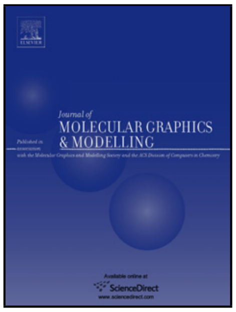
Accepted Manuscript



Title: Comparing Sixteen Scoring Functions for Predicting Biological Activities of Ligands for Protein Targets

Author: Weijun Xu Andrew J. Lucke David P. Fairlie

PII: S1093-3263(15)00028-5
DOI: http://dx.doi.org/doi:10.1016/j.jmgm.2015.01.009
Reference: JMG 6510

To appear in: *Journal of Molecular Graphics and Modelling*

Received date: 21-10-2014
Revised date: 22-1-2015
Accepted date: 23-1-2015

Please cite this article as: W. Xu, A.J. Lucke, D.P. Fairlie, Comparing Sixteen Scoring Functions for Predicting Biological Activities of Ligands for Protein Targets, *Journal of Molecular Graphics and Modelling* (2015), http://dx.doi.org/10.1016/j.jmgm.2015.01.009

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Comparing Sixteen Scoring Functions for Predicting Biological Activities of Ligands for Protein Targets

Weijun Xu, Andrew J. Lucke, David P. Fairlie*

Division of Chemistry and Structural Biology, Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

To whom correspondence should be addressed: Professor David Fairlie, Institute for Molecular Bioscience, University of Queensland, Brisbane, Qld 4072, Australia, Tel: +61-733462989; Fax: +61-73346 2990; E-mail: d.fairlie@imb.uq.edu.au

Abstract

Accurately predicting relative binding affinities and biological potencies for ligands that interact with proteins remains a significant challenge for computational chemists. Most evaluations of docking and scoring algorithms have focused on enhancing ligand affinity for a protein by optimizing docking poses and enrichment factors during virtual screening. However, there is still relatively limited information on the accuracy of commercially available docking and scoring software programs for correctly predicting binding affinities and biological activities of structurally related inhibitors of different enzyme classes. Presented here is a comparative evaluation of eight molecular docking programs (Autodock Vina, Fitted, FlexX, Fred, Glide, GOLD, LibDock, MolDock) using sixteen docking and scoring functions to predict the rank-order activity of different ligand series for six pharmacologically important protein and enzyme targets (Factor Xa, Cdk2 kinase, Aurora A kinase, COX-2, pla2g2a, β Estrogen receptor). Use of Fitted gave an excellent correlation (Pearson 0.86, Spearman 0.91) between predicted and experimental binding only for Cdk2 kinase inhibitors. FlexX and GOLDScore produced good correlations (Pearson > 0.6) for hydrophilic targets such as Factor Xa, Cdk2 kinase and Aurora A kinase. By contrast, pla2g2a and COX-2 emerged as difficult targets for scoring functions to predict ligand activities. Albeit possessing a high hydrophobicity in its binding site, β Estrogen receptor produced reasonable correlations using LibDock (Pearson 0.75, Spearman 0.68). These findings can assist medicinal chemists to better match scoring

Accepted Manuscript

functions with ligand-target systems for hit-to-lead optimization using computer-aided drug design approaches.

Keywords

Molecular docking; Scoring functions; Hydrophilic versus Hydrophobic targets; Drug design; Enzyme inhibitor

1. Introduction

Lead optimization is important for drug discovery and involves making substantial improvements in ligand specificity, potency and pharmacokinetic properties over weakly potent hits typically identified from virtual or high throughput screening. Lead development via chemical modification is often guided by available ligand SAR, 2D or 3D similarity-based fragment searches, 3D-pharmacophore model building and structure-based design. To accelerate lead optimization, reduce labor and minimize costs, reliable computational methods that accurately predict compound binding affinity and/or functional potency are highly desirable. A variety of approaches to calculate ligand binding affinity have been developed and reviewed[1, 2]. Molecular dynamic (MD) simulations, Monte Carlo (MC) simulations, free energy perturbation (FEP) and thermodynamic integration methods can all be used to calculate binding free energies that are similar to experimentally determined values[3-5]. MM/PBSA calculations, pioneered by Kollman and coworkers, use a combination of molecular mechanics and continuum solvation to compute binding free energies for the binding complexes between bound and unbound states[6]. A related approach, MM/GBSA, has been used in studies of protein-ligand interactions and applied to diverse targets[7, 8]. Although some encouraging results have been produced[9] from free energy calculations, these approaches are computationally expensive and impractical for routine evaluations of binding affinity predictions. Comparing ligands is therefore mainly done using molecular docking and scoring functions to identify and rank ligand binding poses in a binding pocket. Scoring functions rank each pose of a ligand relative to other poses typically that corresponding to a crystal structure. These scores are commonly used not only to rank individual ligand poses, but also to compare different ligand scores for identifying the potentially more potent ligands (some scoring functions produce a binding energy value). Computational methods

Accepted Manuscript

are useful tools in medicinal chemistry, but suffer from difficulties in predicting protein conformational changes and still require considerable further refinements to improve their effectiveness in drug design and ligand optimization strategies *in silico*[10].

In the last decade, evaluation of the performance of docking and scoring functions has focused predominantly on two measures. Firstly, it has sought accurate reproductions of co-crystalized ligand binding poses in protein crystal structures. Ligand docking is most accurate if the top ranked pose has a heavy atom root-mean-square deviation (RMSD) < 2.0 Å from the location of the crystalized ligand[11]; and this has been shown to be achievable with several common docking programs[12-14]. Software programs for ligand docking are constantly improving and can now achieve heavy atom rmsd values within 1 Å for some targets[15]. A second approach to validate docking and scoring algorithms involves examining the enrichment factor (EF) after virtual screening. The EF is defined as the accumulated ratio of active ligands found above a certain percentile of the ranked database containing active and inactive ligands. A higher EF value at a defined percentile (e.g. $EF^{2\%}$) usually indicates a better scoring function[11]; this measure has been used many times to evaluate scoring functions[16-19]. The area under the curve (AUC) of receiver operator[20] characteristics is usually employed to reflect the enrichment (CSAR 2011-2012)[21]. Scoring functions have also been evaluated for accuracy in predicting experimental binding affinity or biological activity. This is still challenging due to the reproducibility of ligand binding or activity data measured experimentally (often under different conditions) in different laboratories[11], and especially because some scoring functions lack terms such as solvation energy and configurational entropy which affect affinity of ligand binding[2], and uncertainties in protein conformations which are extraordinarily difficult to computationally predict at the present time.

A large number of docking and scoring comparisons have been reported, comparing RMSD values, EF values[12, 14-16, 19, 22-34] and less frequently predicting and ranking ligand binding affinity[35-38]. Wang et al. comparatively evaluated 11 scoring functions (four scoring functions in LigFit module in Cerius2: LigScore, PLP, PMF, and LUDI; four scoring functions implemented in CScore module in SyByl: F-Score, G-Score, D-Score, ChemScore, scoring functions in AutoDock program, and two standalone scoring functions: Drug-Score and X-Score)

Accepted Manuscript

ACCEPTED MANUSCRIPT

for effectiveness in molecular docking, by assessing their ability to reproduce experimentally determined binding conformations and affinities of 100 protein-ligand complexes[15]. Autodock was used to generate docking conformations and re-scored by other scoring functions. Results showed that six scoring functions achieved a success rate of 66%-76% using RMSD 2.0 Å as the chief criterion. However, only four scoring functions were able to give a ranking correlation of 0.5 – 0.7 when they were applied to predict the experimentally determined binding affinities for the protein-ligand complexes. Warren et al. evaluated 10 different docking programs incorporating 37 scoring functions against 8 proteins of 7 protein families with aproximately 1300 ligands; binding mode, virtual screening and binding affinity prediction were examined[19]. Nineteen docking protocols were able to predict accurate ligand conformations of 136 protein ligand complexes for which crystal structures were available. However, none of the scoring functions usefully predicted ligand-binding affinity. The study indicated that the goal of accurately predicting ligand affinities was beyond the capacity of all of the scoring functions at that time.

There have been relatively limited reports on comparisons of docking, scoring and binding affinity predictions on multiple defined series of congeneric compounds. A few representative examples are referred to herein. Pearlman and Charifson[39] examined a series of p38 MAP kinase inhibitors and found a good correlation between experimental ligand binding affinities determined via free energy grid calculations compared to Chemscore, PLPScore and Dock energy ligand scores. Lyne[40] accurately predicted relative inhibitory potencies of members of a series of kinase inhibitors (p38, Aurora A, Cdk2 and Jnk3) using molecular docking followed by MM-GBSA scoring (Pearson correlation: 0.71 – 0.84). Rapp et al.[41] applied a molecular mechanics approach when examining 12 protein targets with their congeneric inhibitors. Prime energy calculations were included in the scoring and produced good correlations between predicted binding scores and experimental binding affinities (r^2: 0.25 – 0.82). These reports suggest that the inclusion of MM-GBSA based scoring correlates well with ligand binding affinity. It is not clear how broadly applicable this method is though, as reports have generally examined only kinase proteins with a small number of congeneric ligands.

Recently, the Community Structure-Activity Resource (CSAR) conducted a blinded exercise in evaluating the docking and relative ranking of congeneric compounds against four different protein targets; 20 groups worldwide being invited

Accepted Manuscript

ACCEPTED MANUSCRIPT

to submit their hypothesis on the choice of the best scoring functions for both ligand docking and ranking[21]. It was found that relative ranking was the most difficult and most groups did not achieve a high correlation between computationally predicted ligand pose scores and experimental binding activity data. However, many docking programs were able to differentiate between active and inactive compounds against one target, the urokinase protein.

The current study is aimed at comparing the performance of several scoring functions from eight different molecular docking programs (commercially available and free trial versions) in predicting experimental biological activities of ligands for their protein targets. The scoring functions were applied to six pharmaceutically important protein targets each against a set of ligands for which biological activities have been reported in the literature. Table 1 summarizes these six target proteins, the number of ligands to be used for this computational study, the range of experimental inhibition constants covered by the ligand set, and the literature references from which the data was taken. We chose proteins considered to be difficult targets for ligand docking and for which experimental data on ligand binding affinity or protein inhibition was available based on similar experimental conditions. The aim of this study was to examine a variety of docking and scoring functions for their capacity to correctly predict relative rank order of biological activity or binding affinity of ligands to hydrophilic and hydrophobic protein targets. As well we wanted to examine whether possible correlations between predicted and experimental results were useful in “lead” optimization studies and to identify an optimized docking scoring protocol for virtual screening across different target proteins.

Table 1: Selected Literature Compounds

Target protein	Number of Compounds	Experimental data (pK_i and pIC_{50} range)	Reference
Factor Xa	33	5.8-10.9 (pK_i)	[42-45]
cdk2 kinase	24	5.3-8.3 (pIC_{50})	[46]
Aurora A Kinase	21	5.1-8.4 (pIC_{50})	[47]
COX-2	22	5.1-8.1 (pIC_{50})	[48-50]
pla2g2a	29	4.7-7.7 (pIC_{50})	[51]
β Estrogen Receptor	25	5.7-8.9 (pIC_{50})	[52]

2. Materials & Methods

2.1. Protein targets

Accepted Manuscript

Factor Xa*:* Factor Xa is a trypsin-like serine protease enzyme that is an important target for antithrombosis due to its role in the coagulation cascade[53]. The crystal structure shows the ligand binding site is a shallow solvent-exposed groove, except for a deep S1 pocket that prefers to bind positively charged or basic groups[43]. Factor Xa has been reported in several studies on scoring functions[19, 31, 41].

Cyclin-Dependent Kinase 2*:* The cyclin-dependent kinases (Cdks) are a family of serine-threonine protein kinases which control cell cycle proliferation in eukaryotic cells[54]. Abnormal activity of Cdks can lead to a loss of cell function checkpoints and are linked to cancer pathology,[55] and are cancer therapeutic targets[56]. The crystal structure of Cdk2 with a bound potent inhibitor: NU6102 shows two key hydrogen bonds are essential for strong binding[57]. This target has also been included in a few previous comparative assessments of scoring functions[5, 24, 40, 41].

Aurora A kinase: Aurora A kinase is a member of the Aurora family of serine/threonine kinase enzymes[58, 59]. It is a key regulator of mitosis in eukaryotic cells and has been shown to be strongly involved in the onset and progression of cancer[60, 61]. Aurora A is over-expressed in human cancers such as pancreatic, breast, colon and ovarian tumors. The search for new inhibitors of Aurora A kinase has been driven by clinical success of current inhibitors in oncological studies[62-65]. Aurora A has a hydrophilic binding site, containing charged amino acids which form salt bridges to ligands[47].

