JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

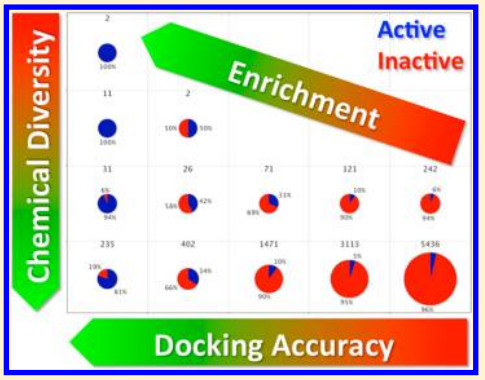
Best of Both Worlds: On the Complementarity of Ligand-Based and Structure-Based Virtual Screening

Fabio Broccatelli and Nathan Brown*

†Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, London SM2 5NG, United Kingdom

Supporting Information

ABSTRACT: Virtual screening with docking is an integral component of drug design, particularly during hit finding phases. While successful prospective studies of virtual screening exist, it remains a significant challenge to identify best practices a priori due to the many factors that influence the final outcome, including targets, data sets, software, metrics, and expert knowledge of the users. This study investigates the extent to which ligand-based methods can be applied to improve structure-based methods. The use of ligand-based methods to modulate the number of hits identified using the protein−ligand complex and also the diversity of these hits from the crystallographic ligand is discussed. In this study, 40 CDK2 ligand complexes were used together with two external data sets containing both actives and inactives from GlaxoSmithKline (GSK) and actives and decoys from the Directory of Useful Decoys (DUD). Results show how ligand-based modeling can be used to select a more appropriate protein conformation for docking, as well as to assess the reliability of the docking experiment. The time gained by reducing the pool of virtual screening candidates via ligand-based similarity can be invested in more accurate docking procedures, as well as in downstream labor-intensive approaches (e.g., visual inspection) maximizing the use of the chemical and biological information available. This provides a framework for molecular modeling scientists that are involved in initiating virtual screening campaigns with practical advice to make best use of the information available to them.



INTRODUCTION

Virtual screening is a well-established hit finding strategy in drug discovery,[1] whereby scientists with access to large compound libraries cherry-pick compounds for high-throughput screening.[2] Vast collections of chemical structures available from compound vendors can also be screened in silico to obtain a short list of compounds for purchase and testing.[2−4] Generally, the aim of virtual screening campaigns is to explore and maximally exploit the available chemical space at minimum cost to identify virtual hits that can be experimentally confirmed and followed up as part of a drug discovery program. While the number of experimentally validated hits retrieved is an important component of virtual screening campaign success, an equally important, and often neglected, component is the chemical diversity of such hits.[5] A sufficiently high number of diversified hit scaffolds will, most likely, provide a more solid foundation for a drug discovery program and mitigate the risk of chemical series-specific issues such as poor physicochemical property space, challenging synthetic accessibility, poor selectivity, adverse metabolism, or toxicity profiles. Computational chemists rely on a number of different families of algorithms to identify novel hits, which mainly differ in the level of chemical abstraction, protein and/or ligand awareness, and computational expense. Different algorithms are often compared in chemoinformatics studies.[6−8] While this practice can provide valuable insights into the different methodologies used in virtual screening, it has intrinsic limitations that prevent scientists from drawing definitive conclusions on the best computational strategy to approach a novel target. Methodological bias may come from the expertise of the users, from the use of decoys (rather than inactives), from difficulty in assigning the appropriate parameters for the methods applied, and from the potential error in data sets and protein structures.[6−10] Furthermore, performing truly prospective studies in a fair and unbiased manner is challenging because the structures of the known hits are, to some extent, influenced by the choice of protein−ligand complex conformation under study.[6] In our practical experience, different computational tools show complementary features and are used frequently in combination to optimize virtual screening cascades.

