# Integration of Ligand and Structure Based Approaches for CSAR-2014

*Philip Prathipati*[*] *and Kenji Mizuguchi*

National Institutes of Biomedical Innovation, Health and Nutrition, 7-6-8 Saito-Asagi, Ibaraki City, Osaka 567-0085, Japan

ABSTRACT

The prediction of binding poses and affinities is an area of active interest in computer-aided drug design (CADD). Given the documented limitations with either ligand or structure based approaches, we employed an integrated approach and developed a rapid protocol for binding mode and affinity predictions. This workflow was applied to the three protein targets of Community Structure–Activity Resource-2014 (CSAR-2014) exercise: Factor Xa (FXa), Spleen Tyrosine Kinase (SYK) and tRNA (guanine-N(1)-)-methyltransferase (TrmD). Our docking and scoring workflow incorporates compound clustering and ligand and protein structure based pharmacophore modeling, followed by local docking, minimization and scoring. While the former part of the protocol ensures high-quality ligand alignments and mapping, the subsequent minimization and scoring provides the predicted binding modes and affinities. We made blind predictions of docking pose for one, five and 14 ligands docked into one, two and 12 crystal structures of FXa, SYK and TrmD, respectively. The resulting 174 poses were compared with co-crystalized structures (one, five and 14 complexes) made available at the end of CSAR. Our predicted poses were related to the experimentally determined structures with a mean root mean square deviation value of 3.4 Å. Further, we were able to classify high and low affinity ligands with the area under the curve values of 0.47, 0.60 and 0.69 for FXa, SYK and TrmD, respectively, indicating the validity of our approach in at least two of the three systems. Detailed critical analysis of the results and CSAR methodology ranking procedures suggested that a straightforward application of our workflow has limitations, as some of the performance measures do not reflect the actual utility of pose and affinity predictions in the biological context of individual systems.

## INTRODUCTION

The stated goal of the CSAR-2014 exercise is the evaluation of various computer aided drug design (CADD) techniques for pose and affinity prediction. The range of CADD techniques are broadly classified into ligand based drug design (LBDD) and structure based drug design (SBDD) approaches[1]. LBDD approaches are generally applicable to targets with no three-dimensional (3D) structures, are quicker and better interpretable to chemists. Prototypical ligand based approaches such as (3D) Quantitative Structure-Activity Relationships (QSAR) models or pharmacophore based models use physicochemical properties and conformational preferences for explaining the variation in structure-activity relationships (SARs). However, ligand based approaches can only be used when a sufficient amount of activity data are available and their utility for pose prediction depends on the availability of the bioactive conformations of one or more active compounds. Furthermore, selecting a model that reflects the biochemical reality is a difficult issue, as selection criteria that use statistical significance alone might lead to erroneous model prioritization from a pool of possible models.[2] On the other hand, SBDD can be used for making *de novo* predictions and for designing, more rationally, compounds that could interact with selected motifs in the protein's active site.[3, 4] However, this approach, suffers from huge computational costs, the uncertainties of the scoring functions and the problem of binding pose selection. [1,3-6,7] Previous CSAR contests have shown that while SBDD can reproduce ligand binding pose within 2Å, it fails to accurately rank order ligands by affinity.[5] In some instances (such as Urikinase) SBDD was able to classify active ligands from inactives.[5] In general, however, SBDD approaches failed to identify "activity cliffs", which a ligand based approach (such as QSAR) might be able to explain by analyzing the conformational preferences or physicochemical properties such as pKa, where a single log order change can produce several

orders of magnitude variation in activities.[8, 9] In view of the above, we propose an integrated structure and ligand based approach employing a structure-based pharmacophore to address the limitations and provide a rapid protocol for binding mode and affinity predictions. The workflow is also computationally scalable for the virtual screening of large chemical libraries.

In this paper, we report the results of our protocol applied to the three CSAR-2014 datasets in the following sections: (1) Integrated protocol for predicting binding poses and affinities, (2) Cluster analysis of target chemical space, (3) Structure, active site and structural integrity of functional motifs, (4) Protein structure based pharmacophore, (5) Blinded docking pose and affinity predictions, (6) Visual inspection of the binding modes of representative CSAR-2014 ligands and (7) The minimum requirements and applicability domain of the protocol.

## RESULTS AND DISCUSSION

### (1) Integrated protocol for predicting binding poses and affinities:

As discussed in Introduction, the application of ligand or structure based approaches alone presents significant challenges for pose and affinity prediction.[1] Hence, we developed a workflow illustrated in Figure 1, which encompasses a) cluster analysis of the target chemical space, b) conformation generation, c) ligand and protein structure-based pharmacophore generation and mapping (pose prediction) and d) subsequent minimization and scoring (affinity prediction). This workflow addresses the difficulties in: 1) the selection of pharmacophore models, which capture known essential interactions, 2) the automatic selection of docking poses and 3) the computational scalability. The calibration of the scoring function remains an unresolved issue and thus, the predicted affinity values may have to be interpreted with caution. Below, we will describe these workflow elements in more detail.

**(2) Cluster analysis of target chemical space**

Our workflow starts with the construction of the target chemical space, i.e., a collection of all the chemical compounds known to interact with a given target and their associated properties. Cluster analysis helps us identify representative compounds for the generation of pharmacophores and also further assess visually the binding modes and affinities of selected representative drugs and tool compounds. For this part of the study, we employed Weighted Gene Coexpression Network Analysis (WGCNA),[10] an advanced unsupervised machine learning method with elements similar to cluster analysis.[10] In contrast to most cluster analysis methods, WGCNA offers several unique features including a) topological overlap measure (TOM) as a proximity indicator to identify modules, b) tuning the soft thresholding parameter beta such that approximate scale free topology is reached and c) a fuzzy clustering method.[11] Hard-thresholding (rigid similarity cut-off such as 0.7) in cluster analysis has been a significant issue in chemical similarity analysis and often led to unreliable or subjective classification of chemical structures.[12] In addition, among the available range of clustering methods, network based methods are considered to be the best (accurate and reproducible) hence WGCNA was used to cluster compounds. WGCNA can furthermore relate clusters to external attributes such as chemical scaffolds to assess the similarity content of the modules derived using various cluster thresholding parameters.

Scaffolds are one of the major chemical space attributes and constitute the functionally important building blocks of chemical structures.[13,14,15] They have a strong influence on pharmacodynamic and pharmacokinetic profiles. Hence we analyzed the scaffold composition of the various modules identified by WGCNA by using the scaffolds generated by Scaffold Network Generator (SNG) [13] and molBlocks.[16] SNG is an open-source command-line utility that computes the

hierarchical network of scaffolds that define a large set of input molecules, while molBLOCKS is a suite of programs for breaking down sets of small molecules into fragments according to a predefined set of chemical rules, clustering the resulting fragments, and uncovering statistically enriched fragments.

Determining the number of clusters in a data set is a frequent problem in data clustering. In the present study, we applied a method analogous to the 'the elbow method' and visually inspected the relationship between the number of clusters and the total number of enriched scaffolds (Supporting information.2 tables S10-S12; see Materials and Methods for more details) and also examined the consistency between the cluster assignments and known chemical classification. The optimal thresholding cutoffs thus chosen resulted in 279, 71 and 322 clusters for the Fxa, SYK and TrmD global chemical spaces, respectively.

**(3) Structure, active site and structural integrity of functional motifs**

To prioritize the structure based pharmacophores that reflect the biological reality and are applicable to the global chemical space of the targets, understanding the functional motifs is critical.[18] Furthermore, since CSAR-2014 experimental data constitute the functional IC50 (half maximal inhibitory concentration) values, the structural integrity of the functional motifs is critically important for the application of our integrated LBDD and SBDD protocol for a reasonable activity prediction.[19] Hence we began by analyzing the structure, active site and structural integrity of functional motifs in comparison to prototypical crystal structures. A "prototypical" crystal structure of a target is a structure where all the major functional motifs and important secondary structure elements are intact.[19-21] Prototypical structures are often co-crystallized with natural substrates, as these studies are aimed at elucidating the structure-

function mechanisms of the target proteins. Although in the present study, we visually inspected for the conservation of the functional motifs, this process of selecting the prototypical structures could be automated using methods such as 'RINerator'.[22]

In addition, functional motifs in proteins contribute to their structure, and to the specificity of their interactions with other biomolecules including cognate ligands. They are often conserved across the broader superfamily and also constitute "hot spots" that can form strong interactions with ligands.[23] Site directed mutagenesis experiments furthermore indicate the crucial role of these in structure-function modulation of proteins.[24] Thus it is vitally important to review functional motifs and asses their structural integrity for a proper evaluation of pose and affinity predictions.

**(3.1) Functional motifs and structural integrity of FXa**

Factor Xa contains a serine protease domain in a trypsin-like closed two β-barrel fold encompassing the catalytic triad Ser195-His57-Asp102, two essential sub-sites S1 and S4, two disulphide bridges and an autolysis loop (table 1).[25, 26] The prototypical crystal structure (PDB ID, 1FAX) revealed a set of key motifs, including the S4 pocket that contains Aromatic box and cation hole. [2] The structural integrity of the S1 and S4 pockets is critical for the accurate prediction of the binding modes of FXa compounds. Among the three crystal structures provided in CSAR-2014, 'FXa_gtc000401_2.07.pdb' seemed the closest to the prototypical structure (in particular, the conserved autolysis loop) and hence was used for docking (Supporting information.2 figure S1). However, even this structure displays no cationic hole, with the side chain of Glu97 pointing away from the binding site. The other critical S1 pocket, too, is partially intact, with the Gln192 residue pointing away from the ligand binding site (Supporting

information.1 figure S1). These structural variations could introduce significant challenges in predicting the binding modes, especially of non-aromatic P4 fragments.

### (3.2) Functional motifs and structural integrity of SYK

The prototypical crystal structures of active SYK (PDB ID, 4FL1 and 1F6E) revealed a set of key residues (table 2 and Supporting information.2 figures S2 and S3). [27]The structural integrity of the catalytic site, the gate keeper residue (in hydrophobic pocket 2) and the p-loop (also known as the glycine rich loop) are critical for the accurate prediction of the binding modes of SYK compounds. Among the eight crystal structures provided by CSAR-2014, 'SYK_GTC000224.pdb' and 'SYK_GTC000249.pdb' seemed the closest to the prototypical structures and hence were used for docking. The SYK kinase domain of these two structures has a subdomain structure typical of other kinases, with a largely β-sheet N-terminal lobe, a largely α-helical C-terminal lobe, and the active site sandwiched between the two lobes. The N-terminal lobe consists of a five-stranded β-sheet plus a single α-helix equivalent to the C-helix of PKA. The larger C-terminal lobe is predominantly α-helical with three short β-strands: one at the hinge region and two between the activation loop and the main body of the C-lobe.