COX-2*:* Cyclooxygenase-2 is an enzyme involved in the synthesis of eicosanoids from C_{20} polyunsaturated fatty acids in the cyclooxygenase pathways[66]. Over-expression of COX-2 is usually responsible for production of pro-inflammatory prostaglandins. Hence, COX-2 is an attractive target for drug design to combat inflammatory diseases and physiological disorders. The active site of COX-2 contains mainly hydrophobic residues[67].

sPLA2*:* Human secretory phospholipases A2 (sPLA2) are enzymes that catalyze the hydrolysis of the 2-acyl ester of 3-sn-phosphoglycerides to produce arachidonic acid and lysophospholipid. The arachidonate is then metabolized to eicosanoids by cyclooxygenase and lipoxygenase and the later is converted to platelet activating factor[68]. Human sPLA2 group IIa (pla2g2a) has been shown in abnormally high concentrations in synovial fluid from patients with rheumatoid and

Accepted Manuscript

ACCEPTED MANUSCRIPT

osteoarthritis[51]. A high level of pla2g2a has been found to be associated with the severity of arthritis and sepsis[51]. The crystal structure[51] of pla2g2a revealed that the active site is lined by a series of hydrophobic residues Phe5, Ile9, Ala18, Ala19, Try22, Gly23 and Cys45.

β ***Estrogen Receptor****:* Estrogens belong to a family of naturally occurring steroid hormones that mediate the growth, development and maintenance of different tissues in human body[52]. The action of estrogen on different cell types is mediated via estrogen receptors that are members of a superfamily of nuclear receptors that play a role as ligand-activated gene transcription factors. There are two types of estrogen receptors: ERα and ERβ. Although widely expressed in many tissues, ERα is found mainly in uterus, kidney, and ovarian theca cells, whereas ERβ is predominantly expressed in ovarian granulosa cells, lung, bladder, and prostate[52]. Selective ERβ ligands have been found to have utility in treatment of diseases such as inflammatory bowel disease and rheumatoid arthritis[52].

2.2 Preparation of Protein Structures

Target protein crystal structures for Factor Xa (pdb code: 2P16), cdk2 kinase (pdb code: 1H1S), Aurora kinase A (pdb code: 3D14), COX-2 (pdb code: 6COX), Estrogen receptor (pdb code: 1YY4) and Pla2g2a (pdb code: 1J1A) were chosen as their co-crystalized ligands had a corresponding identical or similar ligand in the congeneric ligand set; crystal structures were appropriate for docking with resolution values <3Å and R-values <0.3. Structures were retrieved from the Protein Databank[69, 70] (www.rscb.org) and coordinates of chain “A” from each protein were imported into Maestro (Schrödinger software version 9.4) interface and then prepared using the Protein Preparation Wizard. Missing side chains and hydrogens were added, bond orders were corrected, and disulfide and zero order bonds to metals were created. Remote metal ions not involved in ligand binding were removed, since we considered that their stabilization roles were unlikely to affect ligand docking. H-bond assignments, tautomer and protonation states of amino acids at pH 7.4, were optimized. The prepared structures were then saved for use in docking programs that did not internally prepare proteins (e.g. GOLD).

2.3 SiteMap Calculation for Hydrophobicity of Protein Binding Sites

Accepted Manuscript

ACCEPTED MANUSCRIPT

SiteMap is a tool that defines putative binding sites by analyzing several parameters contributing to binding between a ligand and its receptor[71]. Parameters included in calculations are: site score, size, exposure score, contact, hydrophobic/hydrophilic property[72]. Once protein targets were prepared, the program SiteMap (Schrödinger software version 9.4) was used to evaluate and quantify the hydrophobic and hydrophilic nature of the binding site. Default parameters were used with a single binding site defined as the region of 6 Å about the binding ligand atoms.

2.4 Test Compounds

Compounds for target proteins were selected from each particular research group, either in an original research paper or several papers published on the same target, to ensure consistency of experimental conditions used to determine biological activities. Each compound series contained at least twenty ligands. In addition, except for the COX2 compound set, at least one compound belonging to the series had been co-crystallized with the target protein. Table 1 lists the reference for each compound series, the number of compounds, and the range of the experimental data. When pK_i was not reported, pIC_{50} was used based on a general premise that compounds sharing a similar scaffold should bind to the protein at a site similar to the one identified in the crystal structure. pK_i or pIC_{50} of the compounds spanned a magnitude of at least four fold for biological activities of the compounds.

2.5. Preparation of Ligands

Structures for all ligands were drawn in ChemBioDraw13.0 as a neutral species with the correct stereochemistry and then saved as a 2D sdf file. LigPrep in Schrödinger Suite software (version 9.4) was then used to convert the 2D sdf files into 3D maestro and sdf files. LigPrep generated a single 3D structure per ligand with that was minimized using the OPLS2005 force field and protonation state corrected to pH 7.4 using Epik.

2.6. Molecular Docking:

GOLD: GOLD[73] uses a genetic algorithm and takes into account partial receptor flexibility with full ligand flexibility during conformational searches and docking. Each ligand conformation is analogously encoded as evolution of a

Accepted Manuscript

ACCEPTED MANUSCRIPT

population of possible solutions via genetic operators (viz. mutations, crossovers and migrations) to a final population. The degree of freedom of the ligand is represented as binary strings called genes. These genes make up the “chromosome” which reflects ligand binding pose. In GOLD, the docking site was defined by a search radius of 15 Å around Asp 48 in Factor Xa, 10 Å around Phe 80 in cdk2 kinase, 10 Å around Glu 194 in aurora A kinase, 10 Å around Phe 518 in COX-2, 10 Å around Asp 48 in sPLA2, and 10 Å around Leu 298 in β estrogen receptor. Default parameters were applied with 100% ligand search efficiency. All other parameters were set as default. Each ligand was docked for 10 GA runs but the top 3 poses were saved as final solutions.

GLIDE: Glide[13] uses a series of hierarchical filters to search for possible locations of a ligand in the binding site using a pre-defined grid representation of the rigid receptor. The grid-enclosing box was placed on the centroid of a selected amino acid in the binding site and all other residues within 14 Å were included in considering the binding site. The scaling factor was set to 0.8 according to the default setting and GLIDE was run in extra precision (XP) mode with 10 poses per ligand kept. Docked poses from GLIDE XP were submitted to a PRIME/MM-GBSA calculation using default parameters to determine binding free energies between ligands and receptor. MM-GBSA, energies were estimated based on OPLS-AA force field for molecular mechanics energy (EMM) and the surface-generalised borne model for polar solvation energy, and a non-polar solvation term were also taken into account[74].

FlexX: FlexX is one of the most frequently used docking software programs. It is based on an incremental fragment-based docking approach developed from the Leach and Kuntz algorithm[75]. During the docking process, the whole ligand is broken into small fragments. All base fragments generated from a given ligand serve as starting point for docking[76]. The complete ligand is constructed and mapped into the protein active site after placement of a single base fragment by taking into account entropy, hydrogen bonds, metal acceptor, amide, methyl and aromatic ring[31]. In the current study, the FlexX package was part of the software package LeadIT (BioSolveIT GmbH). For FlexX, the docking set up was prepared according to standard workflow and the binding site was defined as 6.5 Å around the ligand in the crystal structure.

Accepted Manuscript

ACCEPTED MANUSCRIPT

Autodock Vina: Autodock tools were used to convert the Schrödinger prepared target protein pdb files to the Autodock Vina required pdbqt file type. Ligand sdf files were converted to pdb files using OpenBabel and converted to Autodock Vina required pdqt files using Autodock tools. Autodock Vina[77] uses a grid-based approach with the center of the search set as a 20 Å box about the center of the protein bound ligand. Vina search exhaustiveness was set to ten and ten dockings per ligand were performed.

Fitted: FITTED Suite 3.6[78] was used for molecular docking; files were prepared and docking procedures performed as described in the user guide using default parameters unless noted. The grid center for docking was defined by automatic search using the center of the crystallized ligand. The grid size was retained as the default parameters (15 Å) in Fitted. FITTED used a GA based docking approach to dock ligands into a binding site defined as spheres and used RankScore as scoring function. Initially, PREPARE was used to download and prepare the target protein adding hydrogens, optimizing tautomers and water molecules. SMART was used to prepared ligands, ProCESS to setup the proteins for docking and FITTED used to perform the docking. FITTED docked ligands three times by default using the default rigid protein.

Molegro: Molegro Virtual Docker 6.0 (MVD) was used for the preparation of ligand and protein files and for docking with MolDock[79]. MolDock used a hybrid guided differential evolution (DE) algorithm combined with a cavity prediction algorithm for ligand docking. The MolDock scoring function was based on a piecewise linear potential (PLP) modified to take into account H-bond directionality. Top ranked poses were re-ranked using a more complex scoring function that added an sp2-sp2 torsion term and a Lennard-Jones potential term to the score. Protein and ligand files were prepared and the docking performed as described in the Docking Tutorial in the MVD manual. The docking site was set by choosing the bound ligand in the crystal structure and a radius of 15 Å was applied. Docking was run with 10 poses per ligand, with similar poses within 1 Å RMSD being ignored.

Fred: Fred[80] was supplied as part of the OpenEye suite of programs, it docks a multi-conformer library of ligands into the binding site using an exhaustive search algorithm that systematically searches rotations and translations of the conformers with in the binding site. The default scoring function used by Fred is Chemgauss4 a shape based complementarity score between the ligand pose and

Accepted Manuscript

binding site. Docking was performed as described in the OpenEye OEDocking[81] manual using the default parameters unless noted. Omega[82, 83] was used with default settings to generate a library of 200 conformers per ligand for docking. Receptor files were prepared by reading the Maestro prepared pdb files into the make_receptor GUI supplied with Fred. A 20 Å box was centred on the co-crystalized ligand to define the binding site, the shape potential of the binding site was defined as balanced, no constraints were used. Fred was then used to dock the multi-conformer ligand library into the protein receptor file with poses scored by Fred Chemgauss4 score.

Hybrid: Hybrid[84] was supplied as part of the OpenEye suite of programs. Hybrid pose scoring takes into account ligand similarity during the docking process. Protein and ligand file preparation as well as docking were performed in a similar manner to that described for the Fred docking program. Like Fred, Hybrid uses an exhaustive search algorithm that systematically searches rotations and translations of the ligand conformers with in the binding site. During the exhaustive search, ligand poses were scored using the Chemical Gaussian Overlay (CGO) function that takes into account the shape and chemistry of the docked ligand pose relative to the co-crystalized protein ligand. The top ranked CGO poses are then optimized and rescored using the Fred Chemgauss4 score.

Discovery Studio: The LibDock[85] module of Discovery Studio was used for ligand docking. LibDock is based on the algorithm developed by Diller and Merz and this algorithm uses protein binding site features to guide docking. This software is part of Discovery Studio (Accelrys Software Inc). The receptor binding site was automatically searched and determined within LibDock during docking set up. The top 3 poses were kept and re-scored using two empirical scoring functions Jain and Ludi1.

2.7. Statistical Analysis:

Statistical analyses including Pearson and Spearman correlation calculations and outlier identification (ROUT method) were performed using GraphPad Prism version 5.00 for Mac OS X, GraphPad Software, San Diego California USA, www.graphpad.com.

Accepted Manuscript

ACCEPTED MANUSCRIPT

3. Results

An important property of a scoring function is how accurately it predicts the activity of a docked compound. In our comparison of different docking and scoring functions for sets of congeneric ligands against six selected protein targets (Table 2), we aimed to gauge the general performance of some of the more readily accessible scoring functions in predicting both absolute and relative ranking of biological activities for selected ligands against their reported protein targets, five enzymes and one protein receptor. It is notable that, for a virtual screening approach, this correlation does not have to be linear. A scoring function can work well as long as it can provide the correct ranking of candidate molecules[15]. Hence, two commonly used parameters to measure the goodness of correlation between scores from docking and tested biological activities are the Pearson correlation coefficient (R_p) and the Spearman correlation coefficient (R_s). The Pearson correlation is typically employed to provide a linear relationship, whereas the Spearman correlation provides a measurement of the non-parametric relationship between ranks of data. Therefore, the Pearson coefficient is generally a better measurement for absolute predictions while the Spearman coefficient is more appropriate for relative ranking[21]. The Pearson correlation coefficient is calculated as follows:

$$R_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

N is the number of tested complexes, x_i and y_i are the experimentally determined binding energy and the calculated score for the i-th complex, respectively; $\bar{x}$ is an arithmetic average over all the complexes.

The Spearman correlation coefficient measures the correlation between two sets of rankings to provide an index for ranking complexes and is calculated as follows:

$$R_s = 1 - \frac{6 \times \sum_{i=1}^{N}(R_i - S_i)^2}{N^3 - N}$$

where R_i is the rank of complex i determined by its experimental binding constant, while S_i is the rank reflected by a scoring function. N is the total number of tested complexes. For both the Pearson and Spearman coefficients, the values can vary from -1 to 1, while -1 suggests an inverse correlation between two set of ranking variables and 1 suggests a strong positive correlation between them.