In this study, we investigate the use of ligand-based similarity to enhance the performance of ligand docking in virtual screening.[11] It has long been recognized that docking success rate increases with the similarity of the screened molecule to the ligand bound to the protein conformation under test (herein described as the native ligand).[12,13] As the structural biology community continually increases exemplification of the druggable genome with protein X-ray crystal structures, drug discovery programs more frequently have access to multiple

Received: March 13, 2014

   dx.doi.org/10.1021/ci5001604 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

structures, and it is important to understand which structures are likely to be most applicable to virtual screening. The hypothesis under test is that ligand-based similarity methods can be applied to identify the most appropriate protein structures available to a virtual screening campaign. This approach has been previously suggested by Sutherland et al.,[14] who showed that using a protein structure selected based on the similarity to the crystallographic ligand reduces the success rate gap between native docking (docking the native ligand back into the apoprotein structure derived from that ligand) and cross docking (docking a non-native ligand into the protein structure).[12−17] In this study, we expand this analysis to three different docking protocols differing in accuracy and speed, as well as to different external data sets representative of virtual screening scenarios. The comparisons were performed on the set of native ligands, on a set containing confirmed actives and decoy ligands from the directory of useful decoys (DUD) data set,[9] and on a published set of active and inactive ligands for the protein under study (GlaxoSmithKline; GSK).[18] For cross-docked ligands, it was possible to compare the docking poses with the original native poses by calculation of the root-mean-square deviation (RMSD) geometric distance in ångströms (Å) between the poses. However, this was not possible for the DUD and the GSK data sets. Hence, the two data sets were analyzed based on their docking scores. We hypothesized that docking a molecule into the protein structure derived from its most similar crystallographic ligand would be closer in success rate to native ligand docking than to cross docking. We found that this approach significantly enhances the enrichment of active compounds within larger data sets compared to docking on a single protein structure. Furthermore, this approach reduces the computational expense with minimal reduction in success rate compared to ensemble docking (which in this work refers to cross docking applied to all the available protein conformations except for the native one). We also demonstrate how molecular modelers could increase the diversity of virtual screening hits while minimizing loss of accuracy.

The protein target cyclin-dependent kinase 2 (CDK2) has previously been used as a model system in numerous cross docking and virtual screening studies;[6,8−10,12,13,15,17,19] multiple high-quality crystallographic structures are available from the Protein Data Bank (PDB) representative of a diversity of ligand space for CDK2.

Overall, the focus of this study is to identify the best practice that uses the available ligand and structure data and not to compare different computational methods for docking and similarity calculations. Therefore, the algorithms applied in this study are widely adopted as standards for fingerprint similarity (Pipeline Pilot Extended Connectivity Fingerprints),[20] three-dimensional (3D) ligand-based similarity (ROCS Tanimoto combo),[21] and docking (Glide).[22] The choice of the software does not reflect our absolute preferences and represents some of the commonly used virtual screening tools from the scientific literature.

The results from the docking experiments were analyzed using statistical measures, accuracy and enrichment, computational time, and chemical diversity of the identified hits. Limitations regarding the biases previously discussed for virtual screening studies also apply to this study; however, we believe that these results offer informative insights for computational scientists approaching a novel drug discovery project to make the best use of the data available.

MATERIAL AND METHODS

Selection of the CDK2 Structures. Crystallographic PDB structures were selected if they were published after 2010 with a resolution better than 2 Å. The ligands of these 92 structures were extracted and their Morgan fingerprints generated using the RDKit implementation (number of bits = 1024, radius = 2) available in KNIME v2.6.3 for Mac.[23,24] The fingerprints were processed using the RDKit Diversity Picker to select 40 representative ligands; the tool is based on the MaxMin algorithm as presented by Ashton et al.[25] Each of the selected 40 structures was visually inspected and considered acceptable only if the ligand was present in a single conformation, a hinge binder, and exclusively interacting with the protein pocket and/or with water molecules. Structures that failed to pass this further screening were replaced by the closest analogues not included in the original selection using the Tanimoto similarity of the Morgan fingerprints. The PDB codes of the selected structures can be found in Table S1 of the Supporting Information. The mean pairwise similarity using ECFP_4 fingerprints observed for the set of crystallographic ligand was 0.20, which indicates a sufficiently high structural diversity.