However, even these two structures display little structural integrity at the catalytic pocket, with the side chain of Lys402 pointing away from the binding site and no salt bridge interaction between the Lys402 and Asp512, which is required for the catalytic action of the kinase (Supporting information.2 figures S3 and S4). The other critical element, p-loop, though, is intact in 'gtc000249_SYK.pdb' but is completely missing in 'gtc000224.pdb' (Supporting information.2 figure S4). The flexible p-loop constitutes one of the major structure-function motifs of the ATP competitive inhibitors. It serves to close the gap between the N- and C-lobes thereby enabling the stabilization of the kinase in the inactive conformation in which ATP cannot

bind. We will discuss the implications of these structural variations for predicting the binding modes and affinities in the subsequent sections.

### (3.3) Functional motifs of TrmD

TrmD has a homodimeric structure with 29.7kDa per protomer.[28] The AdoMet binding site is located between the dimeric interfaces and forms a characteristic deep trefoil knot at the N-terminus. The AdoMet binding site forces the cofactor into a bent "L" conformation on binding. Most of the prokaryotic methyl transferaces (TrmDs) are classified as Type 1 tRNA methyltransferase.[29] They do not require the tertiary structure of tRNA for their enzymatic action.[29] The prototypical crystal structural of TrmD (Protein Data Bank ID 1P9P) revealed a key set of motifs (Table 3), which were observed in all the TrmD crystal structures submitted in CSAR-2014 competition.

## (4) Protein structure based pharmacophore

A pharmacophore describes the arrangement of molecular features that a ligand must contain to efficaciously bind a receptor.[30-33] Pharmacophore models are of two kinds, ligand based and protein structure-based. The former are derived using representative ligands in their bioactive conformation and the later are derived using molecular features by conversion of protein properties to reciprocal ligand space. In the present work, we derived an integrated pharmacophore model using the active site conformations of the representative compounds (selected by the WGCNA cluster analysis) co-crystalized with each target (figure 2). This model was initially generated as a ligand based model (using the PharmaGist software) and subsequently improved by adding excluded volumes, a feature of a protein structure based model

(using the Pharmer software). The three protein structure based pharmacophores used for generating the docking poses and affinities are described below.

### (4.1) FXa

Essentially, the FXa structure based pharmacophore (figure 2b) captures a) the pi- and hydrogen bonding interactions of '6-chloronaphthalene' with FXa's S1 pocket (Tyr228, Asp189, Ala190, Ser214) (labeled RingAromatic and Acceptor), b) the hydrogen bond acceptor interactions (labeled Acceptor and Acceptor) of the 2-sulfonamido group with Gln192, another critical residue of the S1 pocket, c) the hydrogen bond acceptor interactions (labeled Acceptor) of pyrrolidin-2-one and finally d) the interactions 1-methyl-1,4-diazepane terminal amines with the residues of S4 cationic hole (labeled Ionizable_Positive) as a positive ionizable feature.

### (4.2) SYK

The SYK structure based pharmacophore (figure 2d) captures a) the hydrogen donor and acceptor interactions of the hinge binding fragments with Ala451 and Glu452 (labeled Donor and Acceptor, respectively), b) the hydrophobic interactions of the hinge binding fragments with Met450 (labeled RingAromatic), c) cation pi-interactions of the catalytic site binding fragments with Lys402 (labeled RingAromatic), d) the hydrogen bond donor interactions of the catalytic site binding fragments with Glu420 from the C-helix (labeled Donor) and e) the hydrogen bond acceptor interactions of the hydrophobic pocket 1 fragments with Gly378 of the p-loop (labeled acceptor).

### (4.3) TrmD

Figures 2f illustrates the features of our TrmD pharmacophore models mapped to 'GTC000448A' co-crystallized with 'TRMD_448.pdb'. Essentially, the TRMD pharmacophore model is a mirror image of the interaction diagram of 'GTC000448A' with its cognate protein TRMD_448.pdb' (figure 2e).

**(5) Blinded docking pose and affinity predictions**

To create the docked structures, we minimized pharmacophore mapped conformations within the protein active sites to remove unfavorable contacts with the protein and relax the ligand conformations. The quantitative performance of the docked poses and affinity predictions was performed using the root mean square deviation (RMSD) and an analysis of the area under the curve (AUC) of the receiver operator characteristic (ROC) curves respectively.

**(5.1) Blinded docking pose prediction- quantitative assessment using RMSD values.**

The CSAR-2014 organizers provided the participants with multiple crystal structures (Table 4) for each of the three targets and blinded pose predictions were assessed for one FXa, five SYK and 14 TrmD ligands, respectively. As part of our analysis, we docked 163 FXa ligand structures into one FXa crystal structure 'fxa_gtc000401_2.07.pdb' and the RMSD value was calculated for the ligand used in the blinded docking assessment, 'GTC000101A'. Similarly for SYK we docked 272 SYK inhibitors into two SYK crystal structures, 'gtc000224_SYK.pdb' and 'gtc000249_SYK.pdb' using two structure based pharmacophore models. However, the RMSD values were estimated for the five SYK compounds docked into 'gtc000249_SYK.pdb' only. Thus we present the mean and median of these five RMSD values. Finally, we docked 31 TrmD ligands into the 14 crystal structures using 14 different structure based pharmacophore models. The blinded pose predictions were assessed for 14 of the 31 ligands docked into 12 of the 14

target structures resulting in 168 (= 14 × 12) RMSD values for TrmD. Thus, the total number of poses for which the RMSD was calculated based on our participation was 174 (= 1 + 5 + 168).

The overall mean value of 3.5 Å across 174 predicted poses suggests that protein structure based pharmacophores for pose prediction should be used with caution. The median and mean RMSD values of one FXa (6.03 Å and 6.03 Å) and five SYK (4.28 Å and 4.01 Å) inhibitors predicted by our method were comparatively higher than all the participants' median and mean values for FXa (2.94 Å and 2.98 Å) and SYK (2.58 Å and 2.46 Å) (table 4, figure 3). From the visual analysis of representative compounds of all three targets (presented in the next section), it is clear that most of the representative compounds' binding modes were reasonably close to the expected protein-ligand contacts. Though our protocol worked reasonably well in mapping the compounds to the structure based pharmacophores described above, the subsequent minimization with SMINA led to significant deviations; for instance 'GTC000101A' the FXa compound used in the blinded docking study deviated by 6.03 Å after minimization with SMINA (figures 3a and 3b). The low predicted score of '3.6044' also indicates the minimization did not reach convergence. The lack of adequate conformational sampling or even the solvent-exposed binding pocket of the FXa and absence of sufficient number of anchor residues or functional motifs such as the cation hole could also have contributed to the unusually large RMSD value of 'GTC000101A'.

The significant variation between the predicted and experimental structures for 'gtc000224' docked with 'gtc000224_SYK.pdb' can primarily be attributed to the '1,3‐oxazol‐5‐yl' fragment, which binds in the solvent-exposed region that is located outside the hinge pocket and adjacent to the ribose pocket. In addition, the flexible aminopropan-1-ol fragment of 'gtc000224'

also deviated significantly between the predicted and experimental poses and hence produced an overall RMSD value of 3.49 Å (figures 3c and 3d). The difficulties in predicting the binding modes of fragments at solvent exposed pockets were widely discussed in the literature.

The pose predictions for TrmD with median and mean RMSD values of 3.20 Å and 2.53 Å across 168 predicted poses were comparatively better than the overall median and mean RMSD values (3.84 Å and 3.52 Å) obtained by all the participants of this contest. This result could primarily be attributed to the buried polar binding pocket of TrmD. However, a major difference between the experimentally determined and predicted structures of GTC000448A is the result of an alternative overlay of oxadiazol-3-amine fragment, which resulted in an RMSD of 4.57 Å (figures 3e and 3f). This observation highlights a major issue with protein-structure based pharmacophores for pose predictions, where alternative alignments are possible and hence requiring manual inspection or an extended workflow to include protein-ligand fingerprint analysis. Appropriate tutomeric and protonation state assignments can also greatly reduce the possibility of such alternative docking poses.

**(5.2) Blinded docking rank ordering and classification of actives versus inactivies**

Although our approach can be described as an integrated approach, it is essentially a crude docking method and hence primarily an SBDD approach. While conventional docking approach starts by generating grid potential maps for the protein and ligand and using them to complementarily place the ligands within the protein pockets, the structure based pharmacophore approach uses predefined pharmacophoric features as constrains to dock ligands. As expected with a SBDD protocol, no statistically significant correlation was observed between experimental IC50 and predicted activity values with R2 values of 0.001, 0.013 and 0.164 for

FXa, SYK and TrmD, respectively (table 5). The poor correlation values of our approach could also be the result of the least squares based measures being less robust with respect to outliers. Our rapid and crude docking workflow together with inadequate conformational sampling also produced a large number of outliers, mostly as a result of the SMINA minimized poses, which did not converge to an energy minimum. In some instances the pharmacophore-predicted pose was stuck due to an unfavorable clash with the protein or due to an alternative alignment picked by the pharmacophore. Hence we decided to use the area under the curve (AUC) of receiver operator characteristic (ROC) curve values to quantitatively assess the classification accuracy of the predicted affinities. Furthermore the AUC measures from ROC curves are generally considered to be more robust with respect to outliers and hence can be considered better measures for quantitative assessment. The measured AUC values of all compounds predicted in the competition and the compounds of the individual clusters are presented in tables 6 and 7. In spite of the relatively poor correlation between the experimental and predicted affinities, our protein structure based approach discriminated active from inactive compounds reasonably well for at least two targets, as demonstrated by the AUC values of 0.47, 0.60 and 0.69 for FXa, SYK and TrmD, respectively (where the actives were defined as the compounds with pIC50 > 7.5 for FXa and SYK and with pIC50 > 5.5 for TrmD) (figure 4). Since we varied the “cutoff” values and had multiple structures to choose from for the SYK and TrmD systems, the AUC values of 0.60 and 0.69 for SYK and TrmD may have a level of overfitting and caution needs to be exercised in interpreting these results. For FXa and SYK, we choose a cut-off of ‘7.5’, since this cut-off afforded relatively equal numbers of actives and inactives in comparison to other cut-offs (table 6). For TrmD, the cut-off that afforded relatively equal numbers of actives and inactives was 5.5. In addition, among the two crystal structures of SYK that we used for docking, we

chose ‘gtc000224_SYK‘, as this crystal structure had a better structural integrity of active site motifs, Among the several high scoring TrmD crystal structures, we chose TrmD_448 for the ROC curve depiction, as its co-crystalized ligand was representative of the TrmD ligand set.

Some of the additional insights that can be gained from tables 6 and 7 include the identification of outlier clusters such as ‘deeppink’ FXa cluster (AUC=0.18) and ‘royalblue’SYK cluster (AUC=0.17).

**(6) Visual inspection of the binding modes of representative CSAR-2014 ligands**

The assessment of pose prediction is typically performed using quantitative measures such as RMSD or can be visually inspected for native contacts with selected critical residues implicated in the structure function mechanisms of the proteins. Previous CSAR exercises have highlighted the issues with RMSD and the need for multiple approaches to assess the performance of docking poses. The notable issues include the unreliability of RMSD values in a flexible docking protocol, the artificially low RMSD values of small molecules and high RMSD values of symmetric molecules.[5] In addition, docking typically fails to identify native poses in solvent exposed and flexible pockets such as the S2-S3 pockets in FXa and the p-loop pocket in SYK. Hence, a visual inspection of the binding modes of representative members of clusters of FXa, SYK and TrmD could throw further light on the general quality of docking and the conservation rate for ligand-residue contacts.