Accepted Manuscript

ACCEPTED MANUSCRIPT

It was found that most of the docking packages examined here docked the congeneric ligands into the correct binding site of their targets, with the core structural features of each ligand tending to superimpose (Fitted docked ligands, Figure 1). The capacity of each docking program to successfully re-dock the bound crystal structure ligand into the native-binding conformation was tested using rmsd of heavy atoms against the bound crystal structure ligand. It was found that most of the docking programs were able to reproduce acceptable native ligand conformations with heavy atom rmsd ≤ 2 Å (Supporting Information Table S7), most successfully achieving re-docking poses of crystal ligands with rmsd < 1 Å (Table S7). Only a small number of exceptions were noted in particular, Autodock-Cdk2 kinase rmsd 2.2 Å, DS Libdock-Aurora kinase rmsd 2.5 Å, DS Libdock-Pla2g2a rmsd 3.3 Å and GoldScore-Pla2g2a rmsd 5.2 Å. Only GOLDScore failed to consistently reproduce ligand docking poses found in crystal structures for pla2g2a. However, it should be noted that even ligands that poorly reproduce the native ligand pose as defined by a crystal structure (and measured by rmsd threshold values) can still provide valuable information to a medicinal chemist. Alternative ligand poses in an active site may provide other plausible space-filling orientations or alternative contacts with active site residues that suggest further chemical modifications to the ligand [31].

Furthermore, crystal structures often only capture a single snapshot of the ligand bound protein complex, and whether such a static structure is always a real reflection of the ligand efficiency data obtained in solution is questionable. Instead of targeting a single docking pose of a given ligand on a single receptor, looking for the most populated alternatives from an ensemble of docking solutions within the active binding site may be more effective. It was beyond the scope of this study to fully examine the “docking power” of each program through parameter manipulation, but we provide here the docked poses of the two best performing and two worst performing scoring functions on a compliant target: cdk2 kinase (Figure 2) and a difficult target: COX-2 (Figure 3). When scoring functions gave a negative value, these were made positive to ensure a more positive score represented a higher pK_i or pIC_{50}. Correlation plots between docking scores (representing binding affinity) and pK_i or pIC_{50} (representing experimental inhibitor potencies) were calculated and Figure 4 displays the best correlating scoring function for each target protein. Correlation plots of all the scoring functions are included in Supporting Information. Pearson correlation coefficients and Spearman ranking correlation coefficients are

Accepted Manuscript

listed for each series in Table 3. In addition, a correlation heatmap of all scoring functions on each target is depicted in Figure 5.

Table 2: Six Protein Targets and Relative Hydrophobicities

Sitemap calculated relative hydrophobicity of active sites from 6 targets in this study. A balance of > 6.0 indicates high hydrophobicity and likely lipophilicity.

Protein Targets	Hydrophobic	Hydrophilic	Balance
Factor Xa	1.3	0.7	1.8
Cdk2 Kinase	1.4	1.0	1.4
Aurora A Kinase	1.8	1.1	1.6
COX-2	3.4	0.5	6.8
pla2g2a	1.6	0.9	1.8
Estrogen Receptor	4.4	0.3	13.3

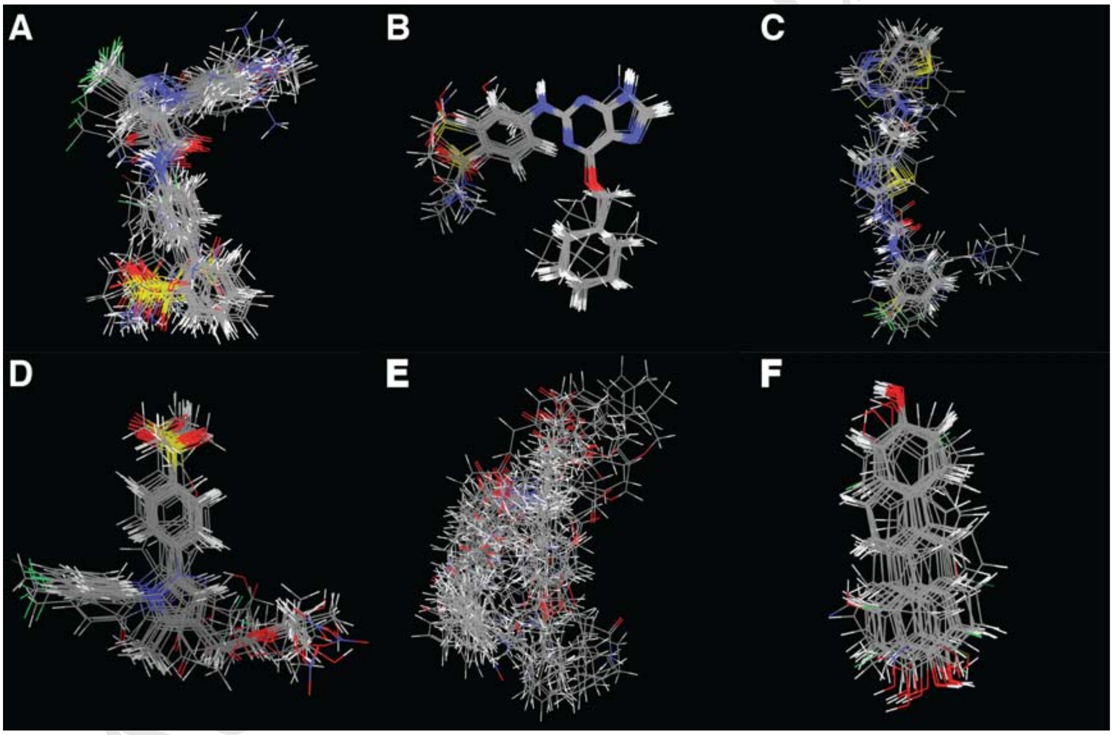
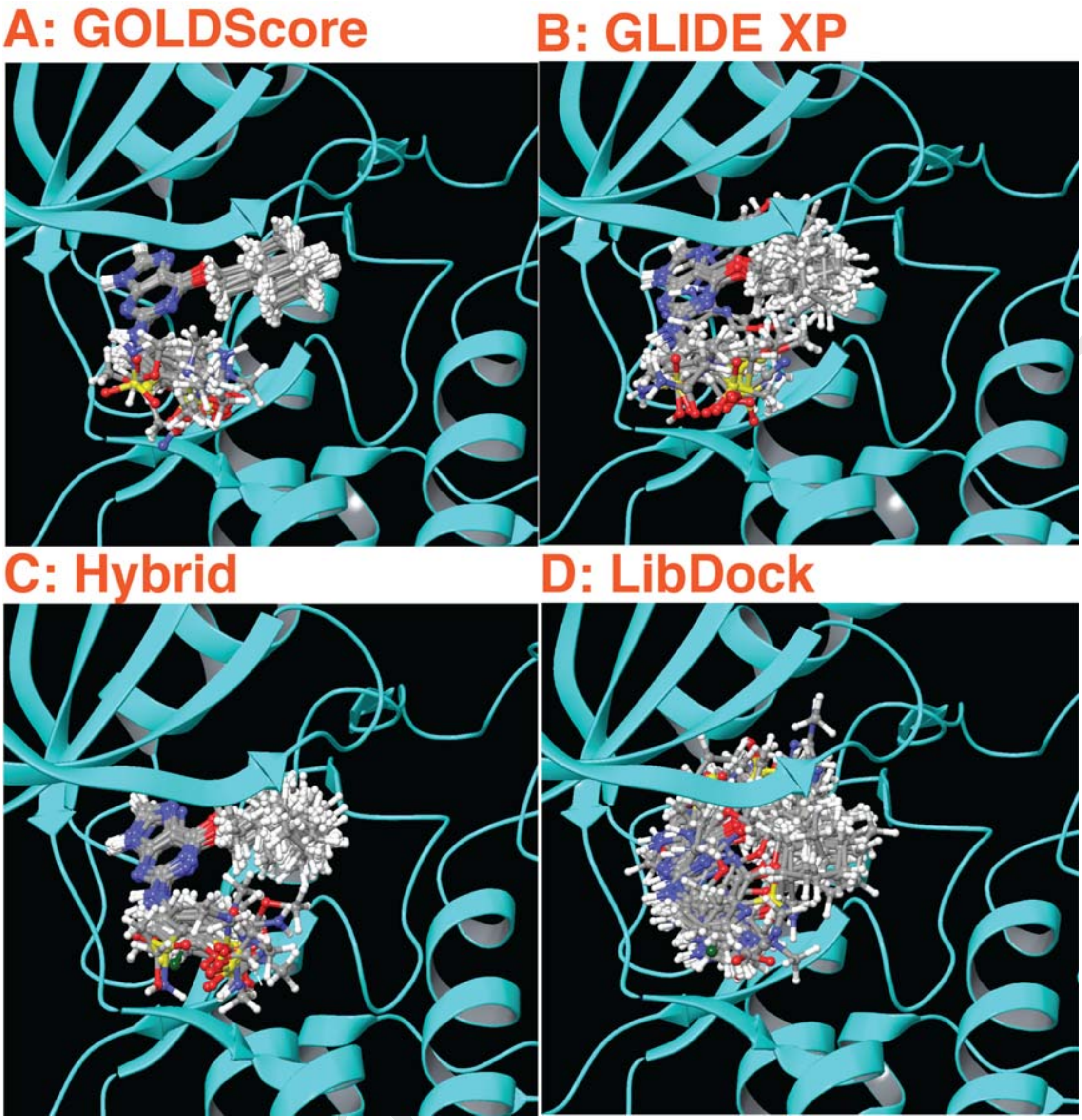


Figure 1: Superimposed view of docked ligands in protein active site derived by Fitted docking program. Ligands for A: Factor Xa ligands; B: cdk2 kinase ligands; C: Aurora A kinase ligands; D: COX-2 ligands; E: pla2g2a ligands; F: Estrogen receptor ligands.

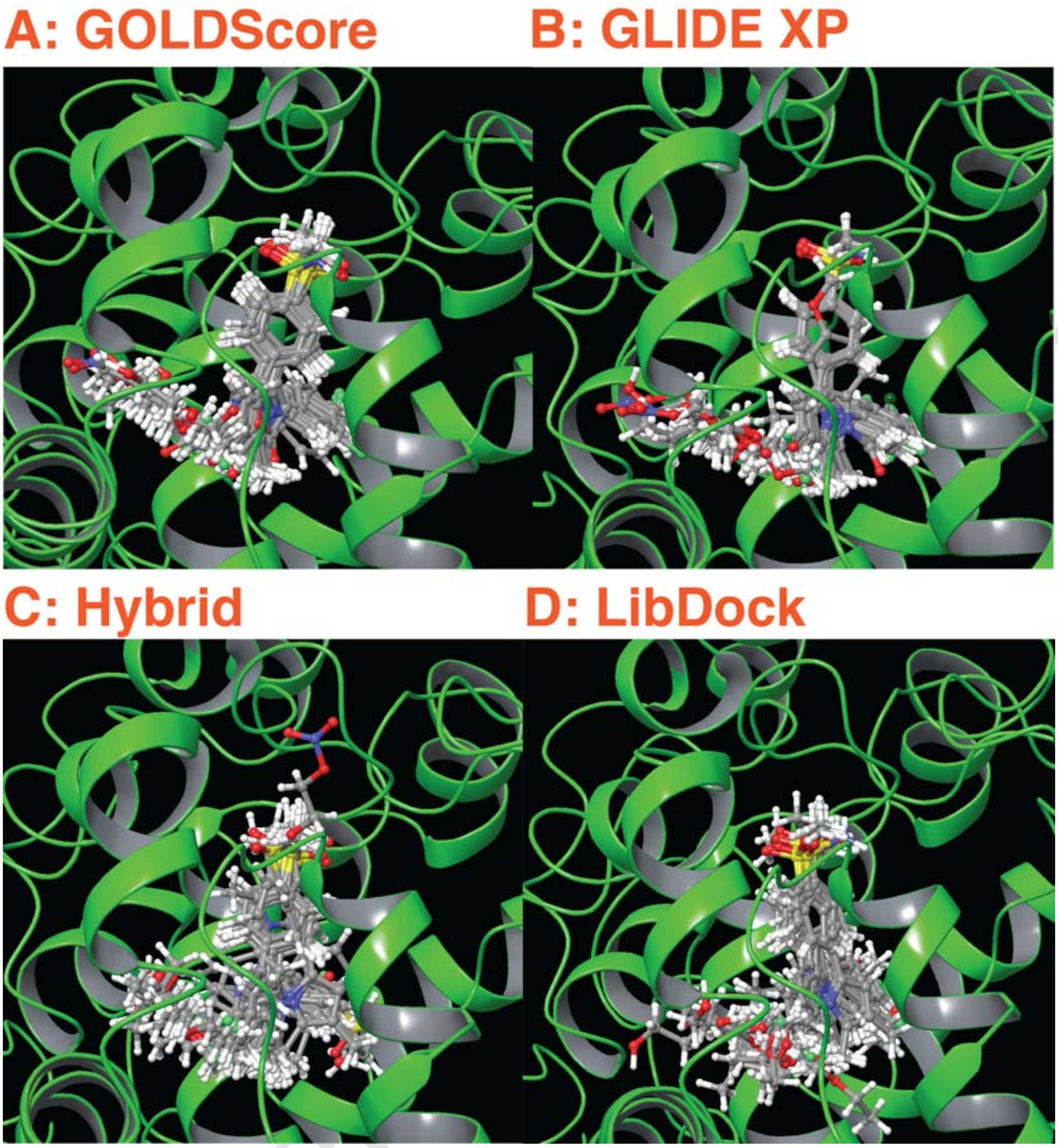
Accepted Manuscript



1
2 **Figure 2:** Docked poses of cdk2 kinase ligands by A: GOLDScore, B: GLIDE XP, C: Hybrid, D:
3 LibDock