Assembly and Preparation of Data Sets. Several authors have reported a strong dependency between virtual screening performance and the data sets used.[6] In this study, a standard data set was used for CDK2 virtual screening studies (DUD data set) and a data set recently disclosed by GlaxoSmithKline scientists extracted from ChEMBL (GSK).[19,26] The DUD data set was originally presented by Huang et al.[9] and was downloaded from the DUD Web site.[27] Compounds available in the DUD data set are either binders or decoys that have similar physicochemical properties to the binders and can be reported in different protonation states. After structure standardization and removal of duplicates using MOE extensions for KNIME,[28] 1829 compounds remained for CDK2 in the DUD data set, of which 50 were actives. The GSK data sets originally included 367 compounds reported with percent inhibition data for CDK2 at two concentrations and measured using two different protocols (four data points for each molecule). The four single-point data sets showed comparable outcomes; the data sets were harmonized and reduced to 24 actives and 270 inactives as follows: Inactive compounds were defined as those with less than 10% inhibition, while the actives were those compounds with greater than 20% inhibition. Compounds were excluded where the outcome of the four assays was not unanimous (72 compounds).

The three data sets (DUD, GSK, Native Ligands) were standardized (stereochemistry, keep largest fragment) using the Standardize Molecules component in PipelinePilot v8.0.1.500 for Windows; the molecules were minimized using CORINA in Pipeline Pilot. Compounds to be docked were further processed in KNIME using the Schrödinger tools Tautomerizer (maximum of two tautomers) and Epik (solvent H_2O, predicted states at pH 7 ± 1).[29] For compounds to be used for 3D similarity search, 25 conformations were produced using OMEGA v2.4.6 (OpenEye) for Mac with standard options (except for −strictstereo false, to allow molecules with unspecified stereocenters to be processed).[30]

Ligand Similarity, Docking, Data Production, and Elaboration. For the molecules in the three data sets, the Tanimoto fingerprint similarity to the crystallographic ligands was calculated using PipilinePilot Extended Connectivity

dx.doi.org/10.1021/ci5001604 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Fingerprints (ECFP_4). 3D Similarity was calculated using ROCS v3.1.1 (OpenEye) with standard options (Tanimoto Combo similarity).[21] The protein−ligand complexes were inspected manually to assign the appropriate protonation state and bond order; the process was guided by chemical judgment and by the data reported in the original publications. All the water molecules were removed. The PDB structures were prealigned using the KNIME node Align Binding Sites (Schrödinger).[29] The grids were produced using the KNIME node Glide Grid Generation (Schrödinger) after preparing the proteins with the (Schrödinger) KNIME node Protein Preparation Wizard.[29] Docking was performed using the KNIME node Glide Ensemble Docking (Schrödinger), with three different protocols (HTVS, SP, and XP), docking flexibly, penalizing the nonplanar conformations of amide bonds, and adding Epik state penalties to the docking score.[29] While in the first part of the study (native docking and cross docking experiments), three different binding hypotheses were produced and compared to the crystallographic ligand conformation; for the remaining part of the study, only the pose with the highest score was considered. The RMSD of the heavy atoms of the docked ligand from the prealigned crystallographic ligand was calculated using the KNIME node RMSD (MOE).[28] The docking experiment was considered successful if the resulting pose had a RMSD of less than 2 Å from the crystallographic ligand. This criterion was selected because it has been used by a number of other investigators in recent studies.[12,13,17,19] Data were elaborated and plots were produced using the standard KNIME nodes, Microsoft Excel Office 2011[31] (bar charts), Vortex v2013.08.24252[32] (boxplots, scatter plots, pie chart plots) and RStudio v0.98.484 (package enrich vs v0.0.5 for calculating AUC and BEDROC at α = 20).[33,34] An exhaustive description of the BEDROC metric and its utility for comparing virtual screening methods has been previously presented by Truchon and Bayly.[35] A number of data failed to be produced using the methodology described above, particularly docking and 3D similarity. However, greater than 96% expected data points were produced; hence, the reasons for calculation failures were only marginally investigated. Calculation failures in the docking experiments were mainly due to the lack of good solutions identified by algorithms (as demonstrated by the higher number of missing values when using the low precision protocol HTVS) and to the mismatch between the protonation states calculated by the software and through the manual inspection step, which made it impossible to calculate the RMSD in cross-docking experiments.