**(6.1) FXa**

Both the predicted poses of various FXa inhibitors and the crystal structures of potent inhibitors in complex with FXa (figures 3a and 3b, supporting information.2 tables S13 to S16) revealed

that the inhibitors adopt L-shaped binding conformations with two major anchoring points at S1 and S4 sites.

In the present study, we analyzed the binding modes and enriched fragments of CSAR-2104 FXa compounds comprehensively in relation to the global FXa chemical space. Such an analysis should help us infer the novelty in the CSAR-2104 FXa chemotypes, the difficulties in predicting the binding modes, and the physicochemical property profiles in comparison to well-studied FXa drugs and tool compounds.

In general, our protocol, despite poorly classifying the experimental activities of the FXa dataset (overall AUC= 0.47), reproduced expected binding modes of CSAR-2014 representative FXa ligands reasonably well (Supporting information.2 tables S13 to S15). In particular, most compounds faithfully reproduced the expected binding modes at the S1 pocket, where the halogen atom of naphthalene or thiophene moiety maintains hydrogen bond acceptor and T-shaped pi staking interactions with the S1 pocket residues (Supporting information.2 tables S13 to S15). In addition, our protocol was able to process the 163 compounds in several CPU minutes (compared to several CPU hours by typical docking packages) and thus it is computationally scalable for larger datasets. This way, it significantly reduced the efforts needed for post docking processing and analysis, which is typically performed manually and involves visual inspection. However, given the limited accuracy of the protein structure based pharmacophore in predicting the FXa inhibitor binding modes and affinities in the blinded docking study, our workflow has to be used with caution.

Since the cation hole in the S4 pocket is not intact in any of the structures provided by the CSAR-2014 organizers, the binding mode predictions for this site deviated from the expected trend. The co-crystallized ligand structure of 'FXa_gtc000401_2.07.pdb' also failed to interact with the sidechain or the backbone atoms of Glu97 and the observed conformation of the P4 fragment could be an artifact of crystal packing (figures 2a and 2b). Our pharmacophore model included features such as positive ionizable and may guide the P4 fragments into the S4 pocket. However, many docked poses (in particular 'darkorchid4' and 'coral3' modules) deviated from the expected placement of the P4 fragments, since our protocol also involved minimizing the ligand in the active site and found no anchor for the non-aromatic P4 fragments to engage with the S4 pocket (which contains both cation hole and the aromatic box in the prototypical structure). The 'tan' and 'darkorchid4' modules performed the worst in terms of predicting both the binding mode and binding affinities of the representative compound. Several reasons could be attributed; (1) adequate conformational sampling of FXa ligands was not performed, (2) the lack of structural integrity of the S4 and S1 motifs and (3) the conformational preferences of the compounds, which manifested during the SMINA minimization steps.

Furthermore, in contrast to natural substrate/inhibitor binding geometries, as well as the vast majority of synthetic inhibitor complexes with serine proteinases,[45] many compounds (GTC000051A_set3, GTC000047A_set2, GTC000104A_set3, GTC000075A_set3, GTC000101A_set2, GTC000091A_set3, GTC000103A_set3, B22617, and B50848802) formed no hydrogen bond with Gly-216. This can also be observed in the protein ligand interaction profiles of PDB IDs 1EZQ and 2P93 and explains the considerable variation of the P2 and P3 fragments in FXa inhibitor design. Thus, the geometric restraints of binding in the S1 and aryl-binding S4 pockets have allowed considerable variation of the P3 fragments in inhibitor design.

The co-crystallized and predicted binding modes of the compounds described demonstrate the feasibility of ignoring hydrogen bonding at Gly-216 and point to the existence of several possible alternative binding modes of the P3 fragments in particular and other fragments in general.[34]

The WGCNA cluster analysis and subsequent scaffold enrichment analysis revealed that 'N-(2-oxo-1-phenylpyrrolidin-3-yl)-2-(thiophen-2-yl)ethene-1-sulfonamide' is a common scaffold across most of the CSAR-2014 FXa modules, except for the 'brown4' module, which contains the 'N-thiophene-2-sulfonamide' as the common scaffold.

The homogeneity in chemical clusters can also be verified by structural similarity between the representative compound of each module and the enriched fragment or scaffold (Tables S13-S16). The compounds and their attributes (including physicochemical properties) of all the modules, the enriched scaffolds and the statically enriched scaffolds are described in supporting information.1 tables S1-S3.

### (6.2) SYK

Both the predicted poses of various SYK inhibitors and the crystal structures of potent inhibitors in complex with SYK (Supporting information.2 tables S17 to S19) revealed that the inhibitors adopt flat binding conformations with three major anchoring points at the hinge (Met450, Ala451 and Glu452), the catalytic pocket (Lys 402, Asp512 and Glu420) and the p-loop (Gly378 and Ser 379).

Our protocol reproduced the experimental binding modes for the five newly crystallized compounds reasonably well (with an RMSD of around four angstroms) and produced the

expected binding modes for most of the other compounds. In particular, most compounds faithfully produced the expected binding modes at the hinge pocket and the catalytic site (Supporting information.2 tables S17 and S18).

The structural integrity of the two kinases in the CSAR-2014 competition seemed to have little influence on the binding modes of the kinase inhibitors, since the inhibitors find additional anchors such as Glu420 in the catalytic pocket. However, since the structural integrity of the p-loop is only partially preserved in 'SYK_GTC000249.pdb' and is completely absent in 'SYK_GTC000224.pdb', the predicted binding modes in hydrophobic pocket 1 deviated from the existing co-crystal structures. In addition, since this pocket is solvent-exposed, it is generally difficult to predict the binding modes accurately without considering solvent effects adequately.

On the other hand, the lack of structural integrity seems to impact the predicted binding affinities. The predicted binding affinities of the CSAR-2014 SYK inhibitors docked to 'SYK_GTC000224.pdb' agreed better with the experimental activities (overall AUC=0.60] compared to those docked to 'SYK_GTC000249.pdb' (overall AUC=0.44] (tables 7). Since the catalytic site is better preserved in 'SYK_GTC000224.pdb' than in 'SYK_GTC000249.pdb' while the p-loop is better preserved in SYK_GTC000249.pdb' than in 'SYK_GTC000224.pdb', the catalytic site appears to be more important in predicting the experimentally measured inhibitory activities. The predicted activities are presented in supporting information.1 table S4 to S6 together with the module assignments and enriched fragments and significantly enriched fragments.

**(6.3) TrmD**

Both the predicted poses of various TrmD inhibitors and the crystal structures of representative inhibitors in complex with TrmD (figures 3e and 3f; supporting information.2 tables S20 and S21) revealed that the inhibitors adopt L-shaped binding conformations with two major anchoring points at the adenine binding pocket and the ribose site.

In general, our protocol reasonably reproduced the experimental affinity (AUC = 0.69) and the binding modes of the 14 newly co-crystalized compounds (with the mean root mean square distance of around 2.53 Å). In particular, most compounds faithfully reproduced the expected binding modes at the adenine binding pocket, where a pair of hydrogen bond acceptor and donor maintain interactions with the hydroxyl of Ser132 (figures 3e and 3f; supporting information.2 table S21).

**(7) The minimum requirements and applicability domain of the protocol**

The major strengths and limitations of our structure based pharmacophore approach were discussed in the various subsections of 5.1 to 5.3 and 6.1 to 6.3. In the following sections, we discuss the minimum requirements and applicability domain of the structure based pharmacophore approach for pose and affinity prediction.

The minimal requirements of our method are broadly akin to those of a typical protein-ligand docking study; they include polar and buried active sites of known structure with bound ligands, as highlighted in Meslamani et al. (2012)[18] and Smith et al. (2011).[35] In addition, our method is particularly sensitive to the structural integrity of key motifs, because our structure-based pharmacophore approach aims to optimize interactions with these motifs.[36] The importance of structural motifs was highlighted in structure-based drug design of kinases, where the protein structures are typically classified into DFG-in, DFG-out, and so on, and appropriate structures

are chosen for docking type1, type2 and type1.5 inhibitors.[37] However, in a real life scenario in which no such knowledge of structural classification is available, it is possible to use snapshots from a molecular dynamics trajectory[19-21], side chain rotamer exploration or a method such as that of Kufareva and Abagyan[38], who presented a computational protocol (Deletion-Of-Loop asp-PHe-gly-IN (DOLPHIN) kinase models) for converting multiple available DFG-in structures of various kinases into accurate and specific models of their type-II-bound state.

Protein-based pharmacophore models are of two kinds: those derived from the protein binding atoms and those derived using prior knowledge of active ligands. Our method belongs to the latter and is restricted to targets with at-least one co-crystal. Our method is similar in principle to Tropsha et al's CSAR-2013 submission[39], where protein ligand fingerprints were used to discriminate binders from non-binders. Furthermore Koes et al.[7] reported the application of an older version of SMINA for scoring co-crystallized ligands in the CSAR 2012 exercise with reported median ranked performance. However, in the present article we report for the first time the application of a protein-structure based pharmacophore approach incorporating (1) WGCNA clustering, (2) SNG-derived chemical scaffold analysis and (3) a combined protein structure based docking tool (Pharmagist + Pharmer + SMINA) for the prediction of binding modes and affinities of protein-ligand interactions in the CSAR exercise.

**Conclusion**

In the present study, we report the application of an integrated ligand and structure based approach to the prediction of binding modes and affinities for FXa, SYK and TrmD inhibitors. We highlighted major advantages of this approach such as the derivation of pharmacophore

models applicable to the global target chemical space, the automated selection of binding modes, and computational scalability, together with limitations such as limited sampling rates (due to pre-computed conformers). The performance of our protein structure pharmacophore approach was affected in part by the SMINA minimization, which often did not converge to a minimum. Future work will address these issues. We also highlighted the relevance of structural integrity of the S1 and S4 pockets for accurate pose and affinity predictions of FXa compounds, and the catalytic site integrity for the pose and affinity predictions of SYK compounds.

## Materials and methods

### FXa CSAR-2014 and FXa bindingDB ligand dataset

The CSAR-2014 FXa dataset consisted of three sets,of $pIC_{50}$ SAR data points for a total of 163 compounds with some overlap between the sets. The pIC50 values ranged from ~5 to ~9. This dataset was combined with the BindingDB Fxa dataset consisting of 6141 compounds, as downloaded from the most recent version of bindingDB. The activity data in bindingDB is reported in terms of Ki, Kd and IC50 at nano-molar concentrations. Using an approximation of the Cheng-Prusoff equation Ki = IC50/2, the IC50 values were converted to ki values (IC50_to_ki)**.**[40] For consistency, the averages of these three values (ki, kd and IC50_to_ki) were converted to log scales. Along with the SAR data, several physicochemical properties and scaffolds were computed and presented along with some useful attributes extracted from bindingDB (Supporting information.1 tables S1 to S3).