Accepted Manuscript

ACCEPTED MANUSCRIPT



1
2 **Figure 3:** Docked poses of COX-2 ligands by A: GOLDScore, B: GLIDE XP, C: Hybrid, D: LibDock

Accepted

ACCEPTED MANUSCRIPT

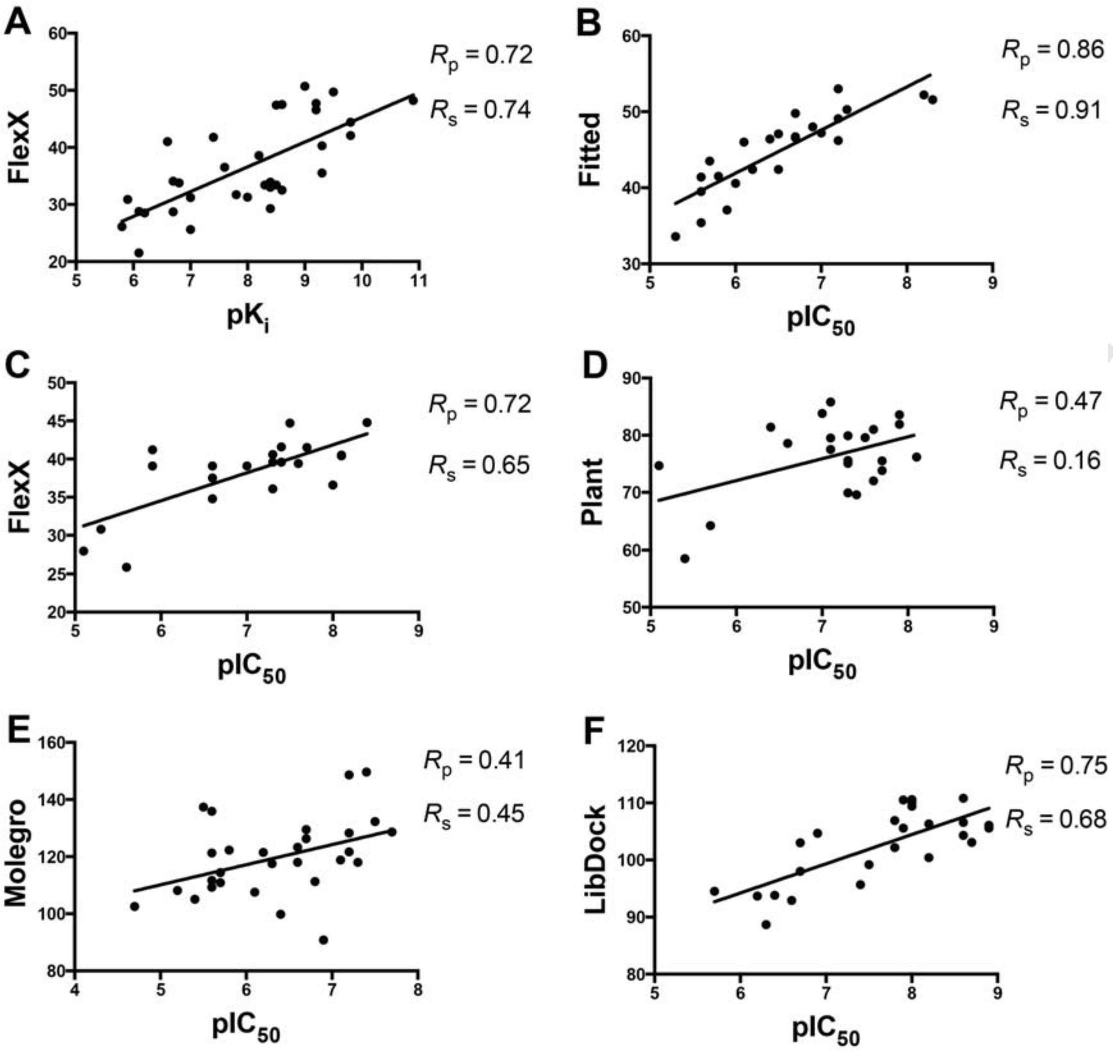


Figure 4: Plot of best performing scoring function values vs experimental protein inhibition by ligands for 6 protein targets. A: FlexX vs pK$_i$ for Factor Xa. B: Fitted vs pIC$_{50}$ for Cdk2 kinase. C: FlexX vs pIC$_{50}$ for Aurora A kinase. D: Plant vs pIC$_{50}$ for COX-2. E: Molegro vs pIC$_{50}$ for pla2g2a. F: LibDock vs pIC$_{50}$ for Estrogen Receptor. Pearson (R_p) and Spearman (R_s) coefficients.

Accepted

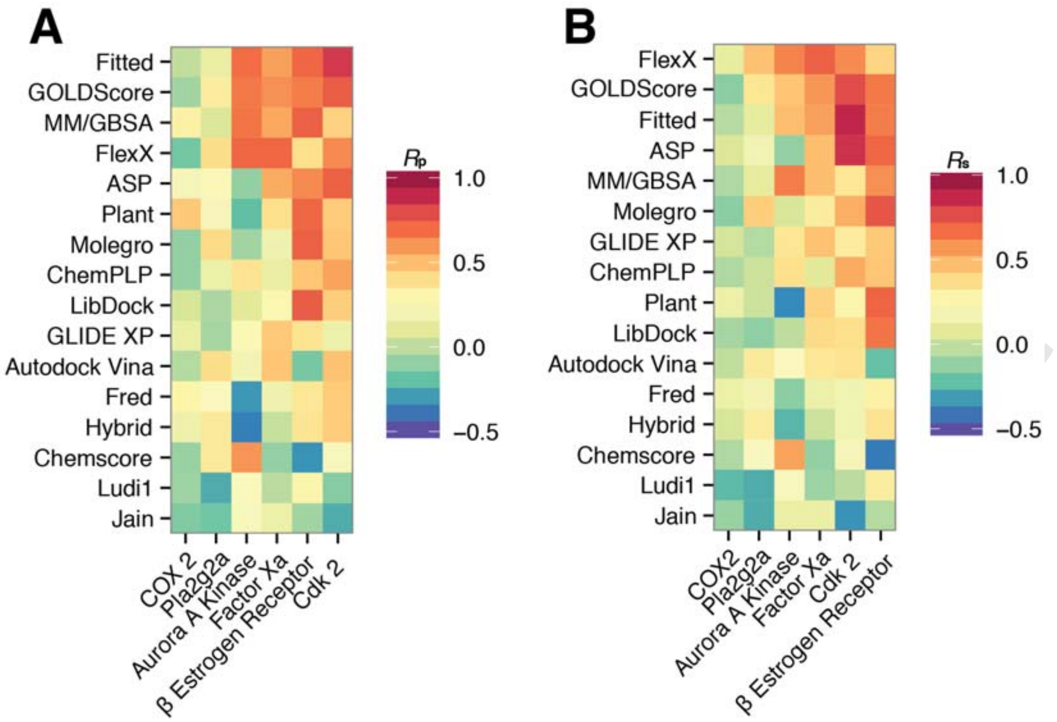


Figure 5: Heatmap correlations of selected scoring functions on protein targets. A: Pearson correlation coefficient. B: Spearman ranking coefficient. Y axis: Scoring functions (strongest to weakest) as ranked from top to bottom. X axis: Protein targets gaining summative correlations (lowest to highest) as ranked from left to right. Pearson correlation coefficient (R_p): linear correlation; Spearman correlation coefficient (R_S): non-parametric relative correlation. Both range from -1 to 1, indicating negatively correlated and positively correlated.

For Factor Xa, FlexX (R_p = 0.72, R_s = 0.74) performed best with relative values for other programs being: GOLDScore (R_p = 0.62, R_s = 0.60), Fitted (R_p = 0.58, R_s = 0.58), ASP (R_p = 0.55, R_s = 0.50), MM-GBSA (R_p = 0.56, R_s = 0.50) and GLIDE XP (R_p = 0.47, R_s = 0.49) generated moderate correlations in both Pearson coefficient and Spearman ranking coefficient. Autodock Vina (R_p = 0.48, R_s = 0.36), Molegro (R_p = 0.18, R_s = 0.34), Plant (R_p = 0.39, R_s = 0.44), Fred (R_p = 0.18, R_s = 0.16), Hybrid (R_p = 0.02, R_s = 0.04), LibDock (R_p = 0.29, R_s = 0.41) and Jain (R_p = 0.16, R_s = 0.15) gave low correlations. In comparison, two empirical based scoring functions in GOLD software, ChemScore (-ve correlations) and ChemPLP (correlations < 0.15) failed to produce comparable correlations as compared to GOLDScore and ASP score. Ludi1 (R_p = -0.01, R_s = -0.18) also produced negative correlation for this target.

For Cdk2 kinase, positive correlations between the predicted scores from docking and experimentally measured activities were obtained by most of the scoring functions applied. Fitted (R_p = 0.86, R_s = 0.91) gave the best correlations for Cdk2

Accepted Manuscript

kinase. GOLDScore and ASP outperformed the rest by achieving a Pearson correlation of 0.75 and 0.74 and a high Spearman correlation of 0.80 and 0.88 respectively. Both FlexX ($R_\mathrm{p} = R_\mathrm{s} = 0.63$) and ChemPLP ($R_\mathrm{p} = 0.57$, $R_\mathrm{s} = 0.55$) gave reasonable correlations. Autodock Vina ($R_\mathrm{p} = 0.49$, $R_\mathrm{s} = 0.38$), Molegro ($R_\mathrm{p} = 0.48$, $R_\mathrm{s} = 0.54$), Plant ($R_\mathrm{p} = 0.46$, $R_\mathrm{s} = 0.30$), Fred ($R_\mathrm{p} = 0.46$, $R_\mathrm{s} = 0.19$), Hybrid ($R_\mathrm{p} = 0.46$, $R_\mathrm{s} = 0.19$) and LibDock ($R_\mathrm{p} = 0.45$, $R_\mathrm{s} = 0.39$) achieved lower correlations. GLIDE XP ($R_\mathrm{p} = 0.16$, $R_\mathrm{s} = 0.34$) gave very poor correlation but rescoring with MM-GBSA ($R_\mathrm{p} = 0.44$, $R_\mathrm{s} = 0.36$) significantly improved the observed correlation. GLIDE XP incorrectly scored compounds **52** and **53**, giving these two ligands as outliers. However, MM-GBSA rescoring eliminated the outliers, possibly accounting for the improved performance of MM-GBSA over GLIDE XP. Chemscore produced a weak correlation ($R_\mathrm{p} = 0.23$, $R_\mathrm{s} = 0.22$) for cdk2 kinase. The only two scoring functions generating negative correlations on this target were Jain ($R_\mathrm{p} = -0.25$, $R_\mathrm{s} = -0.32$) and Ludi1 ($R_\mathrm{p} = -0.13$, $R_\mathrm{s} = -0.11$).

For Aurora A kinase, FlexX produced the best linear correlation and second best ranking correlation ($R_\mathrm{p} = 0.72$, $R_\mathrm{s} = 0.65$). Fitted Score performed reasonably well on this target by achieving a Pearson correlation of 0.70. Prime: MM-GBSA ($R_\mathrm{p} = 0.68$, $R_\mathrm{s} = 0.66$), GOLDScore ($R_\mathrm{p} = 0.67$, $R_\mathrm{s} = 0.48$) and GOLD: ChemScore ($R_\mathrm{p} = 0.61$, $R_\mathrm{s} = 0.57$) also generated good correlations on this target by achieving $R_\mathrm{p} > 0.6$. The highest Spearman correlation was achieved by MM-GBSA. GLIDE XP ($R_\mathrm{p} = 0.28$, $R_\mathrm{s} = 0.37$), Autodock Vina ($R_\mathrm{p} = 0.20$, $R_\mathrm{s} = 0.26$) and the 3 scoring functions from DS: LibDock ($R_\mathrm{p} = 0.1$, $R_\mathrm{s} = 0.0$), Jain ($R_\mathrm{p} = 0.23$, $R_\mathrm{s} = 0.15$), and Ludi1 ($R_\mathrm{p} = 0.26$, $R_\mathrm{s} = 0.23$) all produced weak correlations on this target. ASP ($R_\mathrm{p} = -0.1$, $R_\mathrm{s} = -0.1$), Molegro ($R_\mathrm{p} = -0.07$, $R_\mathrm{s} = 0.07$), Plant ($R_\mathrm{p} = -0.21$, $R_\mathrm{s} = -0.34$), Fred ($R_\mathrm{p} = -0.31$, $R_\mathrm{s} = -0.12$) and Hybrid ($R_\mathrm{p} = -0.37$, $R_\mathrm{s} = -0.23$) generated negative correlations. Compound **74** was a notable outlier in Fred, Hybrid.

