907
13. Bordogna A, Pandini A, Bonati L (2011) Predicting the accuracy of protein–ligand docking on homology models. *J Comput Chem* 32:81–98
14. Sutherland JJ, Nandigam RK, Erickson JA, Vieth M (2007) Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J Chem Inf Model* 47:2293–2302
15. Kolb P, Phan K, Gao ZG, Marko AC, Sali A, Jacobson KA (2012) Limits of ligand selectivity from docking to models: In silico screening for A1 adenosine receptor antagonists. *PLoS One* 7:e49910
16. Teague SJ (2003) Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov* 2:527
17. McNae IW, Kan D, Kontopidis G, Patterson A, Taylor P, Worrall L (2005) Studying protein–ligand interactions using protein crystallography. *Crystallogr Rev* 11:61–71
18. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31:455–461
19. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein–ligand docking using GOLD. *Proteins* 52:609–623
20. Davis IW, Baker D (2009) RosettaLigand docking with full ligand and receptor flexibility. *J Mol Biol* 385:381–392
21. Amaro RE, Baudry J, Chodera J, Demir Ö, McCammon JA, Miao Y, Smith JC (2018) Ensemble docking in drug discovery. *Biophys J* 114:2271–2278
22. Evangelista Falcon W, Ellingson SR, Smith JC, Baudry J (2019) Ensemble docking in drug discovery: how many protein configurations from molecular dynamics simulations are needed to reproduce known ligand binding? *J Phys Chem B* 123:5189–5195
23. Yang J, Baek M, Seok C (2019) GalaxyDock3: Protein–ligand docking that considers the full ligand conformational flexibility. *J Comput Chem* 40:2739–2748
24. Erlanson DA, McDowell RS, O'Brien T (2004) Fragment-based drug discovery. *J Med Chem* 47:3463–3482
25. Chen H, Zhou X, Wang A, Zheng Y, Gao Y, Zhou J (2015) Evolutions in fragment-based drug design: the deconstruction–reconstruction approach. *Drug Discov Today* 20:105–113
26. Erlanson DA, Fesik SW, Hubbard RE, Jahnke W, Jhoti H (2016) Twenty years on: the impact of fragments on drug discovery. *Nat Rev Drug Discov* 15:605
27. Rees DC, Congreve M, Murray CW, Carr R (2004) Fragment-based lead discovery. *Nat Rev Drug Discov* 3:660
28. Chen Y, Shoichet BK (2009) Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat Chem Biol* 5:358
29. Joseph-McCarthy D, Campbell AJ, Kern G, Moustakas D (2014) Fragment-based lead discovery and design. *J Chem Inf Model* 54:693–704
30. Böhm HJ (1992) The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J Comput Aided Mol Des* 6:61–78
31. Wang R, Gao Y, Lai L (2000) LigBuilder: a multi-purpose program for structure-based drug design. *Mol Model Ann* 6:498–516
32. Majeux N, Scarsi M, Apostolakis J, Ehrhardt C, Caflisch A (1999) Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins* 37:88–105
33. Hao GF, Jiang W, Ye YN, Wu FX, Zhu XL, Guo FB, Yang GF (2016) ACFIS: a web server for fragment-based drug discovery. *Nucleic Acids Res* 44:W550–W556
34. Yamaotsu N, Hirono S (2018) In silico fragment-mapping method: a new tool for fragment-based/structure-based drug discovery. *J Comput Aided Mol Des* 32:1229–1245
35. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG (2011) Using graph theory to analyze biological networks. *BioData Min* 4:10
36. Ko E, Kang M, Chang HJ, Kim D (2017) Graph-theory based simplification techniques for efficient biological network analysis. in 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService). 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService) 277–280
37. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) Protein flexibility predictions using graph theory. *Proteins* 44:150–165
38. Canutescu AA, Shelenkov AA, Dunbrack RL Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* 12:2001–2014
39. Mitchell EM, Artymiuk PJ, Rice DW, Willett P (1990) Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J Mol Biol* 212:151–166