RESULTS AND DISCUSSION

Native Docking versus Cross Docking. During virtual screening campaigns, docking is often used to provide binding mode hypotheses for compounds with unknown activity against the target of interest and to score these binding hypotheses. For this approach to be successful, two fundamental requirements must be met: the ligand must fit into the non-native protein conformation and high docking scores must be associated with poses that are close to those experimentally observed.

In this study, Glide was used to dock 40 ligands into their native CDK2 protein conformation (native docking) and also into the remaining 39 protein conformations (cross docking). In order to produce an RMSD value for cross-docking experiments, the 40 protein conformations were prealigned. The experiment was repeated using three different docking options, which mainly differ in the balance of accuracy versus speed. HTVS docking may be used to screen a million compounds in approximately 17 days using a single processor.[29] This speed range is applicable to a docking-based virtual screening campaign using several thousands to millions of virtual compounds. SP docking is the same as HTVS in terms of scoring function but uses a more exhaustive conformational sampling, with an estimated 10-fold increase in the computational time required compared to HTVS.[29] Finally, XP is based on a different scoring function and different criteria for the conformational search component. This option is the slowest (240 times slower with respect to HTVS) and the most prone to penalize the lack of conformational complementarity between ligand and protein.[29]

Results for the cross-docking experiments are reported as the percentage of successful poses observed, where success is measured as an RMSD of below 2 Å between the crystallographic and docked poses and is referred to as docking success rate (Figure 1). We observed no significant difference between

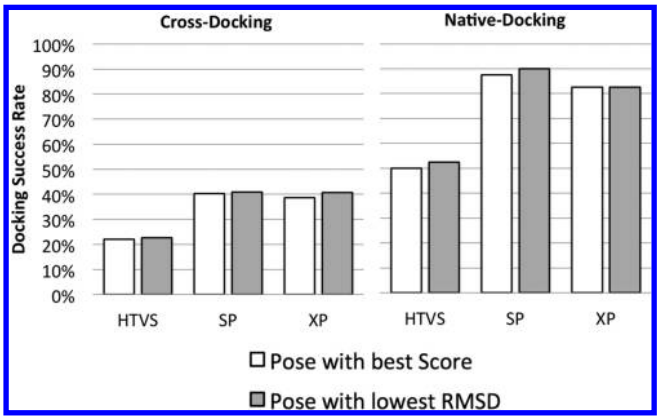


Figure 1. Docking success rate for three Glide protocols: HTVS, SP, and XP. Docking pose prediction is considered correct if the RMSD from the crystallographic ligand is below 2 Å.

considering one or three poses resulting from a docking experiment; the RMSD associated with the highest scoring pose is, in most cases, the lowest or comparable to the lowest. Hence, using one binding pose during virtual screening campaigns was accepted as a reasonable approximation.

Glide docking consistently demonstrated a 2-fold (2.19 ± 0.07) reduction in the docking success rate in cross docking with respect to native docking; this was independent of the docking protocol adopted. The two most accurate docking protocols (SP and XP) showed comparable success rate in native docking (SP = 87.5% and XP = 82.5%) and in cross docking (SP = 40% and XP = 38.7%), while the docking success rate for HTVS was significantly lower (50% in native protein conformations and 22% in cross docking). Overall, the approximately 10-fold increase in computational time from HTVS to SP increases the docking success rate by greater than 30% in native docking and greater than 15% in cross docking. This is exclusively due to the more exhaustive conformational sampling.

Unfortunately, in a prospective application, the user will have no knowledge of the RMSD, and the success of the docking experiment must be judged based on the docking score alone. Therefore, the distribution of the successful docking experiments was analyzed with respect to the docking score range, where a lower docking score suggests a better docking pose,

dx.doi.org/10.1021/ci5001604 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

and the docking success rate observed in each of these ranges (Figure 2). For this data set, it appears that across the three