COX-2 appeared to be the most difficult target for scoring functions to predict both absolute activities and relative ranking between activity and scores in this study. Shown in Table 2, Pearson correlation and Spearman ranking coefficients each received six negative results from all scoring functions applied. Almost half of the scoring functions negatively correlated with compounds biological activities. For the scoring functions which gave positive correlations, none of them achieved a Pearson correlation higher than 0.5 ($R_\mathrm{p} > 0.5$), with the highest of 0.47 achieved by Plant from Molegro. Unfortunately, the highest Spearman ranking coefficient obtained from

Accepted Manuscript

Plant and Fred scores was only 0.16, indicating poor ranking ability of scoring functions for COX-2 ligands. Furthermore, Discovery Studio was only able to successfully dock 15 of 22 ligands due mainly to steric clashes between the ligands and active site receptor residues. Compared to other targets evaluated here, COX-2 was characterized by 92% hydrophobic residues in its active site[24], reflecting a bottleneck faced by all scoring functions to deal with protein-ligand interactions mainly involving mainly hydrophobic contacts.

For pla2g2a, none of the scoring functions produced a correlation or ranking coefficient >0.5 for the docking of flexible, lipid-like, hydrophobic inhibitors that were also substrate analogues. Molegro produced the highest Pearson correlation (R_p = 0.41, R_s = 0.45). Autodock Vina (R_p = 0.40, R_s = 0.35) and FlexX (R_p = 0.40, R_s = 0.49) generated equivalent second highest Pearson correlations for this target. Fitted, Fred, Hybrid and all scoring functions from GOLD produced slightly positive correlations. GLIDE XP score (R_p = -0.06, R_s = -0.03), together with the 3 scoring functions from Discovery Studio, negatively correlated with biological activities of the ligands. Although MM-GBSA rescoring increased the R_p and R_s, the overall low correlation indicated the scoring functions in GLIDE did not perform well for this target.

For β estrogen receptor, most of the scoring functions were able to give good correlations with the exception of Chemscore, Autodock Vina, and Jain score. Seven scoring functions, LibDock (R_p = 0.75, R_s = 0.68), Molegro (R_p = 0.74, R_s = 0.77), Plant (R_p = 0.72, R_s = 0.73), MM-GBSA (R_p = 0.74, R_s = 0.62), Fitted (R_p = 0.72, R_s = 0.66), GOLDScore (R_p = 0.66, R_s = 0.67) and ASP (R_p = 0.63, R_s = 0.72) performed well compared to the rest by achieving both Pearson and Spearman correlation over 0.6. GLIDE XP (R_p = 0.38, R_s = 0.47), FlexX (R_p = 0.39, R_s = 0.43), Fred (R_p = 0.36, R_s = 0.32), Hybrid (R_p = 0.38, R_s = 0.38) and Ludi1 (R_p = 0.30, R_s = 0.34) generated weak correlations for this target. Both Pearson and Spearman coefficients from Chemscore (R_p = -0.35, R_s = -0.4) and Autodock Vina (R_p = -0.16, R_s = -0.20) were negative, reflecting an inverse correlation with the binding afinities of the ligands. Compound **146** was an outlier from GLIDE XP scoring, but rescoring from MM-GBSA improved correlations.

Accepted Manuscript

Accepted Manuscript

Table 3. Correlations between docking scores and experimentally determined binding affinity/biological activity given by 16 scoring functions.

	Factor Xa		**CDK2**		**Aurora kinase**		**COX-2**		**Pla2g2a**		**Estrogen**		**Sum**[a]	
scoring functions	$\boldsymbol{R_p}$	$\boldsymbol{R_s}$	$\boldsymbol{R_p}$	$\boldsymbol{R_s}$	$\boldsymbol{R_p}$	$\boldsymbol{R_s}$	$\boldsymbol{R_p}$	$\boldsymbol{R_s}$	$\boldsymbol{R_p}$	$\boldsymbol{R_s}$	$\boldsymbol{R_p}$	$\boldsymbol{R_s}$	$\boldsymbol{R_p}$	$\boldsymbol{R_s}$
GOLD: GOLDScore	0.62	0.60	0.75	0.80	0.67	0.48	-0.07	-0.12	0.34	0.37	0.66	0.67	2.97	2.80
GOLD: Chemscore	-0.10	-0.10	0.23	0.22	0.61	0.57	-0.09	-0.04	0.35	0.24	-0.32	-0.38	0.68	0.51
GOLD: ChemPLP	0.14	0.10	0.57	0.55	0.38	0.39	-0.10	-0.04	0.16	0.04	0.48	0.48	1.63	1.52
GOLD: ASP	0.55	0.50	0.74	0.88	-0.10	-0.10	0.22	0.08	0.27	0.19	0.63	0.72	2.31	2.27
GLIDE: XP	0.47	0.49	0.16	0.34	0.28	0.37	0.15	0.06	-0.06	-0.03	0.38	0.47	1.38	1.70
Prime-mmGBSA	0.56	0.50	0.44	0.36	0.68	0.66	0.32	-0.04	0.08	0.12	0.74	0.62	2.82	2.22
FlexX	0.72	0.74	0.63	0.63	0.72	0.65	-0.17	0.13	0.40	0.49	0.39	0.43	2.69	**3.07**
Autodock Vina	0.48	0.36	0.49	0.38	0.20	0.26	-0.03	0.01	0.40	0.35	-0.16	-0.20	1.38	1.16
Fitted	0.58	0.58	0.86	0.91	0.70	0.50	0.01	-0.02	0.14	0.12	0.72	0.66	**3.01**	2.75
Molegro	0.18	0.34	0.48	0.54	-0.07	0.07	-0.10	-0.12	0.41	0.45	0.74	0.77	1.64	2.05
Plant	0.39	0.44	0.46	0.30	-0.21	-0.34	0.47	0.16	0.22	0.04	0.72	0.73	2.05	1.33
Fred: Chemgauss4	0.18	0.16	0.46	0.19	-0.31	-0.12	0.31	0.16	0.26	0.20	0.36	0.32	1.26	0.91
Hybrid: Chemgauss4	0.02	0.04	0.46	0.19	-0.37	-0.23	0.17	0.07	0.35	0.34	0.38	0.38	1.01	0.79
DS: LibDock	0.29	0.41	0.45	0.39	0.1	0	0.07	-0.06	-0.05	-0.11	0.75	0.68	1.61	1.31
DS: Jain	0.16	0.15	-0.25	-0.32	0.23	0.15	-0.14	-0.09	-0.18	-0.25	-0.07	-0.03	-0.25	-0.39
DS: Ludi1	-0.01	-0.08	-0.13	-0.01	0.26	0.23	-0.08	-0.22	-0.26	-0.25	0.3	0.34	0.08	-0.09
Sum[b]	4.79	4.75	6.73	6.29	3.18	3.16	1.09	0.29	3.32	2.92	5.72	5.67		

[a] Sum of Pearson and Spearman correlations of individual scoring function on all targets.

[b] Sum of Pearson and Spearman correlations for each target from all scoring functions.

ACCEPTED MANUSCRIPT

4. Discussion

In this study, eight different docking programs and sixteen scoring functions accessible to most researchers were compared and assessed through an examination of six proteins and individual ligand sets for which experimental biological activities have been reported by individual research groups used a well-defined set of conditions. Most of the ligands examined were not reported in crystal structures with their target protein. Where they were, the top ranked ligand binding poses derived from each docking method were compared to the ligand orientation in the crystal structure. Most ligands in each sample set docked in a very similar orientation to that found in the crystal structure, except in the large hydrophobic cleft of pla2g2a (Figure 1). However, even unexpected ligand binding modes can be used to explore alternative ligand protein contacts and lead to design of novel new ligands for medicinal chemistry[31]. Furthermore, docking poses and predictions of ligand binding affinities might be improved by introducing protein flexibility via protein ensemble docking[86].

Factor Xa is a serine protease considered to have a hydrophilic binding site and high affinity binding is often achieved by ligands that make hydrogen bonds with the enzyme. The best performing scoring functions were FlexX and GOLDScore (Table 3). FlexX was previously shown to perform well for other hydrophilic protein binding sites (e.g. p38 MAP kinase, thrombin, neuraminidase, gelatinase A) that typically make multiple hydrogen bonds to the ligand[16]. It was encouraging that guanidine-containing compounds (compounds **27**-**33** from SI: Table1) ranked at the top of ligands scored by FlexX. The most potent compound **7** (K_i = 0.013 nM) assessed in an enzyme assay ranked as the 3rd top compound in the FlexX scoring list, indicating a satisfying enrichment effect in the series of compounds chosen. It has been noted that some outliers can significantly impair the performance of some scoring functions, for example as in GOLDScore which ranked compound **7** only 10th, giving GOLDSocre a poorer differentiation for the most active compounds. Chemscore (R_p = -0.10, R_s = -0.10) and ChemPLP (R_p = 0.14, R_s = 0.10) produced the lowest correlation for Factor Xa ligand activity. Chemscore did not differentiate between different types of hydrogen bonds[87], and this may explain why it performed so poorly for Factor Xa.

Accepted Manuscript

ACCEPTED MANUSCRIPT

Docking of congeneric inhibitors of Cdk2 gave good activity correlations with the scoring functions Fitted, GOLDScore, ASP and FlexX. MM-GBSA has been reported to perform well against Cdk2 with a correlation of 0.71 (R_p = 0.71) using 11 ligands[46] by Lyne et al.[40], however, for the 24 ligands and protocol used by us there was a lower correlation (R_p = 0.44) using the same scoring functions. Fitted score (R_p = 0.86), GOLDScore (R_p = 0.75) and ASP (R_p = 0.74) score achieved better correlations compared to Prime: MM-GBSA in Lyne's study. Rapp et al. reported a "Prime-ligand" molecular mechanics approach to correlate the calculated binding energies with the biological activities of the same series of Cdk2 ligands from Lyne's study[41]. They achieved a Spearman correlation (R_s) of 0.75. The high Spearman correlations achieved herein in our study containing more than double of compounds (including the same 11 ligands in both Lyne' and Rapp's study) by Fitted (R_s = 0.91), GOLDScore (R_s = 0.80) and ASP (R_s = 0.88) indicate these scoring functions predict relative potencies of inhibitors for this target more accurately compared to the scoring functions from GLIDE. Meanwhile, FlexX produced 0.63 for both R_p and R_s, suggesting that it is effective for this target protein as well. The mildly hydrophilic nature of the active site of cdk2 may account for the poorer relative predictive value of Chemscore, Glide and Autodock Vina in matching experimental data ranking

Twenty potent and selective Aurora kinase inhibitors derived by converting a 3-trifluoromethyphenyl ring to an aminothiazole central ring[47] were also examined here. The scoring functions FlexX (R_p = 0.72) Fitted, GOLDScore, MM-GBSA, and Chemscore each showed a good correlations (>0.6) with enzyme inhibition data. Two previous studies using MM-GBSA by Lyne and molecular mechanics method by Rapp used compound congeners with differing core structures. Lyne et al. docked only 8 compounds from the series they selected and generated a Pearson correlation of 0.75[40] while Rapp et al. docked 12 compounds from the same series and achieved a stronger correlation of 0.8 and a Spearman ranking correlation (R_s) of 0.83. Rapp et al. also chose a series of compounds similar to those included here and achieved R^2 of 0.49 (Rp of 0.7) and R_s of 0.59. By comparison, our study involved the docking of 21 ligands, for which we found that MM-GBSA achieved a similar R_p (0.68) but a slightly higher R_s (0.66). Notably, FlexX score produced R_p 0.72 and R_s 0.65, which are both better compared to "Prime-ligand" scoring in Rapp's study over a smaller compound series. It was noted that in the crystal structure of Aurora kinase bound to its ligand, hydrogen bonding appears to play an important role to stabilize

Accepted Manuscript

ACCEPTED MANUSCRIPT

high affinity ligand binding to the receptor. This further supports the rationale that FlexX performs well for target proteins in which the active site has a degree of hydrophilic character.

In contrast with hydrophilic targets such as Factor Xa, where the active binding pocket is quite solvent exposed, the active site of COX-2 has a deeply buried hydrophobic ligand-binding site that makes predominantly hydrophobic van der Waals contacts with its ligand through residues such as F518, W387, Y385, L384, V523, F381, L352, V349, Y355, L359, L531, and V116. None of the scoring functions examined here for COX-2 ligands gave a good correlation between docking score and experimental inhibitor potency. In previous COX-2 inhibitor docking enrichment studies, FlexX scoring was found to be ineffective as compared to knowledge-based scoring functions such as Drugscore[16], while ICM has been reported to be better for COX-2 ligand enrichment than GOLD, GLIDE and FlexX in Chen's study[31], but was not examined here. Hydrogen bonds do not play a major role in the strong binding of ligands to COX-2, and scoring functions (e.g. FlexX, GOLDScore, Fitted) that performed well on other protein targets did not perform nearly as well with COX-2. An explanation for this may be that for compounds to penetrate deep into a hydrophobic ligand-binding pocket, they need to overcome a large entropy penalty to desolvate. Such desolvation terms are either not explicitly included in the scoring functions or are not currently accurate enough to correctly contribute to the score. Furthermore, the poor performance of all scoring functions examined here may highlight the lack of optimal terms in equations used to calculate predicted protein-ligand interactions that have strong hydrophobic contributions. Finally, the difference in pIC_{50} lies mostly within 1 to 1.5 units, which is within the error range of scoring functions. This could be another cause of COX-2 being less compliant with scoring functions.