### SYK CSAR-2014 and SYK bindingDB ligand dataset

The CSAR-2014 SYK dataset consisted of a single set of $pIC_{50}$ nM SAR data points for 248 compounds. The pIC50 values ranged from ~5 to ~9. This dataset was combined with the BindingDB SYK dataset consisting of 2500 compounds, as downloaded from the most recent version of bindingDB. The activity data in bindingDB were averaged and converted to log scales as described above. Along with the SAR data, several physicochemical properties and scaffolds were computed and presented with other useful attributes extracted from bindingDB (Supporting information.1 tables S4 to S6).

**TrmD CSAR-2014 and TrmD bindingDB ligand dataset**

The CSAR-2014 TRMD dataset consisted of a set of $pIC_{50}$ SAR data points for 31 compounds. The pIC50 values ranged from ~4 to ~9. This dataset was combined with the 50 TrmdD ligands from AstraZeneca, 6600 anti-Mtb compounds, and 300 antibacterial compounds, along with 600 random (non-antibacterial) compounds and 350 (non-antibacterial) drugs. (Supporting information.1 tables S7 to S9).

**Chemical cluster analysis using WGCNA**

WGCNA, a clustering algorithm that selects clustering cutoffs such that the resulting correlation network follows a scale-free distribution,[11] was used to find sets of structurally similar compounds. Using this approach, clusters of compounds with similar structures were grouped into modules (labeled by color). Though the WGCNA methodology is typically used for studying relationships between gene expression levels,[10] here we used these techniques for studying structural similarities between compounds. WGCNA supports the assembly of both unsigned (the Pearson correlation of an unsigned similarity measure is defined as $s_{ij}^{unsigned} = abs(cor(x_i,y_j)$, where chemical fingerprint profiles xi and xj consist of the bitvector fingerprints of

length 4096 for compounds i and j) and unsigned (the Pearson correlation of an unsigned similarity measure is defined as $s_{ij}^{signed} = (1+cor(x_i,x_j))/2))$ networks.

Since WGCNA was applied for the first time to the cluster analysis of chemoinformatics datasets, we also elaborate on two critical aspects of our approach. Firstly, the determination of similarity thresholds in chemoinformatics (or in genomics) has been subjective and hence we use cluster enrichment analysis for selecting the appropriate thresholding parameters. Cluster enrichment analysis is a popular method for comparing the similarity content of different sets (either compounds or genes) and calculates enriched entities (such as scaffolds in chemoinformatics and gene ontology terms in genomics). In the present study, we used scaffolds as entities for enrichment analysis (see below for the definition of scaffolds). Given a set of clusters, let $n_i$ the number of scaffolds enriched in cluster $i$ compared to all the other clusters (with FDR $p<0.05$ from Fisher's exact test). We then calculated the total number of enriched scaffolds by summing up $n_i$ for all the clusters. We examined the relationship between the number of clusters and the total number of enriched scaffolds (Supporting information.2 Tables S10-S12). Generally, as the number of clusters increased, the total number of enriched scaffolds increased up to a point (sometimes known as the "elbow"), where it started to plateau or decrease. While the elbow method uses relative variance explained at various cutoffs, we used enrichment of chemical scaffolds to identify the optimal number of clusters.

For FXa and SYK this point was 0.35 for both the unsigned and signed networks. Between the unsigned and signed networks, we selected the one, which led to clusters more consistent with the known chemical class information as described in the bindingDB entity names. Thus, we selected thresholding cutoffs of unsigned_0.35 for Fxa and signed_0.35 for SYK. For TrmD, no elbow point was identified and thus, we examined the consistency between the resulting clusters

and known antibacterial classification and selected a thresholding cutoff of unsigned_0.45. In this manner, we aimed to choose the threshold that provided the highest number of enriched entities with the least number of clusters.

Secondly, enriched fragments presented in the current study are essentially the same as chemical scaffolds. Fragmentation of chemical structures is often performed using two basic approaches: (1) fragmentation using a predefined set of chemical rules and (2) fragmentation using the concepts of scaffold tree and scaffold networks. In a typical SAR study, enriched fragments are subsequently determined by binning the compounds into active and inactive compounds and determining the fragments that map to active in comparison to inactive compounds. However in the present study, we performed chemical clustering and scaffold (or enriched fragment) analysis to characterize the chemical space; we did not perform structure-function analysis. Enriched scaffolds refer to scaffolds enriched in compounds of a given cluster versus compounds of all other clusters. Hence our definition of enriched fragments differs from that generally used in the literature.

For FXa, we compiled 6141 compounds from BindingDB, and merged this FXa target chemical space with the 163 CSAR-2014 FXa compounds, for which the binding poses and affinities should be predicted. The global FXa chemical space (BindingDB + CSAR-2014 FXa chemical space) was clustered using WGCNA[10] with ECFP_12 (extended connectivity fingerprint of length 12) fingerprint[17] (Supporting information.1 table S1).

In the present study, we explored both signed and unsigned correlation networks derived at various merging thresholding parameters (0.0005, 0.005, 0.05, 0.1, 0.25, 0.35, 0.45). Since an unsigned network with a merging thresholding parameter of 0.35 presents the least number of

clusters with the highest number of enriched fragments or scaffolds (Supporting information.2 table S10), the results with this setting are discussed in this article. Supporting information.1 table S1 shows the 279 modules (cluster assignment of various compounds) and supporting information.1 table S2 shows the results of the SNG's scaffold enrichment analysis of various modules. These modules (clusters) were used to identify representative compounds for pharmacophore model generation (see the next section), for assessing the binding mode and affinity prediction, and for chemical space analysis (Supporting information.2 tables S13-S16).

A similar protocol using WGCNA and SNG has implemented for clustering the SYK chemical space (Supporting information.1 table S4). The SYK chemical space constitutes the 248 CSAR-2014 SYK inhibitors and 2182 SYK binding DB compounds. Since a signed network with a merging threshold parameter of 0.35 presents the least number of clusters with the highest number of enriched fragments or scaffolds (Supporting information.2 table S11), the results with this setting are discussed in this article. Supporting information.1 table S4 shows the 71 modules (cluster assignment of various compounds) and Supporting information.1 tables S5 and S6 shows the results of the SNG's scaffold enrichment analysis of various modules. These modules (clusters) were used to identify representative compounds for pharmacophore model generation (see the next section), for assessing the binding mode and affinity prediction, and for chemical space analysis (Supporting information.2 tables S17-S19).

For clustering the TrmD chemical space we first merged 31 TrmD CSAR-2014 ligands with 33 PubChem derived TrmD ligands together with 7820 antibacterial compounds. Very few TrmD chemotypes are reported in the literature and hence we used 7820 antibacterial compounds, which also contained 312 antibacterial decoys and 527 human drugs used in previous publications. Since an unsigned network with a merging thresholding parameter of 0.45 with the

ECFP12 fingerprints presented the clustering most consistent with the known antibacterial classification (Supporting information.2 table S12), the results with this setting are discussed in this article (Supporting information.1 table S7). We initially analyzed the scaffold composition of the resulting modules by using SNG but we found no enriched fragments, presumably because SNG was unsuitable for diverse and sparse chemical space such as the antibacterial chemical space. Hence, we used molBlock (which uses retro-synthetic rules) for deriving scaffolds. Supporting information.1 table S7 shows the 322 modules (cluster assignment of various compounds) and supporting information.1 tables S8 and S9 shows the results of the molBlock's scaffold enrichment analysis of various modules. These modules (clusters) were used for assessing the binding mode and affinity prediction, and for presenting the results of the *in silico* forward chemical space analysis (Supporting information.2 table S21).

**Scaffold calculation using SNG and molBlocks**

The input datasets which includes CSAR-2014 Fxa,SYK and TrmD datasets together with the bindingDB FXa, SYK and TrmD sets were pre-processed using Discovery studio version 3.5 and subsequently submitted to SNG (FXA, SYK and TrmD) and molBlocks (TrmD) for scaffold tree and network generation. The output was parsed and used for enrichment analysis using WGCNA. The preprocessing steps included assigning bond orders and perceiving aromaticity to canonicalize the structures.

**Enrichment analysis**

The WGCNA functions 'userListEnrichment'[41] and 'enrichmentAnalysis' (of the 'anRichment' package) were used to measure enrichment between WGCNA derived modules and sets of

compounds with common scaffolds generated by the SNG package. Significant enrichment was measured using a hypergeometric test.

**Pharmacophore generation**

The bioactive conformations of around 30 ligands (described in Supporting information.2 tables S16,S19 and S20) co-crystallized with the three target proteins served as input for pharmacophore identification using the PharmaGist software.[42] Among the multiple pharmacophore schemes that were generated we choose the one closely resembling the known interactions of the ligands with the target proteins.

**Docking and scoring**

The CSAR-2014 Fxa.SYK and TrmD dataset provided by the organizers was subjected to multiple (50) conformation generation using the default settings of the FROG2 package.[43] The docking was performed using the Pharmer software[44] and the chemical pharmacophoric features derived using the PharmaGist software served as constrains to guide the placement of the ligands during docking. The protein backbone of the residues other than the interacting residues in the active site served as excluded volumes. These docked structures were subjected to SMINA optimization and scoring.

## ASSOCIATED CONTENT

Supporting Information Available: Twenty one additional tables: Table S1. The IDs, module assignments, activities, predicted physicochemical properties, smiles, aliases of the global FXa chemical space, Table S2. The complete results of the FXa module enrichment analysis using SNG fragments, Table S3. The results of the significantly enriched fragments of the FXa modules, Table S4. The IDs, module assignments, activities, predicted physicochemical properties, smiles, aliases of the global SYK chemical space, Table S5. The complete results of the SYK module enrichment analysis using SNG fragments, Table S6. . The results of the significantly enriched fragments of the SYK modules, Table S7. The IDs, module assignments, activities, predicted physicochemical properties, smiles, aliases of the global TrmD chemical space, Table S8. The complete results of the TrmD module enrichment analysis using SNG fragments, Table S9. The results of the significantly enriched fragments of the TrmD modules, Table S10. The number of modules and enriched fragments (at an FDR cutoff < 0.05) based on WGCNA clustering of 6304 Fxa compounds. Table S11. The number of modules and enriched fragments (at an FDR cutoff < 0.05) based on WGCNA clustering of 2455 SYK compounds. Table S12. The number of modules and enriched fragments (at an FDR cutoff < 0.05) based on WGCNA clustering of 7885antibacterial and TrmD compounds. Table S13. Binding mode of representative FXa compounds from ‘orange’, ‘linen’, ‘darkorchid4’, ‘paleturquoise3’, ‘goldenrod3’ clusters. Table S14. Binding mode of representative FXa compound from the ‘coral3’ cluster. . The CSAR compounds from this cluster have no structural analogs PDB or drugbank Table S15. Binding mode of representative FXa compounds from ‘brown4, ‘tan, ‘coral2, ‘deeppink’ clusters. Table S16. 2D & 3D Binding modes of 39 other representative FXa drugs compounds not covered in CSAR-2014. Table S17. Binding mode of representative SYK

compounds docked into CSAR-2014 pdb file 'SYK_GTC000249.pdb' from 'brown, 'linen', 'royalblue' and 'blue, modules (supplementary tables S6 to S8). Table S18. Binding mode of representative SYK compounds docked into CSAR-2014 pdb file 'SYK_ GTC000224.pdb' from 'brown, 'linen', 'royalblue' and 'blue, modules. Table S19. 2D & 3D Binding modes of other representative SYK drugs and compounds under clinical investigation. Table s20. The protein structure based pharmacophore mapped to the 2D interaction diagram of various CSAR-2014 TrmD crystal structures. Table S21. Binding mode of representative TrmD compounds docked into CSAR-2014 pdb file 'TRMD_448.pdb' from 'brown, 'linen', 'royalblue' and 'blue, modules. Four additional figures: Figure S1. Structural integrity of key FXa motifs. Figure S2. The 2D interaction profile of '4FL1' active site with its co-crystal phosphoaminophosphonic acid-adenylate ester (ANP). Figure S3. Structural integrity of key SYK motifs. Figure S4. Cartoon representation of the structural integrity of key SYK motifs. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

* Phone: +81 72-641-9890 . Fax: +81 72-641-9881. **E-mail:** philip@nibiohn.go.jp.