40. Deng H, Chen G, Yang W, Yang JJ (2006) Predicting calcium-binding sites in proteins: a graph theory and geometry approach. *Proteins* 64:34–42
41. Raymond JW, Willett P (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Comput Aided Mol Des* 16:521–533
42. Ji Y, Xu X, Stormo GD (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinform* 20:1591–1602
43. Reinharz V, Soulé A, Westhof E, Waldispühl J, Denise A (2018) Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families. *Nucleic Acids Res* 46:3841–3851
44. Mallek S, Boukhris I, Elouedi Z (2015) Predicting proteins functional family: a graph-based similarity derived from community detection. *Intell Syst* 2014:629–639
45. Ferguson CS, Tyndale RF (2011) Cytochrome P450 enzymes in the brain: emerging evidence of biological significance. *Trends Pharmacol Sci* 32:708–714
46. De Montellano PRO (2005) Cytochrome P450: structure, mechanism, and biochemistry. Springer, Berlin
47. Salazar NC, Chen J, Rockman HA (2007) Cardiac GPCRs: GPCR signaling in healthy and failing hearts. *BBA Biomembranes* 1768:1006–1018
48. Salon JA, Lodowski DT, Palczewski K (2011) The significance of G protein-coupled receptor crystallography for drug discovery. *Pharmacol Rev* 63:901–937
49. Insel PA, Tang CM, Hahntow I, Michel MC (2007) Impact of GPCRs in clinical medicine: monogenic diseases, genetic variants and drug targets. *BBA Biomembranes* 1768:994–1005
50. Hartl FU, Bracher A, Hayer-Hartl M (2011) Molecular chaperones in protein folding and proteostasis. *Nature* 475:324–332
51. Fujisawa T, Filippakopoulos P (2017) Functions of bromodomain-containing proteins and their roles in homeostasis and cancer. *Nat Rev Mol Cell Biol* 18:246
52. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. *Nat Med* 26:450–452
53. Rothan HA, Byrareddy SN (2020) The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *J Autoimmun* 109:102433
54. DeLano WL (2002) Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography* 40:82–92
55. Studio Discovery (2008) “Discovery Studio.” Accelrys [2.1]
56. Origin(Pro), Version 2019. OriginLab Corporation, Northampton
57. Webb B, Sali A (2014) Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics* 47:5.6
58. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci* 15:2507–2524

59. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M (2008) On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* 3:1503–1507
60. Landrum G (2013) Rdkit documentation. Release 1:1–79
61. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 30:2785–2791
62. Csizmadia P (1999) MarvinSketch and MarvinView: molecule applets for the World Wide Web. In Proceedings of ECSOC-3, the third international electronic conference on synthetic organic chemistry Sept. 1–30:67–369
63. Jain AK (2010) Data clustering: 50 years beyond K-means. *Patt Recogn Lett* 31:651–666
64. Bholowalia P, Kumar A (2014) EBK-means: a clustering technique based on elbow method and k-means in WSN. *Int J Comput Appl* 105:17
65. Orlin J (1977) Contentment in graph theory: covering graphs with cliques. *Indagationes Mathematicae (Proceedings)*. *Indag Math* 406–424
66. Makino K, Uno T (2004) New algorithms for enumerating all maximal cliques. in Scandinavian workshop on algorithm theory. *SWAT* 260–272
67. Bron C, Kerbosch J (1973) Algorithm 457: finding all cliques of an undirected graph. *Commun ACM* 16:575–577
68. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612
69. Cao Y, Li L (2014) Improved protein–ligand binding affinity prediction by using a curvature-dependent surface-area model. *Bioinform* 30:1674–1680
70. Neudert G, Klebe G (2011) DSX: a knowledge-based scoring function for the assessment of protein–ligand complexes. *J Chem Inf Model* 51:2731–2745
71. Riniker S, Landrum GA (2015) Better informed distance geometry: using what we know to improve conformation generation. *J Chem Inf Model* 55:2562–2574

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.