For pla2g2a, SiteMap calculations predicted that this target is hydrophilic (balance of 1.80), but its active site is extremely hydrophobic and accommodates highly flexible phospholipid substrates. The SiteMap calculations may take into account the degree of exposure of the active site to the solvent of this enzyme and hence tends to assign too much hydrophilicity. The pla2g2a inhibitors were all synthesized and tested for activities within our group and so we are confident in comparisons of experimental inhibitory data between compounds in the series. This enzyme tends to catalyze aggregated substrates such as micelles, vesicles, membranes

Accepted Manuscript

ACCEPTED MANUSCRIPT

and monolayers [88]. Twenty-nine small organic inhibitors, that were structural analogues of the native glycerolphospholipid substrates and contained long chain aryl groups, were docked into pla2g2a. The two best performing scoring functions, Molegro (R_p = 0.41, R_s = 0.45) and FlexX (R_p = 0.40, R_s = 0.49), did not generate impressive Pearson or Spearman correlation coefficients for this target. Autodock Vina produced the same Pearson correlation (R_p = 0.40) as FlexX, but with a lower ranking correlation coefficient (R_s = 0.35). Several factors might conceivably affect the performance of the scoring functions for this target. First, the presence of a central catalytic Ca^{2+} ion, which coordinates to a carboxylate and an amide oxygen from each inhibitor as well as Asp 49 and Gly 30 enzyme residues in the active site, could present a challenge to scoring functions. Evaluating interactions with a metal ion involves estimating force field parameters that are still somewhat uncertain for metal-ligand protein complexes. Second, the relatively high number of rotatable C-C bonds enhances ligand flexibility and hence poses uncertainties for scoring functions in conformational sampling of different ligands. Third, there are few interactions made between the inhibitor and the very greasy active site of the enzyme, so any error in ligand orientation or enzyme residue location can profoundly affect affinity predictions for inserted ligands.

Based on SiteMap calculations of relative hydrophobicity of protein targets selected here, the binding site of the estrogen receptor was shown to be the most hydrophobic. Estrogen receptor inhibitors tend to be planar, low molecular weight phenyl-naphthalene derivatives. LibDock (R_p = 0.75) performed best in the correlation of docking scores with activities for the examined ligands followed by Molegro and MM-GBSA (R_p = 0.74). Glide has been shown to be effective for enrichment studies with the Estrogen receptor[31]. However, we found that GLIDE XP score generated a low correlation (0.38) with ligand activity, although this improved upon rescoring with MM-GBSA (R_s = 0.74). In discordance with the poor performance from GOLD in enriching ER ligands concluded by Chen et al.[31], GOLDScore (R_p = 0.66, R_s = 0.67) and ASP (R_p = 0.63, R_s = 0.72) produce good correlations in our hands. It is a bit surprising that, being the most hydrophobic target, scoring functions were able to give reasonable correlations with activities for the ligands examined. The ligands used were relatively more rigid and smaller molecules compared to those for the other five targets, consistent with the performance of

Accepted Manuscript

scoring functions not only being affected by the nature of the protein binding site but also by the nature of the ligands being docked.

The docking programs examined here have thus produced better correlations between pose scores and biological activity for the more hydrophilic vs hydrophobic protein targets. The Estrogen receptor was the exception with the ligands being smaller and more rigid, whereas for COX-2 and pla2g2a targets, their ligands were generally larger with more rotatable bonds contributing to higher ligand flexibility.

Predicting ligand binding affinity for protein targets with current pose scoring functions is limited[19, 33, 89]. The most recent CSAR 2012 exercise asked 20 computational labs to submit binding affinity predictions for four protein targets. Overall success was measured using the sum of both Pearson correlation and Spearman ranking correlation (R_p and R_s) as measuring criteria, a total of $R_p = 4.0$ or $R_s = 4.0$ indicated a perfect prediction and a total of R_p or $R_s > 2.0$ was considered as good performance. Only one group produced a sum $R_p > 2.0$ and 2 groups were able to achieve a sum of $R_s > 2.0$[21]. In a similar fashion, we consider a total of 6.0 for both Pearson correlations and Spearman ranking correlations as perfect predictions since 6 targets were examined here. Hence, only values >3.0 were considered as acceptable performance from the scoring functions. Fitted gave the best Pearson correlations total R_p value of 3.07, followed by GOLDScore (total $R_p = 2.97$), MM-GBSA (total $R_p = 2.82$) and FlexX (total $R_p = 2.69$). The highest Spearman correlation coefficient was achieved by FlexX (total $R_s = 3.01$), followed by GOLDScore (total $R_s = 2.80$) and Fitted (total $R_s = 2.75$). Overall, Fitted, FlexX and GOLDScore were the three best overall scoring functions in predicting the relative potencies for congeneric compounds whereas Jain score was the worst and generated anti-correlations across all six targets.

The correlation between docking scores and activities was also summarized (Table 2) for each protein target to assess the suitability of each target for ligand binding affinity prediction using a docking methodology. None of the protein targets gave a sum of correlations ≥8.0. Cdk 2 kinase obtained the highest sum of R_p (6.73) and R_s (6.29) values from all scoring functions. It also received the highest Pearson correlation from almost half of the scoring functions applied, indicating that this target is perhaps better suited for the prediction of ligand binding affinity by current scoring functions. β-estrogen receptor and factor Xa received the two highest R_p values from all scoring functions. Such results may suggest the applicability of the top

Accepted Manuscript

ACCEPTED MANUSCRIPT

performing scoring functions on other protein targets belonging to the superfamilies of selected targets in this study.

GOLDScore was observed to generally perform better for hydrophilic targets. It achieved Pearson correlations > 0.6 for Factor Xa, cdk2 kinase and aurora A kinase. Our findings are in agreement with Kontoyianni's evaluation of five docking programs using 69 diverse protein-ligand complexes[24]. On hydrophobic targets, GOLDScore did not produce as positive results as for hydrophilic targets. One possible reason for this may be the lack of an explicit term in its scoring functions for hydrophobic interaction, which is an important element for hydrophobic protein binding sites and complementary ligands[32]. The ASP scoring function performed well on all the targets except Aurora A kinase, the poor performance in this target impaired the overall performance of ASP scoring. However, it was still the second best scoring function after the GOLD package. ChemPLP was only able to produce minor correlations for some of the targets in this study. In GOLD software, Chemscore was found to be the weakest scoring function in predicting ligand binding affinity/biological activity.

GLIDE XP score was not as discriminatory as GOLDScore of the nature of the active site of the protein. This echoes Kontoyianni's findings[24] in their comparative study in docking performance. Overall, XP score did not produce significant correlations for the targets here. However, one notable finding in this study is the performance of MM-GBSA for improving the predictive accuracy of compound binding or activity. In MM-GBSA, energies were estimated based on OPLS-AA force field for molecular mechanics energy (EMM) and the surface-generalised borne model for polar solvation energy, and a non-polar solvation term was also taken into account[74]. Although we observed a general trend that rescoring by MM-GBSA increased the correlation between predicted scores and biological activities, we were not able to obtain as dramatic an improvement as reported by Lyne [40]. Considering the larger number of ligands in the dataset used in our study, outliers may have impaired the performance of MM-GBSA scoring. Hence, further studies are needed to verify its usefulness against other ligands.

FlexX was the only scoring function to perform better towards the three hydrophilic targets. This scoring function also produced the second highest Pearson correlation for inhibitors of pla2g2a. FlexX has previously been found to perform well on hydrophilic targets, such as neuraminidase[16]. FlexX may be the docking package

Accepted Manuscript

of choice if lead optimization is being performed on hydrophilic protein targets like serine protease or kinases that share similar binding sites to Factor Xa and Aurora A kinase respectively.

Three scoring functions were evaluated from Discovery Studio software in this work. However, none performed impressively except for LibDock score on β estrogen receptor. Jain and Ludi1 produced low or negative correlations on the majority of the targets.

This study has compared both free and low cost commercial docking software available for ligand docking and scoring. Autodock Vina (free), Fitted, Fred and Molegro (available for academic license) were also included in our studies. Encouragingly, Fitted software outperformed all others in generating a sum of Pearson correlation of 3.01. It also achieved the best result for cdk2 kinase (R_p 0.86, R_s 0.91). Intriguingly, Plant score from Molegro software performed best for COX-2, whereas Molegro re-rank score performed best for sPLA2. This suggests that it may be of potential use in scoring hydrophobic ligands for hydrophobic protein active sites. Scoring functions from Autodock Vina and Fred did not generate any correlation > 0.5 on any target, indicating that the scoring functions from these packages are not well suited for rank-ordering of compound potencies, at least for the protein-ligand sets chosen here. The use of these packages for lead ligand optimization based on predicted compound activities seems to require further scoring function optimization.

As a final cautionary note, the currently available scoring functions do not usually include terms that take into account aromatic-aromatic or π-cation or halogen-protien interactions[90-92]. Many drugs contain halogen atoms introduced during lead optimization for pharmacokinetic or metabolic reasons[93-96]. None of the scoring functions used here are able to accurately deal with halogens. Liu et al. recently developed the first halogen bonding scoring function and showed moderate success in docking, ranking and scoring power[94]. Future scoring function development and optimization should incorporate consideration of these interactions.

5. Conclusion

Eight docking programs and sixteen scoring functions most accessible to medicinal chemists were compared for their accuracy in predicting experimental

Accepted Manuscript

inhibitory activities against six unrelated protein targets. Given the simplicity of sampling and scoring at lower computational cost compared to calculating free energies, the results were reasonably impressive for some of the scoring functions. However, the ability of scoring functions to correctly rank compounds remains challenging on the basis of results herein. Both commercial and free academic docking programs were able to produce good correlations on some targets like factor Xa, Cdk2 kinase, and Aurora kinase. We note that the nature of the active site of the proteins, the choice of scoring functions and the set of ligands used for comparisons, all affected the performance in scoring and ranking compounds. For targets with very hydrophobic active site cavities, such as COX-2 and Pla2g2a, none of the scoring functions examined were able to accurately predict or rank compounds according to experimentally reported inhibitor potencies. This may be a result of the types of ligands studied here. For medicinal chemists who use these approaches to optimize their leads for potency, docking programs like Fitted, FlexX, and GOLD are likely to be most effective for protein targets such as kinases and serine proteases. In general, the docking and scoring functions need to be matched to the protein target and ligand series for optimum results. No program used was effective for all six protein-ligand data sets sampled in this study.

6. Acknowledgements

Funding was provided by the National Health and Medical Research Council of Australia through a grant (APP1025883) and a Senior Principal Research Fellowship (1027369) to DF; a Queensland Government CIF grant; and the Australian Research Council for a grant (DP130100629) and a Centre of Excellence in Advanced Molecular Imaging (CE140100011).

7. References:

[1] H. Gohlke, G. Klebe, Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors, Angew. Chem. Int. Ed. Engl. 41 (2002) 2644-2676.

[2] M.K. Gilson, H.X. Zhou, Calculation of protein-ligand binding affinities, Annu. Rev. Biophys. Biomol. Struct. 36 (2007) 21-42.

[3] C.R. Guimaraes, D.L. Boger, W.L. Jorgensen, Elucidation of fatty acid amide hydrolase inhibition by potent alpha-ketoheterocycle derivatives from Monte Carlo simulations, J. Am. Chem. Soc. 127 (2005) 17377-17384.