### Notes

The authors declare no conflict of interest.

## ACKNOWLEDGMENT

This work was in part supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (Grant Numbers 25430186 and 25293079) and from Japan Agency

for Medical Research and Development (“The adjuvant database project”) to K.M. We thank Drs. Shandar Ahmad and Yayoi Natsume for the critical reading of the manuscript and other members for the discussions.

**FIGURES**

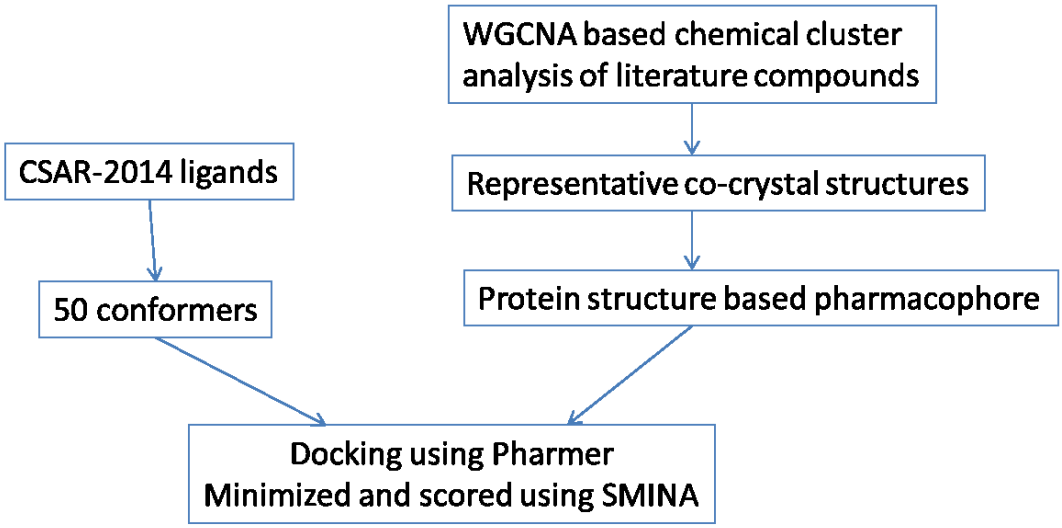



**Figure 1.** The proposed workflow used for docking and scoring the CSAR-2014 compounds.

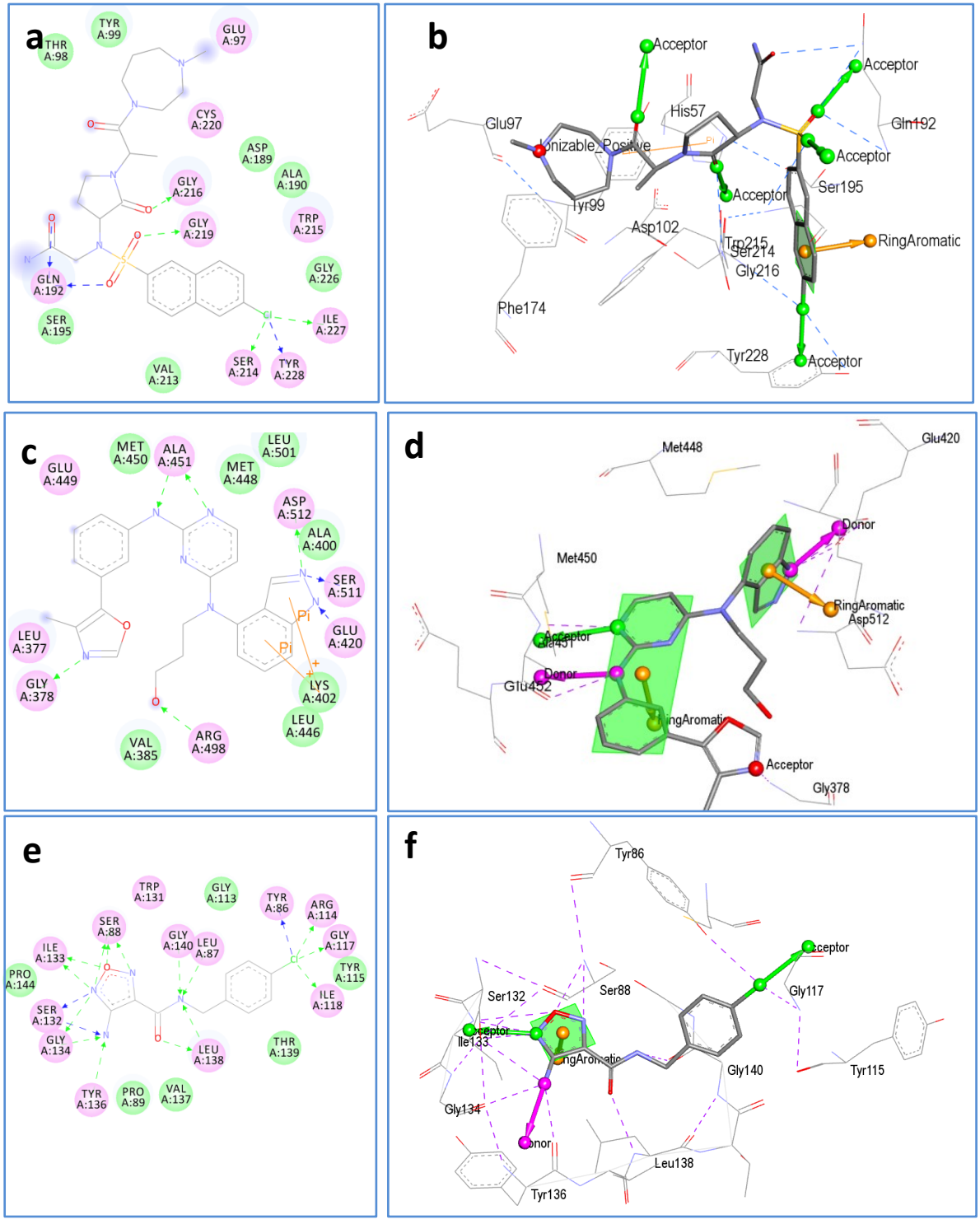



**Figure 2.** (a) The 2D interaction profile of 'FXa_gtc000401_2.07.pdb' active site with its co-crystal structure. (b) The 3D interaction diagram 'FXa_gtc000401_2.07.pdb' with its co-crystal

structure together with the ligand based pharmacophore that was used for docking CSAR-2014 FXa ligands. (c) The 2D interaction profile of ‘gtc000224_SYK.pdb’ active site with its co-crystal structure ‘GTC000224A’. (d) The 3D interaction diagram ‘gtc000224_SYK.pdb’ with its co-crystal structure ‘GTC000224A’ together with the ligand based pharmacophore that was used for docking CSAR-2014 SYK ligands. (e) The 2D interaction profile of ‘TRMD_448.pdb’ active site with its co-crystal structure ‘GTC000448A’. (f) The 3D interaction diagram ‘TRMD_448.pdb’ with its co-crystal structure ‘GTC000448A’together with the ligand based pharmacophore that was used for docking CSAR-2014 TrmD ligands.

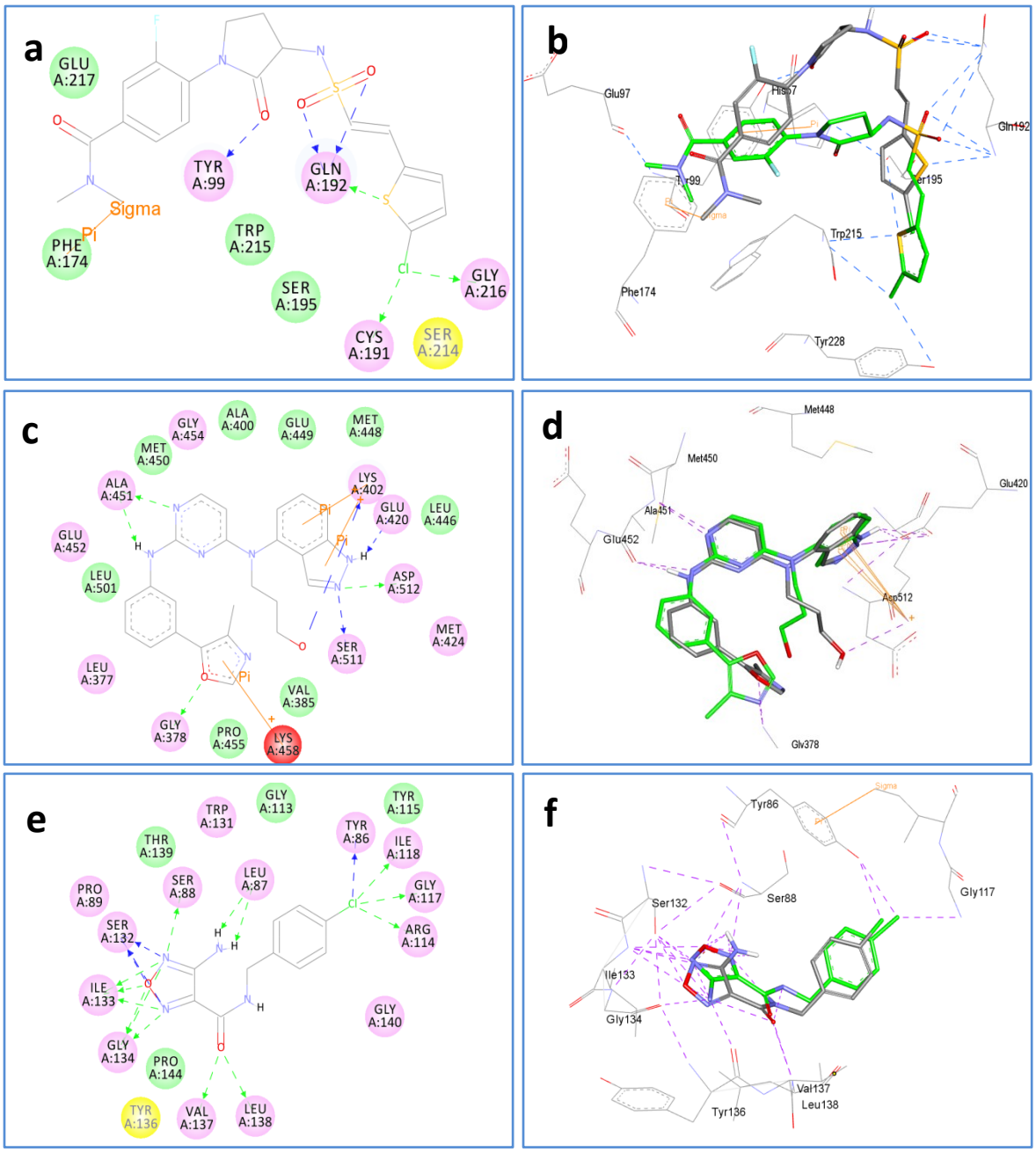



**Figure 3.** (a) The 2D interaction profile of the predicted pose of 'GTC000101' with 'fxa_gtc000401_2.07.pdb', (b) The 3D interaction diagram 'FXa_gtc000401_2.07.pdb' with its cognate co-crystal structure (in green) together the predicted pose of 'GTC000101', (c) The 2D interaction profile of the predicted pose of 'GTC000224A' with 'gtc000224_SYK.pdb', (d) The 3D interaction diagram 'gtc000224_SYK.pdb' with its co-crystallized (in green) and the

predicted pose of ‘GTC000224A’, (e) The 2D interaction profile of the predicted pose of ‘GTC000448A’ with ‘gtc000224_SYK.pdb’. (f) The 3D interaction diagram ‘TrmD_448.pdb’ with its co-crystallized (in green) and the predicted pose of ‘GTC000448A’.

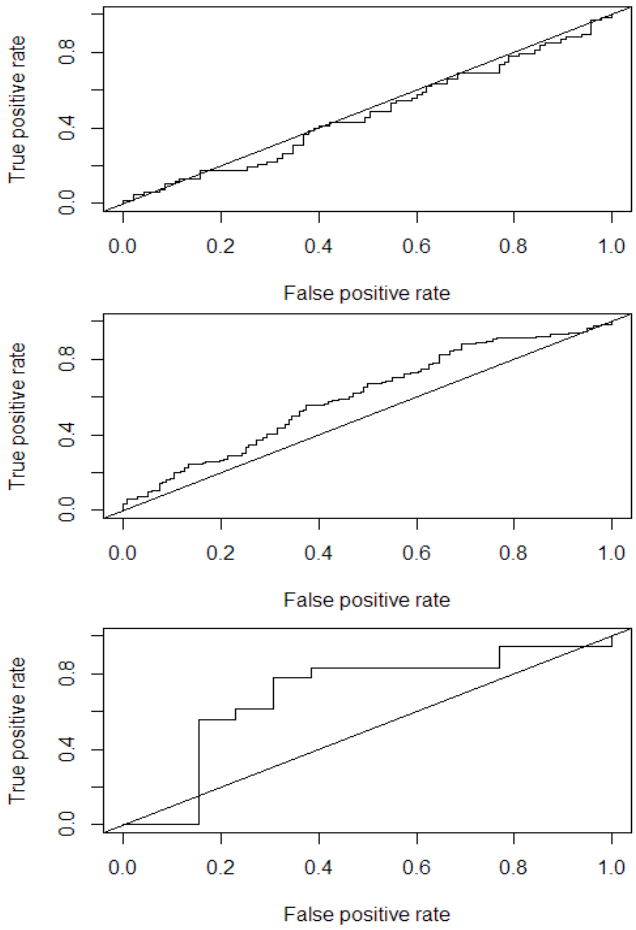



**Figure 4.** The ROC curves for FXa, SYK and TrmD depecting the prediction quality of the protein structure based pharmacophore approach.

TABLES

**Table 1.** The motifs/binding sites, residues that constitute these motifs/binding sites and their structural integrity status in the CSAR-2014 FXa crystal structure 'FXa_gtc000401_2.07.pdb'.

| motif/binding site | Residues | FXa_gtc000401_2.07.pdb |
|---|---|---|
| Catalytic triad | Asp102,His57 & Ser195 | Intact |
| S4 | Phe174,Glu97,Tyr99,Trp215 | Partially intact, but Glu97 is pointing away |
| Aromatic Box | Phe174,Tyr99,Trp215 | intact |
| Catation hole | Backbone carbonyl & sidechain of glu97 | Not present |
| S1 | Gly219,Asp189,Tyr228 & Gln 192 | Partially intact, but Gln192 is pointing away |
| Disulphide bridge 1 | Cys191-Cys220 | Intact |
| Disulphide bruidge 2 | Cys58-Cys42 | Intact |
| Autolysis loop | Arg143-Arg154 | Intact |
| Overall structure | 2 beta barrels + 3 alpha helicies | Intact |

**Table 2.** The motifs/binding sites, residues that constitute these motifs/binding sites and their structural integrity status in 3FQE, "SYK_GTC000224.pdb" and "SYK_GTC000249.pdb".

| Motifs/binding sites | Residues | 3FQE | SYK_GTC000224 | SYK_GTC000249 |
|---|---|---|---|---|
| Hinge binding pocket | Ala451 ,Met450 | Intact | Intact | Intact |
| Gate keeper | Met 448 | Intact | Partially intact | Partially intact |
| hydrophobic pocket 1 (left of hinge) | Glu452,Gly454,Leu377 | Intact | Intact | Intact |
| hydrophobic pocket 2 (right of hinge) | Met448,Val433,Glu449 | Intact | Intact | Intact |
| phosphorylation site (autocatalysis) | Tyr525 | Intact | Intact | Intact |
| DFG motif | Asp512-Gly514 | Intact | Not intact | Not intact |
| catalytic lysine | Lys402 | Intact | Not intact | Not intact, pointing away from the binding site |
| p-loop | Gly378-Thr384 | Intact | missing | Intact |
| activation loop | Asp512-Glu542 | Partially intact | Partially intact | Partially intact |
| allosteric site | - | - | | |
| C-alpha helix | Pro411-Gln425 | Intact | Intact | Intact |

**Table 3.** The motifs/binding sites, residues that constitute these motifs/binding sites in 1P9P and the major mutagenesis experiments.

| TrmD motif/binding site/structural element | residues | Major mutagenesis experiments within these motifs |
|---|---|---|
| Active site/ B4 | Leu85-Gln91 | G91 → A: Loss of activity; no effect on tRNA binding. |
| Active site/SpoU | Gly113-Gly117 | R114 → A: Loss of activity<br>Y115 → A: Increases Km for S-adenosyl-L-methionine 24-fold.<br>G117 → A: Loss of activity. |
| Active site/Adenine | Ser131-Gly140 | D135 → A: Loss of activity; no effect on tRNA binding.<br>Y136 → A: Increases Km for S-adenosyl-L-methionine 68-fold. |
| Riobose binding pocket | Leu87,Tyr86,Arg114,Gly117,Ser88 | R114 → A: Loss of activity. |
| Salt-bridge | Arg220-Asp135 | D135 → A: Loss of activity; no effect on tRNA binding. |
| Salt-bridge | Arg114-Glu142 | R114 → A: Loss of activity. |
| Salt-bridge | Arg121-Asp119 | |
| G37-Adinine binding site | Glu116,Asp169,Arg154 | |

**Table 4.** Overview of our pose prediction performance using CSARdock 2014 benchmark exercise phase 2 dataset.

| | Total number of ligands docked | Number of crystal structures from CSAR | Median RMSD Å | Mean RMSD Å | Median RMSD from All groups Å | Mean RMSD from all groups Å |
|---|---|---|---|---|---|---|
| FXa | 163 | 3 | 6.03 | 6.03 | 2.94 | 2.98 |
| SYK | 276 | 8 | 4.28 | 4.01 | 2.58 | 2.46 |
| TrmD | 31 | 14 | 3.20 | 2.53 | 3.84 | 3.52 |

**Table 5.** Correlation between experimental activities and our scoring using CSARdock 2014 benchmark exercise phase 2 dataset.[a]

| Target proteins | # of inhibitors | Correlation Coefficient (r2) | Pearson r | Spearman ρ | Kendall τ |
|---|---|---|---|---|---|
| fxa_gtc000401_2.07 | 163 | 0.001 | 0.028 (0.181 to -0.127) | 0.067 (0.218 to -0.088) | 0.045 (0.148 to -0.059) |
| gtc000224_SYK | 276 | 0.013 | -0.112 (0.007 to -0.228) | -0.186 (-0.068 to -0.299) | -0.122 (-0.043 to -0.200) |
| gtc000249_SYK | 276 | 0.002 | 0.046 (0.164 to -0.074) | 0.218 (0.329 to 0.100) | 0.15 (0.228 to 0.072) |
| TrmD_445 | 31 | 0.045 | -0.211 (0.155 to -0.526) | 0.225 (0.540 to -0.145) | 0.177 (0.417 to -0.064) |
| TrmD_447 | 31 | 0.006 | -0.077 (0.285 to -0.420) | -0.11 (0.255 to -0.448) | -0.063 (0.184 to -0.310) |
| TrmD_448 | 31 | 0.001 | -0.031 (0.327 to -0.381) | -0.273 (0.096 to -0.577) | -0.159 (0.083 to -0.401) |
| TrmD_452 | 31 | 0.000 | 0.005 (0.359 to -0.350) | 0.317 (0.609 to -0.051) | 0.238 (0.472 to 0.003) |
| TrmD_453 | 31 | 0.164 | -0.406 (-0.060 to -0.664) | 0.152 (0.482 to -0.216) | 0.098 (0.344 to -0.148) |
| TrmD_456 | 31 | 0.043 | -0.207 (0.159 to -0.523) | -0.248 (0.122 to -0.558) | -0.159 (0.083 to -0.401) |
| TrmD_459 | 31 | 0.041 | -0.202 (0.164 to -0.519) | 0.202 (0.522 to -0.168) | 0.181 (0.421 to -0.059) |
| TrmD_460 | 31 | 0.015 | -0.122 (0.243 to -0.456) | 0.251 (0.560 to -0.119) | 0.185 (0.425 to -0.054) |
| TrmD_464 | 31 | 0.133 | -0.365 (-0.013 to -0.637) | 0.117 (0.454 to -0.249) | 0.076 (0.323 to -0.170) |
| TrmD_465 | 31 | 0.049 | -0.221 (0.145 to -0.533) | -0.4 (-0.039 to -0.669) | -0.233 (0.001 to -0.468) |

[a] 95% confidence interval in parentheses.