Accepted Manuscript

[4] T. Simonson, G. Archontis, M. Karplus, Free energy simulations come of age: protein-ligand recognition, Acc. Chem. Res. 35 (2002) 430-437.
[5] C.R. Guimaraes, M. Cardozo, MM-GB/SA rescoring of docking poses in structure-based lead optimization, J. Chem. Inf. Model. 48 (2008) 958-970.
[6] P.A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, T.E. Cheatham, 3rd, Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models, Acc. Chem. Res. 33 (2000) 889-897.
[7] A.M. Ferrari, G. Degliesposti, M. Sgobba, G. Rastelli, Validation of an automated procedure for the prediction of relative free energies of binding on a set of aldose reductase inhibitors, Bioorg. Med. Chem. 15 (2007) 7865-7877.
[8] G. Barreiro, C.R. Guimaraes, I. Tubert-Brohman, T.M. Lyons, J. Tirado-Rives, W.L. Jorgensen, Search for non-nucleoside inhibitors of HIV-1 reverse transcriptase using chemical similarity, molecular docking, and MM-GB/SA scoring, J. Chem. Inf. Model. 47 (2007) 2416-2428.
[9] J. Fidelak, J. Juraszek, D. Branduardi, M. Bianciotto, F.L. Gervasio, Free-energy-based methods for binding profile determination in a congeneric series of CDK2 inhibitors, J. Phys. Chem. B. 114 (2010) 9516-9524.
[10] A.R. Leach, B.K. Shoichet, C.E. Peishoff, Prediction of protein-ligand interactions. Docking and scoring: successes and gaps, J. Med. Chem. 49 (2006) 5851-5855.
[11] S.Y. Huang, S.Z. Grinter, X. Zou, Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions, Phys. Chem. Chem. Phys. 12 (2010) 12899-12908.
[12] T. Cheng, X. Li, Y. Li, Z. Liu, R. Wang, Comparative assessment of scoring functions on a diverse test set, J. Chem. Inf. Model. 49 (2009) 1079-1093.
[13] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, J. Med. Chem. 47 (2004) 1739-1749.
[14] E. Kellenberger, J. Rodrigo, P. Muller, D. Rognan, Comparative evaluation of eight docking tools for docking and virtual screening accuracy, Proteins 57 (2004) 225-242.
[15] R. Wang, Y. Lu, S. Wang, Comparative evaluation of 11 scoring functions for molecular docking, J. Med. Chem. 46 (2003) 2287-2303.
[16] M. Stahl, M. Rarey, Detailed analysis of scoring functions for virtual screening, J. Med. Chem. 44 (2001) 1035-1042.
[17] R. Teramoto, H. Fukunishi, Consensus scoring with feature selection for structure-based virtual screening, J. Chem. Inf. Model. 48 (2008) 288-295.
[18] T. Tuccinardi, G. Poli, V. Romboli, A. Giordano, A. Martinelli, Extensive consensus docking evaluation for ligand pose prediction and virtual screening studies, J. Chem. Inf. Model. 54 (2014) 2980-2986.
[19] G.L. Warren, C.W. Andrews, A.M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus, S. Senger, G. Tedesco, I.D. Wall, J.M. Woolven, C.E. Peishoff, M.S. Head, A critical assessment of docking programs and scoring functions, J. Med. Chem. 49 (2006) 5912-5931.
[20] N. Triballeau, F. Acher, I. Brabet, J.P. Pin, H.O. Bertrand, Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4, J. Med. Chem. 48 (2005) 2534-2547.

Accepted Manuscript

[21] K.L. Damm-Ganamet, R.D. Smith, J.B. Dunbar, Jr., J.A. Stuckey, H.A. Carlson, CSAR Benchmark Exercise 2011-2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series, J. Chem. Inf. Model. 53 (2013) 1853-1870.

[22] E. Perola, W.P. Walters, P.S. Charifson, A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance, Proteins 56 (2004) 235-249.

[23] X. Hu, S. Balaz, W.H. Shelver, A practical approach to docking of zinc metalloproteinase inhibitors, J. Mol. Graph. Model. 22 (2004) 293-307.

[24] M. Kontoyianni, L.M. McClellan, G.S. Sokol, Evaluation of docking performance: comparative data on docking algorithms, J. Med. Chem. 47 (2004) 558-565.

[25] P. Ferrara, H. Gohlke, D.J. Price, G. Klebe, C.L. Brooks, 3rd, Assessing scoring functions for protein-ligand interactions, J. Med. Chem. 47 (2004) 3032-3047.

[26] C. Bissantz, G. Folkers, D. Rognan, Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations, J. Med. Chem. 43 (2000) 4759-4767.

[27] M. Kontoyianni, G.S. Sokol, L.M. McClellan, Evaluation of library ranking efficacy in virtual screening, J. Comput. Chem. 26 (2005) 11-22.

[28] Z. Zhou, A.K. Felts, R.A. Friesner, R.M. Levy, Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets, J. Chem. Inf. Model. 47 (2007) 1599-1608.

[29] R.D. Smith, J.B. Dunbar, Jr., P.M. Ung, E.X. Esposito, C.Y. Yang, S. Wang, H.A. Carlson, CSAR benchmark exercise of 2010: combined evaluation across all submitted scoring functions, J. Chem. Inf. Model. 51 (2011) 2115-2131.

[30] C.P. Mpamhanga, B. Chen, I.M. McLay, D.L. Ormsby, M.K. Lindvall, Retrospective docking study of PDE4B ligands and an analysis of the behavior of selected scoring functions, J. Chem. Inf. Model. 45 (2005) 1061-1074.

[31] H. Chen, P.D. Lyne, F. Giordanetto, T. Lovell, J. Li, On evaluating molecular-docking methods for pose prediction and enrichment factors, J. Chem. Inf. Model. 46 (2006) 401-415.

[32] R. Wang, Y. Lu, X. Fang, S. Wang, An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes, J. Chem. Inf. Comput. Sci. 44 (2004) 2114-2125.

[33] J.B. Cross, D.C. Thompson, B.K. Rai, J.C. Baber, K.Y. Fan, Y. Hu, C. Humblet, Comparison of several molecular docking programs: pose prediction and virtual screening accuracy, J. Chem. Inf. Model. 49 (2009) 1455-1474.

[34] X. Li, Y. Li, T. Cheng, Z. Liu, R. Wang, Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes, J. Comput Chem. 31 (2010) 2109-2125.

[35] R. Wang, L. Lai, S. Wang, Further development and validation of empirical scoring functions for structure-based binding affinity prediction, J. Comput. Aided. Mol. Des. 16 (2002) 11-26.

[36] P. Tao, L. Lai, Protein ligand docking based on empirical method for binding affinity estimation, J. Comput. Aided. Mol. Des. 15 (2001) 429-446.

[37] S. Makino, T.J. Ewing, I.D. Kuntz, DREAM++: flexible docking program for virtual combinatorial libraries, J. Comput. Aided. Mol. Des. 13 (1999) 513-532.

[38] E.X. Esposito, K. Baran, K. Kelly, J.D. Madura, Docking of sulfonamides to carbonic anhydrase II and IV, J. Mol. Graph. Model. 18 (2000) 283-289, 307-288.

Accepted Manuscript

[39] D.A. Pearlman, P.S. Charifson, Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system, J. Med. Chem. 44 (2001) 3417-3423.

[40] P.D. Lyne, M.L. Lamb, J.C. Saeh, Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring, J. Med. Chem. 49 (2006) 4805-4808.

[41] C. Rapp, C. Kalyanaraman, A. Schiffmiller, E.L. Schoenbrun, M.P. Jacobson, A molecular mechanics approach to modeling protein-ligand interactions: relative binding affinities in congeneric series, J. Chem. Inf. Model. 51 (2011) 2082-2089.

[42] Q. Han, C. Dominguez, P.F. Stouten, J.M. Park, D.E. Duffy, R.A. Galemmo, Jr., K.A. Rossi, R.S. Alexander, A.M. Smallwood, P.C. Wong, M.M. Wright, J.M. Luettgen, R.M. Knabb, R.R. Wexler, Design, synthesis, and biological evaluation of potent and selective amidino bicyclic factor Xa inhibitors, J. Med. Chem. 43 (2000) 4398-4415.

[43] D.J. Pinto, M.J. Orwat, S. Koch, K.A. Rossi, R.S. Alexander, A. Smallwood, P.C. Wong, A.R. Rendina, J.M. Luettgen, R.M. Knabb, K. He, B. Xin, R.R. Wexler, P.Y. Lam, Discovery of 1-(4-methoxyphenyl)-7-oxo-6-(4-(2-oxopiperidin-1-yl)phenyl)-4,5,6,7-tetrahydro-1H -pyrazolo[3,4-c]pyridine-3-carboxamide (apixaban, BMS-562247), a highly potent, selective, efficacious, and orally bioavailable inhibitor of blood coagulation factor Xa, J. Med. Chem. 50 (2007) 5339-5356.

[44] J.R. Pruitt, D.J. Pinto, R.A. Galemmo, Jr., R.S. Alexander, K.A. Rossi, B.L. Wells, S. Drummond, L.L. Bostrom, D. Burdick, R. Bruckner, H. Chen, A. Smallwood, P.C. Wong, M.R. Wright, S. Bai, J.M. Luettgen, R.M. Knabb, P.Y. Lam, R.R. Wexler, Discovery of 1-(2-aminomethylphenyl)-3-trifluoromethyl-N- [3-fluoro-2'-(aminosulfonyl)[1,1'-biphenyl)]-4-yl]-1H-pyrazole-5-carboxyamide (DPC602), a potent, selective, and orally bioavailable factor Xa inhibitor(1), J. Med. Chem. 46 (2003) 5298-5315.

[45] M.L. Quan, P.Y. Lam, Q. Han, D.J. Pinto, M.Y. He, R. Li, C.D. Ellis, C.G. Clark, C.A. Teleha, J.H. Sun, R.S. Alexander, S. Bai, J.M. Luettgen, R.M. Knabb, P.C. Wong, R.R. Wexler, Discovery of 1-(3'-aminobenzisoxazol-5'-yl)-3-trifluoromethyl-N-[2-fluoro-4- [(2'-dimethylaminomethyl)imidazol-1-yl]phenyl]-1H-pyrazole-5-carboxyamide hydrochloride (razaxaban), a highly potent, selective, and orally bioavailable factor Xa inhibitor, J. Med. Chem. 48 (2005) 1729-1744.

[46] I.R. Hardcastle, C.E. Arris, J. Bentley, F.T. Boyle, Y. Chen, N.J. Curtin, J.A. Endicott, A.E. Gibson, B.T. Golding, R.J. Griffin, P. Jewsbury, J. Menyerol, V. Mesguiche, D.R. Newell, M.E. Noble, D.J. Pratt, L.Z. Wang, H.J. Whitfield, N2-substituted O6-cyclohexylmethylguanine derivatives: potent inhibitors of cyclin-dependent kinases 1 and 2, J. Med. Chem. 47 (2004) 3710-3722.

[47] J.D. Oslob, M.J. Romanowski, D.A. Allen, S. Baskaran, M. Bui, R.A. Elling, W.M. Flanagan, A.D. Fung, E.J. Hanan, S. Harris, S.A. Heumann, U. Hoch, J.W. Jacobs, J. Lam, C.E. Lawrence, R.S. McDowell, M.A. Nannini, W. Shen, J.A. Silverman, M.M. Sopko, B.T. Tangonan, J. Teague, J.C. Yoburn, C.H. Yu, M. Zhong, K.M. Zimmerman, T. O'Brien, W. Lew, Discovery of a potent and selective aurora kinase inhibitor, Bioorg. Med. Chem. Lett. 18 (2008) 4880-4884.

[48] M. Anzini, A. Di Capua, S. Valenti, S. Brogi, M. Rovini, G. Giuliani, A. Cappelli, S. Vomero, L. Chiasserini, A. Sega, G. Poce, G. Giorgi, V. Calderone, A. Martelli, L. Testai, L. Sautebin, A. Rossi, S. Pace, C. Ghelardini, L. Di Cesare Mannelli, V. Benetti, A. Giordani, P. Anzellotti, M. Dovizio, P. Patrignani, M. Biava, Novel analgesic/anti-inflammatory agents: 1,5-diarylpyrrole nitrooxyalkyl ethers and

Accepted Manuscript

related compounds as cyclooxygenase-2 inhibiting nitric oxide donors, J. Med. Chem. 56 (2013) 3191-3206.

[49] M. Anzini, M. Rovini, A. Cappelli, S. Vomero, F. Manetti, M. Botta, L. Sautebin, A. Rossi, C. Pergola, C. Ghelardini, M. Norcini, A. Giordani, F. Makovec, P. Anzellotti, P. Patrignani, M. Biava, Synthesis, biological evaluation, and enzyme docking simulations of 1,5-diarylpyrrole-3-alkoxyethyl ethers as selective cyclooxygenase-2 inhibitors endowed with anti-inflammatory and antinociceptive activity, J. Med. Chem. 51 (2008) 4476-4481.

[50] M. Biava, G.C. Porretta, A. Cappelli, S. Vomero, F. Manetti, M. Botta, L. Sautebin, A. Rossi, F. Makovec, M. Anzini, 1,5-Diarylpyrrole-3-acetic acids and esters as novel classes of potent and highly selective cyclooxygenase-2 inhibitors, J. Med. Chem. 48 (2005) 3428-3432.

[51] K.A. Hansford, R.C. Reid, C.I. Clark, J.D. Tyndall, M.W. Whitehouse, T. Guthrie, R.P. McGeary, K. Schafer, J.L. Martin, D.P. Fairlie, D-Tyrosine as a chiral precusor to potent inhibitors of human nonpancreatic secretory phospholipase A2 (IIa) with antiinflammatory activity, Chembiochem 4 (2003) 181-185.

[52] R.E. Mewshaw, R.J. Edsall, Jr., C. Yang, E.S. Manas, Z.B. Xu, R.A. Henderson, J.C. Keith, Jr., H.A. Harris, ERbeta ligands. 3. Exploiting two binding orientations of the 2-phenylnaphthalene scaffold to achieve ERbeta selectivity, J. Med. Chem. 48 (2005) 3953-3979.

[53] M.L. Quan, J.M. Smallheer, The race to an orally active Factor Xa inhibitor: recent advances, Curr. Opin. Drug. Discov. Devel. 7 (2004) 460-469.

[54] C.E. Arris, F.T. Boyle, A.H. Calvert, N.J. Curtin, J.A. Endicott, E.F. Garman, A.E. Gibson, B.T. Golding, S. Grant, R.J. Griffin, P. Jewsbury, L.N. Johnson, A.M. Lawrie, D.R. Newell, M.E. Noble, E.A. Sausville, R. Schultz, W. Yu, Identification of novel purine and pyrimidine cyclin-dependent kinase inhibitors with distinct molecular interactions and tumor cell growth inhibition profiles, J. Med. Chem. 43 (2000) 2797-2804.

[55] M. Hall, G. Peters, Genetic alterations of cyclins, cyclin-dependent kinases, and Cdk inhibitors in human cancer, Adv. Cancer. Res. 68 (1996) 67-108.

[56] D.H. Walker, Small-molecule inhibitors of cyclin-dependent kinases: molecular tools and potential therapeutics, Curr. Top. Microbiol. Immunol. 227 (1998) 149-165.

[57] T.G. Davies, J. Bentley, C.E. Arris, F.T. Boyle, N.J. Curtin, J.A. Endicott, A.E. Gibson, B.T. Golding, R.J. Griffin, I.R. Hardcastle, P. Jewsbury, L.N. Johnson, V. Mesguiche, D.R. Newell, M.E. Noble, J.A. Tucker, L. Wang, H.J. Whitfield, Structure-based design of a potent purine-based cyclin-dependent kinase inhibitor, Nat. Struct. Biol. 9 (2002) 745-749.

[58] M. Carmena, W.C. Earnshaw, The cellular geography of aurora kinases, Nat Rev Mol Cell Biol 4 (2003) 842-854.

[59] T. Marumoto, D. Zhang, H. Saya, Aurora-A - a guardian of poles, Nat. Rev. Cancer. 5 (2005) 42-50.

[60] H. Katayama, W.R. Brinkley, S. Sen, The Aurora kinases: role in cell transformation and tumorigenesis, Cancer. Metastasis. Rev. 22 (2003) 451-464.

[61] O. Gautschi, J. Heighway, P.C. Mack, P.R. Purnell, P.N. Lara, Jr., D.R. Gandara, Aurora kinases as anticancer drug targets, Clin. Cancer. Res. 14 (2008) 1639-1648.

[62] P.D. Andrews, Aurora kinases: shining lights on the therapeutic horizon?, Oncogene 24 (2005) 5005-5015.

[63] F. Girdler, K.E. Gascoigne, P.A. Eyers, S. Hartmuth, C. Crafter, K.M. Foote, N.J. Keen, S.S. Taylor, Validating Aurora B as an anti-cancer drug target, J. Cell. Sci. 119 (2006) 3664-3675.

Accepted Manuscript

[64] E.A. Harrington, D. Bebbington, J. Moore, R.K. Rasmussen, A.O. Ajose-Adeogun, T. Nakayama, J.A. Graham, C. Demur, T. Hercend, A. Diu-Hercend, M. Su, J.M. Golec, K.M. Miller, VX-680, a potent and selective small-molecule inhibitor of the Aurora kinases, suppresses tumor growth in vivo, Nat. Med. 10 (2004) 262-267.

[65] N. Keen, S. Taylor, Aurora-kinase inhibitors as anticancer agents, Nat. Rev. Cancer. 4 (2004) 927-936.

[66] V. Rajakrishnan, V.R. Manoj, G. Subba Rao, Computer-aided, rational design of a potent and selective small peptide inhibitor of cyclooxygenase 2 (COX2), J. Biomol. Struct. Dyn. 25 (2008) 535-542.

[67] R.G. Kurumbail, A.M. Stevens, J.K. Gierse, J.J. McDonald, R.A. Stegeman, J.Y. Pak, D. Gildehaus, J.M. Miyashiro, T.D. Penning, K. Seibert, P.C. Isakson, W.C. Stallings, Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents, Nature 384 (1996) 644-648.

[68] N. Fox, M. Song, J. Schrementi, J.D. Sharp, D.L. White, D.W. Snyder, L.W. Hartley, D.G. Carlson, N.J. Bach, R.D. Dillard, S.E. Draheim, J.L. Bobbitt, L. Fisher, E.D. Mihelich, Transgenic model for the discovery of novel human secretory non-pancreatic phospholipase A2 inhibitors, Eur. J. Pharmacol. 308 (1996) 195-203.

[69] H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide Protein Data Bank, Nat. Struct. Biol. 10 (2003) 980.

[70] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic. Acids. Res. 28 (2000) 235-242.

[71] T.A. Halgren, Identifying and characterizing binding sites and assessing druggability, J. Chem. Inf. Model. 49 (2009) 377-389.

[72] D. Ramamoorthy, E. Turos, W.C. Guida, Identification of a new binding site in E. coli FabH using Molecular dynamics simulations: validation by computational alanine mutagenesis and docking studies, J. Chem. Inf. Model. 53 (2013) 1138-1156.

[73] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking, J. Mol. Biol. 267 (1997) 727-748.

[74] Y. Diao, W. Lu, H. Jin, J. Zhu, L. Han, M. Xu, R. Gao, X. Shen, Z. Zhao, X. Liu, Y. Xu, J. Huang, H. Li, Discovery of diverse human dihydroorotate dehydrogenase inhibitors as immunosuppressive agents by structure-based virtual screening, J. Med. Chem. 55 (2012) 8341-8349.

[75] A.R. Leach, I.D. Kuntz, Conformational-Analysis of Flexible Ligands in Macromolecular Receptor-Sites, J. Comput. Chem. 13 (1992) 730-748.

[76] R.T. Kroemer, Structure-based drug design: docking and scoring, Curr. Protein. Pept. Sci. 8 (2007) 312-328.

[77] O. Trott, A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2010) 455-461.

[78] C.R. Corbeil, P. Englebienne, N. Moitessier, Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0, J. Chem. Inf. Model. 47 (2007) 435-449.

[79] R. Thomsen, M.H. Christensen, MolDock: a new technique for high-accuracy molecular docking, J. Med. Chem. 49 (2006) 3315-3321.

[80] M. McGann, FRED pose prediction and virtual screening accuracy, J. Chem. Inf. Model. 51 (2011) 578-596.

Accepted Manuscript

ACCEPTED MANUSCRIPT

[81] OEDocking. version 3.0.1, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, http://www.eyesopen.com, 2010□.

[82] P.C. Hawkins, A.G. Skillman, G.L. Warren, B.A. Ellingson, M.T. Stahl, Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database, J. Chem. Inf. Model. 50 (2010) 572-584.

[83] Omega2. version 2.5.1.4, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, http://www.eyesopen.com, 2010.

[84] G.B. McGaughey, R.P. Sheridan, C.I. Bayly, J.C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.F. Truchon, W.D. Cornell, Comparison of topological, shape, and docking methods in virtual screening, J. Chem. Inf. Model. 47 (2007) 1504-1519.

[85] D.J. Diller, K.M. Merz, Jr., High throughput docking for library design and library prioritization, Proteins 43 (2001) 113-124.

[86] S.F. Sousa, P.A. Fernandes, M.J. Ramos, Protein-ligand docking: current status and future challenges, Proteins 65 (2006) 15-26.

[87] M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, R.P. Mee, Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes, J. Comput. Aided. Mol. Des. 11 (1997) 425-445.

[88] D.L. Scott, S.P. White, Z. Otwinowski, W. Yuan, M.H. Gelb, P.B. Sigler, Interfacial catalysis: the mechanism of phospholipase A2, Science 250 (1990) 1541-1546.

[89] R. Kim, J. Skolnick, Assessment of programs for ligand binding affinity prediction, J. Comput. Chem. 29 (2008) 1316-1331.

[90] Y. Lu, Y. Liu, Z. Xu, H. Li, H. Liu, W. Zhu, Halogen bonding for rational drug design and new drug discovery, Expert Opin Drug Discov 7 (2012) 375-383.

[91] Y. Lu, T. Shi, Y. Wang, H. Yang, X. Yan, X. Luo, H. Jiang, W. Zhu, Halogen bonding--a novel interaction for rational drug design?, J. Med. Chem. 52 (2009) 2854-2862.

[92] Z. Xu, Z. Liu, T. Chen, T. Chen, Z. Wang, G. Tian, J. Shi, X. Wang, Y. Lu, X. Yan, G. Wang, H. Jiang, K. Chen, S. Wang, Y. Xu, J. Shen, W. Zhu, Utilization of halogen bond in lead optimization: a case study of rational design of potent phosphodiesterase type 5 (PDE5) inhibitors, J. Med. Chem. 54 (2011) 5607-5611.

[93] M.Z. Hernandes, S.M. Cavalcanti, D.R. Moreira, W.F. de Azevedo Junior, A.C. Leite, Halogen atoms in the modern medicinal chemistry: hints for the drug design, Curr. Drug. Targets. 11 (2010) 303-314.

[94] Y. Liu, Z. Xu, Z. Yang, K. Chen, W. Zhu, A knowledge-based halogen bonding scoring function for predicting protein-ligand interactions, J. Mol. Model. (2013).

[95] A. Merino, A.K. Bronowska, D.B. Jackson, D.J. Cahill, Drug profiling: knowing where it hits, Drug. Discov. Today. 15 (2010) 749-756.

[96] A. Mirza, R. Desai, J. Reynisson, Known drug space as a metric in exploring the boundaries of drug-like chemical space, Eur. J. Med. Chem. 44 (2009) 5006-5011.

Accepted Manuscript

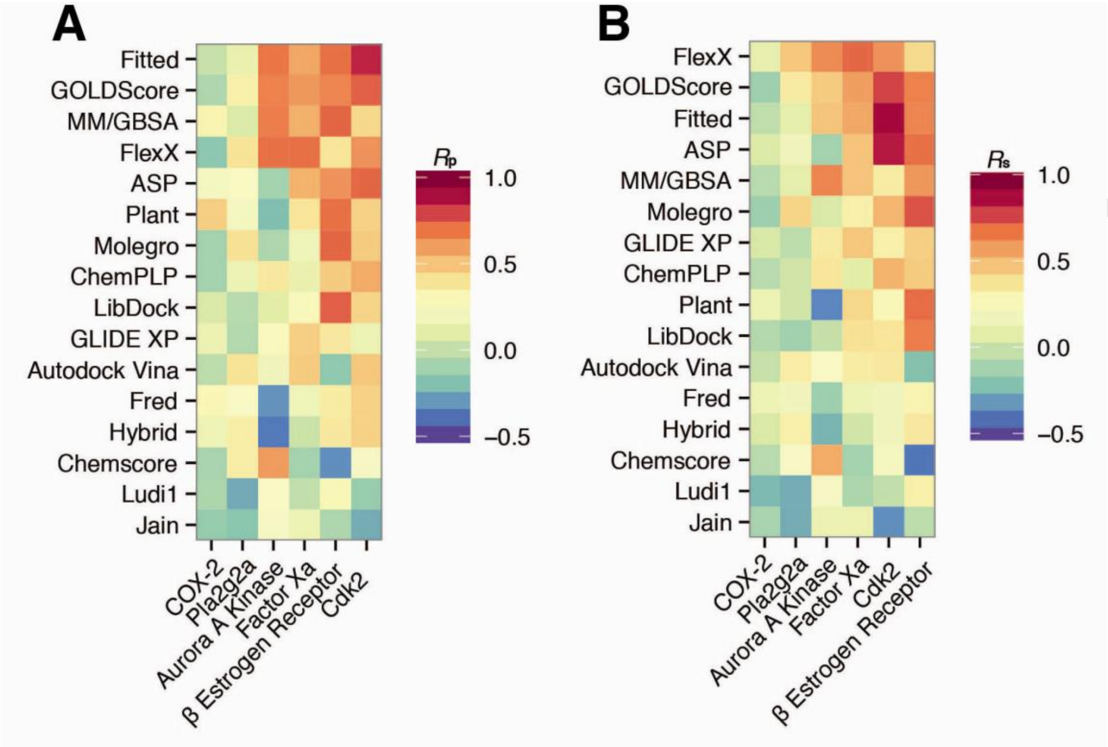
Highlights

- Eight docking programs and sixteen scoring functions were examined.
- Fitted, FlexX and GOLDScore outperformed the other programs here.
- Hydrophilic targets such as factor Xa, Cdk2 kinase and Aurora kinase are amenable to current scoring functions.
- Hydrophobic targets such as COX-2 and Pla2g2a represent challenges to scoring functions.
- No program used was effective for all six protein-ligand data sets sampled in this study.

Accepted Manuscript

ACCEPTED MANUSCRIPT

Graphical Abstract



Accepted Manuscript