**Table 6.** AUC Values Derived from ROC Curves for All Ligands and Clusters of FXa and SYK.[a]

| | cutoff[b]_6.0 | cutoff[b]_6.5 | cutoff[b]_7.0 | cutoff[b]_7.5 |
|---|---|---|---|---|
| fxa_gtc000401_2.07_all | 0.44 (140,23) | 0.44 (117,46) | 0.44 (117,46) | 0.47 (93,70) |
| fxa_gtc000401_2.07_orange | - (49,0) | 0.65 (48,1) | 0.55 (39,10) | 0.58 (26,13) |
| fxa_gtc000401_2.07_brown4 | - (2,0) | 1 (1,1) | 1 (1,1) | - (0,2) |
| fxa_gtc000401_2.07_tan | 0.62 (31,12) | 0.65 (12,31) | 0.68 (6,37) | 0.7 (5,38) |
| fxa_gtc000401_2.07_linen | 0.75 (2,6) | 0.75 (2,6) | 0.75 (2,6) | -(0,6) |
| fxa_gtc000401_2.07_darkorchid4 | -(4,0) | -(4,0) | -(4,0) | -(4,0) |
| fxa_gtc000401_2.07_paleturquoise 3 | - (0,6) | - (0,6) | -(0,6) | -(0,6) |
| fxa_gtc000401_2.07_coral2 | -(22,0) | -(22,0) | -(22,0) | 0.57 (21,1) |
| fxa_gtc000401_2.07_coral3 | -(17,0) | -(17,0) | 0.21 (11,6) | 0.53 (5,12) |
| fxa_gtc000401_2.07_deeppink | 0 (10,1) | 0.63 (8,3) | 0.63 (5,6) | 0.18 (4,7) |
| fxa_gtc000401_2.07_goldenrod3 | .0(3,0) | .0(3,0) | .0(3,0) | .0(3,0) |
| | | | | |
| gtc000224_SYK_all | 0.53 (258,14) | 0.55 (226,46) | 0.60 (180,92) | 0.60 (136,136) |
| gtc000224_SYK_brown | 0.47 (98,9) | 0.42 (69,29) | 0.45 (34,64) | 0.41 (11,87) |
| gtc000224_SYK_royalblue | 0.27 (10,3) | 0.33 (3,10) | 0.17 (1,12) | 0.17 (1,12) |
| gtc000224_SYK_blue | 0.72 (157,1) | 0.67 (154,4) | 0.55 (144,14) | 0.51 (122,34) |
| | | | | |
| gtc000249_SYK_all | 0.28 (258,14) | 0.29(226,46) | 0.40(180,92 ) | 0.44(136,136 ) |
| gtc000249_SYK_brown | 0.27 (98,9) | 0.29(69,29) | 0.38 (34,64) | 0.43 (11,87) |
| gtc000249_SYK_royalblue | 0.4 (10,3) | 0.30 (3,10) | 0.17 (1,12) | 0.17 (1,12) |
| gtc000249_SYK_blue | 0.52 (157,1) | 0.67 (154,4) | 0.55 (144,14) | 0.52(122,34) |

[a.] The number of active and nonactive ligands at the cut-off are given in parentheses. [b] Cut-off is an adjustable parameter for (9-pIC50) values based on which actives and inactives are defined

**Table 7.** AUC Values Derived from ROC Curves for All Ligands and Clusters of TrmD docked to multiple crystal structures.[a]

| | cutoff[b]_5.5 | cutoff[b]_6.0 | cutoff[b]_6.5 | cutoff[b]_7.0 | cutoff[b]_7.5 |
|---|---|---|---|---|---|
| TrmD_ 445_all | 0.52 (18,13) | 0.56 (12,19) | 0.55 (5,26) | 0.76 (2,29) | 0.76 (2,29) |
| TrmD_ 445 _lightpink3 | 0.64 (11,4) | 0.55 (7,8) | 0.27 (2,13) | -(0,15) | -(0,15) |
| TrmD_ 445 _orangered1 | -(0,2) | -(0,2) | -(0,2) | -(0,2) | -(0,2) |
| TrmD_ 445 _grey | 0.18 (4,7) | 0.28 (2,9) | -(0,11) | -(0,11) | -(0,11) |
| TrmD_ 445 _chartreuse3 | -(3,0) | -(3,0) | -(3,0) | 0.5 (2,1) | 0.5 (2,1) |
| TrmD_ 447_all | 0.69 (18,13) | 0.56 (12,19) | 0.45 (5,26) | 0.62 (2,29) | 0.62 (2,29) |
| TrmD_ 447 _lightpink3 | 0.7 (11,4) | 0.41 (7,8) | 0.31 (2,13) | -(0,15) | -(0,15) |
| TrmD_ 447 _orangered1 | -(0,2) | -(0,2) | -(0,2) | -(0,2) | -(0,2) |
| TrmD_ 447 _grey | 0.54 (4,7) | 0.56 (2,9) | -(0,11) | -(0,11) | -(0,11) |
| TrmD_ 447 _chartreuse3 | -(3,0) | -(3,0) | -(3,0) | 1 (2,1) | 1 (2,1) |
| TrmD_ 448_all | **0.69 (18,13)** | **0.56 (12,19)** | **0.45 (5,26)** | **0.62 (2,29)** | **0.62 (2,29)** |
| TrmD_ 448 _lightpink3 | 0.71 (11,4) | 0.41 (7,8) | 0.31 (2,13) | -(0,15) | -(0,15) |
| TrmD_ 448 _orangered1 | -(0,2) | -(0,2) | -(0,2) | -(0,2) | -(0,2) |
| TrmD_ 448 _grey | 0.54 (4,7) | 0.56 (2,9) | -(0,11) | -(0,11) | -(0,11) |
| TrmD_ 448 _chartreuse3 | -(3,0) | -(3,0) | -(3,0) | 1 (2,1) | 1 (2,1) |
| TrmD_ 452_all | 0.21 (18,13) | 0.21 (12,19) | 0.61 (5,26) | 0.90 (2,29) | 0.90 (2,29) |
| TrmD_ 452 _lightpink3 | 0.18 (11,4) | 0.57 (7,8) | 0.35 (2,13) | -(0,15) | -(0,15) |
| TrmD_ 452 _orangered1 | -(0,2) | -(0,2) | -(0,2) | -(0,2) | -(0,2) |
| TrmD_ 452 _grey | 0 (4,7) | 0.07 (2,9) | -(0,11) | -(0,11) | -(0,11) |
| TrmD_ 452 _chartreuse3 | -(3,0) | -(3,0) | -(3,0) | 0.5 (2,1) | 0.5 (2,1) |
| TrmD_453_all | 0.36 (18,13) | 0.41 (12,19) | 0.69 (5,26) | 0.97 (2,29) | 0.97 (2,29) |
| TrmD_453_lightpink3 | 0.25 (11,4) | 0.48 (7,8) | 0.54 (2,13) | -(0,15) | -(0,15) |
| TrmD_453_orangered1 | -(0,2) | -(0,2) | -(0,2) | -(0,2) | -(0,2) |
| TrmD_453_grey | 0.36 (4,7) | 0 (2,9) | -(0,11) | -(0,11) | -(0,11) |
| TrmD_453_chartreuse3 | -(3,0) | -(3,0) | -(3,0) | 1 (2,1) | 1 (2,1) |
| TrmD_456_all | 0.59 (18,13) | 0.64 (12,19) | 0.64 (5,26) | 0.60 (2,29) | 0.60 (2,29) |
| TrmD_456_lightpink3 | 0.59 (11,4) | 0.64 (7,8) | 0.54 (2,13) | -(0,15) | -(0,15) |
| TrmD_456_orangered1 | -(0,2) | -(0,2) | -(0,2) | -(0,2) | -(0,2) |
| TrmD_456_grey | 0.43 (4,7) | 0.33 (2,9) | -(0,11) | -(0,11) | -(0,11) |
| TrmD_456_chartreuse3 | -(3,0) | -(3,0) | -(3,0) | 1 (2,1) | 1 (2,1) |
| TrmD_459_all | 0.29 (18,13) | 0.44 (12,19) | 0.63 (5,26) | 0.86 (2,29) | 0.86 (2,29) |
| TrmD_459_lightpink3 | 0.16 (11,4) | 0.43 (7,8) | 0.38 (2,13) | -(0,15) | -(0,15) |
| TrmD_459_orangered1 | -(0,2) | -(0,2) | -(0,2) | -(0,2) | -(0,2) |
| TrmD_459_grey | 0.21 (4,7) | 0.06 (2,9) | -(0,11) | -(0,11) | -(0,11) |
| TrmD_459_chartreuse3 | -(3,0) | -(3,0) | -(3,0) | 0 (2,1) | 0 (2,1) |

| | | | | | |
|---|---|---|---|---|---|
| TrmD_460_all | 0.39 (18,13) | 0.34 (12,19) | 0.6 (5,26) | 0.88 (2,29) | 0.88 (2,29) |
| TrmD_460_lightpink3 | 0.61 (11,4) | 0.34 | 0.35 (2,13) | -(0,15) | -(0,15) |
| TrmD_460_orangered1 | -(0,2) | -(0,2) | -(0,2) | -(0,2) | -(0,2) |
| TrmD_460_grey | 0.18 (4,7) | 0 (2,9) | -(0,11) | -(0,11) | -(0,11) |
| TrmD_460_chartreuse3 | -(3,0) | -(3,0) | -(3,0) | 0.5 (2,1) | 0.5 (2,1) |
| TrmD_464_all | 0.38 (18,13) | 0.51 (12,19) | 0.55 (5,26) | 0.97 (2,29) | 0.97 (2,29) |
| TrmD_464_lightpink3 | 0.34 (11,4) | 0.64 (7,8) | 0.35 (2,13) | -(0,15) | -(0,15) |
| TrmD_464_orangered1 | -(0,2) | -(0,2) | -(0,2) | -(0,2) | -(0,2) |
| TrmD_464_grey | 0.39 (4,7) | 0.17 | -(0,11) | -(0,11) | -(0,11) |
| TrmD_464_chartreuse3 | -(3,0) | -(3,0) | -(3,0) | 1 (2,1) | 1 (2,1) |
| TrmD_465_all | 0.68 (18,13) | 0.75 (12,19) | 0.58 (5,26) | 0.48 (2,29) | 0.48 (2,29) |
| TrmD_465_lightpink3 | 0.43 (11,4) | 0.61 (7,8) | 0.69(2,13) | -(0,15) | -(0,15) |
| TrmD_465_orangered1 | -(0,2) | -(0,2) | -(0,2) | -(0,2) | -(0,2) |
| TrmD_465_grey | 0.68 (4,7) | 1 (2,9) | -(0,11) | -(0,11) | -(0,11) |
| TrmD_465_chartreuse3 | -(3,0) | -(3,0) | -(3,0) | 1 (2,1) | 1 (2,1) |

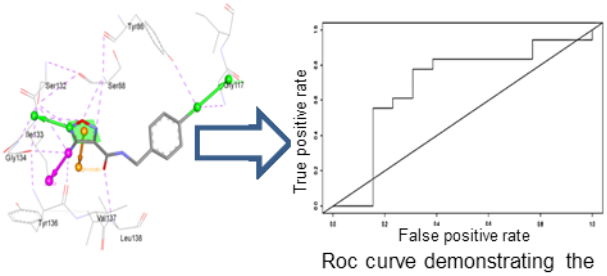
a. The number of active and nonactive ligands at the cut-off is given in parentheses.[b] Cut-off is an adjustable parameter for (9-pIC50) values based on which actives and inactives are defined

**REFERENCES**


1. Prathipati, P.; Dixit, A.; Saxena, A. K., Computer Aided Drug Design: Integration of Structure Based and Ligand Based Approaches. *Curr Comput Aided Drug Des* **2007**, *92*, 29-37, doi: 10.2174/157340907780809516.
2. Saxena, A. K.; Prathipati, P., Collection and Preparation of Molecular Databases for Virtual Screening. *SAR QSAR Environ Res* **2006**, *17*, 371-92.
3. Prathipati, P.; Pandey, G.; Saxena, A. K., CoMFA and Docking Studies on Glycogen Phosphorylase a Inhibitors as Antidiabetic Agents. *J Chem Inf Model* **2005**, *45*, 136-45.
4. Prathipati, P.; Saxena, A. K., Evaluation of Binary QSAR Models Derived from LUDI and MOE Scoring Functions for Structure based Virtual Screening. *J Chem Inf Model* **2006**, *46*, 39-51.
5. Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B., Jr.; Stuckey, J. A.; Carlson, H. A., CSAR Benchmark Exercise 2011-2012: Evaluation of Results From Docking and Relative Ranking of Blinded Congeneric Series. *J Chem Inf Model* **2013**, *53*, 1853-70.
6. Dunbar, J. B., Jr.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y. N.; Kubish, G.; Gestwicki, J. E.; Stuckey, J. A.; Carlson, H. A., CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. *J Chem Inf Model* **2013**, *53*, 1842-52.
7. Koes, D. R.; Baumgartner, M. P.; Camacho, C. J., Lessons Learned in Empirical Scoring With Smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model* **2013**, *53*, 1893-904.
8. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J., Molecular Similarity in Medicinal Chemistry. *J Med Chem* **2014**, 57, 3186-204.
9. Stumpfe, D.; Dimova, D.; Bajorath, J., Composition and Topology of Activity Cliff Clusters Formed by Bioactive Compounds. *J Chem Inf Model* **2014**, *54*, 451-61.
10. Langfelder, P.; Horvath, S., WGCNA: an R package for Weighted Correlation Network Analysis. *BMC Bioinformatics* **2008**, 9, 559, doi: 10.1186/1471-2105-9-559.
11. Zhang, B.; Horvath, S., A General Framework for Weighted Gene Co-expression Network Analysis. *Stat Appl Genet Mol Biol* **2005**, *4*, Article17, doi:10.1.1.471.9599
12. Rivera-Borroto, O. M.; Marrero-Ponce, Y.; Garcia-de la Vega, J. M.; Grau-Abalo Rdel, C., Comparison of Combinatorial Clustering Methods on Pharmacological Data Sets Represented by Machine Learning-selected Real Molecular Descriptors. *J Chem Inf Model* **2011**, *51*, 3036-49.
13. Matlock, M. K.; Zaretzki, J. M.; Swamidass, S. J., Scaffold Network Generator: a Tool for Mining Molecular Structures. *Bioinformatics* **2013**, *29*, 2655-6.
14. Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S., Probing the Bioactivity-relevant Chemical Space of Robust Reactions and Common Molecular Building Blocks. *J Chem Inf Model* **2012**, *52*, 1167-78.
15. Jacoby, E.; Tresadern, G.; Bembenek, S.; Wroblowski, B.; Buyck, C.; Neefs, J. M.; Rassokhin, D.; Poncelet, A.; Hunt, J.; van Vlijmen, H., Extending Kinome Coverage by Analysis of Kinase Inhibitor Broad Profiling Data. *Drug Discov Today* **2015,** *20***,** 652-8.
16. Ghersi, D.; Singh, M., molBLOCKS: Decomposing Small Molecule Sets and Uncovering Enriched Fragments. *Bioinformatics* **2014**, *30*, 2081-3.
17. Hinselmann, G.; Rosenbaum, L.; Jahn, A.; Fechner, N.; Zell, A., jCompoundMapper: An Open Source Java Library and Command-line Tool for Chemical Fingerprints. *J Cheminform* **2011**, *3*, 3, doi: 10.1186/1758-2946-3-3.

18. Meslamani, J.; Li, J.; Sutter, J.; Stevens, A.; Bertrand, H. O.; Rognan, D., Protein-ligand-based Pharmacophores: Generation and Utility Assessment in Computational Ligand Profiling. *J Chem Inf Model* **2012**, 52, 943-55.
19. Husby, J.; Bottegoni, G.; Kufareva, I.; Abagyan, R.; Cavalli, A., Structure-based Predictions of Activity Cliffs. *J Chem Inf Model* **2015**, *55*, 1062-76.
20. Rueda, M.; Bottegoni, G.; Abagyan, R., Consistent Improvement of Cross-docking Results using Binding Site Ensembles Generated with Elastic Network Normal Modes. *J Chem Inf Model* **2009**, *49*, 716-25.
21. Rueda, M.; Bottegoni, G.; Abagyan, R., Recipes for the Selection of Experimental Protein Conformations for Virtual screening. *J Chem Inf Model* **2010**, *50*, 186-93.
22. Doncheva, N. T.; Klein, K.; Domingues, F. S.; Albrecht, M., Analyzing and Visualizing Residue Networks of Protein Structures. *Trends Biochem Sci* **2011**, *36*, 179-82.
23. Barillari, C.; Marcou, G.; Rognan, D., Hot-spots-guided Receptor-based Pharmacophores (HS-Pharm): a Knowledge-based Approach to Identify Ligand-anchoring Atoms in Protein Cavities and Prioritize Structure-based Pharmacophores. *J Chem Inf Model* **2008**, 48, 1396-410.
24. Gibbs, A. C., Elements and Modulation of Functional Dynamics. *J Med Chem* **2014**, *57*, 7819-37.
25. Chan, C.; Borthwick, A. D.; Brown, D.; Burns-Kurtis, C. L.; Campbell, M.; Chaudry, L.; Chung, C. W.; Convery, M. A.; Hamblin, J. N.; Johnstone, L.; Kelly, H. A.; Kleanthous, S.; Patikis, A.; Patel, C.; Pateman, A. J.; Senger, S.; Shah, G. P.; Toomey, J. R.; Watson, N. S.; Weston, H. E.; Whitworth, C.; Young, R. J.; Zhou, P., Factor Xa Inhibitors: S1 binding Interactions of a Series of N-{(3S)-1-[(1S)-1-methyl-2-morpholin-4-yl-2-oxoethyl]-2-oxopyrrolidin-3-yl}sulfon Amides. *J Med Chem* **2007**, *50*, 1546-57.
26. Nar, H., The Role of Structural Information in the Discovery of Direct Thrombin and Factor Xa inhibitors. *Trends Pharmacol Sci* **2012**, *33*, 279-88.
27. Gradler, U.; Schwarz, D.; Dresing, V.; Musil, D.; Bomke, J.; Frech, M.; Greiner, H.; Jakel, S.; Rysiok, T.; Muller-Pompalla, D.; Wegener, A., Structural and Biophysical Characterization of the Syk Activation Switch. *J Mol Biol* **2013**, *425*, 309-33.
28. Ahn, H. J.; Kim, H. W.; Yoon, H. J.; Lee, B. I.; Suh, S. W.; Yang, J. K., Crystal structure of tRNA(m1G37)methyltransferase: Insights into tRNA Recognition. *EMBO J* **2003**, *22*, 2593-603.
29. Christian, T.; Hou, Y. M., Distinct Determinants of tRNA Recognition by the TrmD and Trm5 Methyl Transferases. *J Mol Biol* **2007**, *373*, 623-32.
30. Guner, O. F., History and Evolution of the Pharmacophore Concept in Computer-aided Drug Design. *Curr Top Med Chem* **2002**, *2*, 1321-32.
31. Prathipati, P.; Saxena, A. K., Characterization of Beta3-adrenergic Receptor: Determination of Pharmacophore and 3D QSAR model for Beta3 Adrenergic Receptor Agonism. *J Comput Aided Mol Des* **2005**, *19*, 93-110.
32. Wolber, G.; Seidel, T.; Bendix, F.; Langer, T., Molecule-pharmacophore Superpositioning and Pattern Matching in Computational Drug Design. *Drug Discov Today* **2008**, *13*, 23-9.
33. Acharya, C.; Coop, A.; Polli, J. E.; Mackerell, A. D., Jr., Recent Advances in Ligand-based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. *Curr Comput Aided Drug Des* **2011**, *7*, 10-22.
34. Genheden, S.; Nilsson, I.; Ryde, U., Binding Affinities of Factor Xa Inhibitors Estimated by Thermodynamic Integration and MM/GBSA. *J Chem Inf Model* **2011**, *51*, 947-58.

35. Smith, R. D.; Dunbar, J. B., Jr.; Ung, P. M.; Esposito, E. X.; Yang, C. Y.; Wang, S.; Carlson, H. A., CSAR benchmark exercise of 2010: Combined Evaluation Across all Submitted Scoring Functions. *J Chem Inf Model* **2011**, *51*, 2115-31.
36. Meslamani, J.; Bhajun, R.; Martz, F.; Rognan, D., Computational Profiling of Bioactive Compounds Using a Target-dependent Composite Workflow. *J Chem Inf Model* **2013**, *53*, 2322-33.
37. Chiu, Y. Y.; Lin, C. T.; Huang, J. W.; Hsu, K. C.; Tseng, J. H.; You, S. R.; Yang, J. M., KIDFamMap: a Database of Kinase-inhibitor-disease Family Maps for Kinase Inhibitor Selectivity and Binding Mechanisms. *Nucleic Acids Res* **2013**, *41*, D430-40.
38. Kufareva, I.; Abagyan, R., Type-II kinase Inhibitor Docking, Screening, and Profiling using Modified Structures of Active Kinase States. *J Med Chem* **2008**, *51*, 7921-32.
39. Fourches, D.; Politi, R.; Tropsha, A., Target-specific Native/decoy Pose Classifier Improves the Accuracy of Ligand Ranking in the CSAR 2013 benchmark. *J Chem Inf Model* **2015**, *55*, 63-71.
40. Copeland, R. A., Evaluation of Enzyme Inhibitors in Drug Discovery. A Guide for Medicinal Chemists and Pharmacologists. *Methods Biochem Anal* **2005**, *46*, 1-265.
41. Miller, J. A.; Cai, C.; Langfelder, P.; Geschwind, D. H.; Kurian, S. M.; Salomon, D. R.; Horvath, S., Strategies for Aggregating Gene Expression Data: the collapseRows R function. *BMC Bioinformatics* **2011**, *12*, 322, doi: 10.1186/1471-2105-12-322.
42. Dror, O.; Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J., Novel Approach for Efficient Pharmacophore-based Virtual Screening: Method and Applications. *J Chem Inf Model* **2009**, *49*, 2333-43.
43. Miteva, M. A.; Guyon, F.; Tuffery, P., Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res* **2010**, *38*, W622-7.
44. Koes, D. R.; Camacho, C. J., Pharmer: Efficient and Exact Pharmacophore Search. *J Chem Inf Model* **2011**, *51*, 1307-14.
45. Bhunia, S. S.; Roy, K. K.; Saxena, A. K., Profiling the structural determinants for the selectivity of representative factor-Xa and thrombin inhibitors using combined ligand-based and structure-based approaches. *J Chem Inf Model* **2011**, *51*, 1966-85.




TrmD Protein structure based pharmacophore

Roc curve demonstrating the active versus inactive classification accuracy of 0.69

for Table of Contents use only