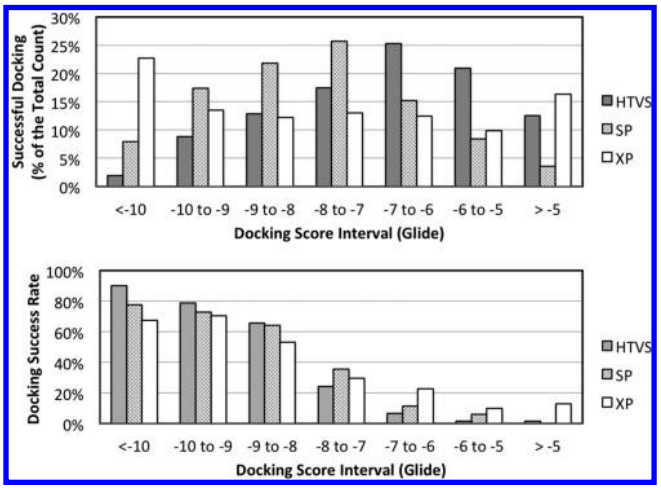


Figure 2. Distribution of the relative percentage of the accurate docking experiments with respect to the docking score interval (upper plot). Variation of the docking success rate with respect to the docking score interval (lower plot).

docking protocols the docking success rate is greater than 50% when the docking score is better than −8 and over 70% when the score is better than −9. Glide SP and XP produce a similar fraction of accurate poses with scores better than −8 (47−48%), while this percentage is only 24% for HTVS. Using high-speed docking seems to reduce the fraction of successful docking poses (RMSD below 2 Å) associated with docking score better than −8 by one-half.

Multiple Protein Conformations in Virtual Screening. The results discussed so far highlight two key points: (1) The conformational complementarity between ligand and protein improves the docking success rate. (2) More thorough ligand conformational sampling also improves success rate while increasing the computational time. Considering the time required for docking calculations and results analysis, in addition to the limitations related to commercial software licenses, it is prohibitive to apply a docking-based virtual screening campaign to screen a large library of compounds (e.g., all commercially available compounds) based on more than one crystal structure. Hence, if several crystal structures are available before starting a virtual screening campaign, it is important to identify the most appropriate method to select the optimal structure. There is a relationship between the similarity to the crystallographic ligand and the probability of successful docking, as suggested by our observed difference in docking success rate between native docking and cross docking and consistent with previous work in this area.[12,13,15−17] Therefore, docking compounds that belong to a congeneric series for which a crystal structure is available is approximately as accurate as native docking. At the opposite end of the spectrum, compounds that are highly dissimilar to the native ligand are more likely to induce unexplored conformational changes in the protein structure; hence, the probability of docking success is lower.

Cross-docking experiments were performed using 39 structures, and the results compared to those obtained with docking on a single structure selected using ligand-based similarity (either ECFP_4 or ROCS Tanimoto Combo, to native docking, and to ensemble cross docking (based on 39 non-native protein conformations, the selected pose is the one associated with the lowest docking score) (Figure 3). These

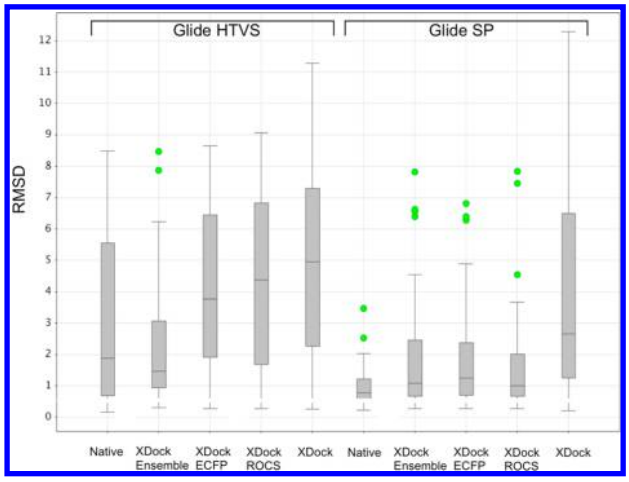


Figure 3. Box plot describing the distributions of the RMSDs of predicted binding modes from crystallographic ligands using different docking approaches and protocols. The box is delimited by the 25th and the 75th percentile values; the line within the box represents the median. The whiskers represent maximum and minimum values for non-outliers. Data points are defined as outliers (green dots) if their distance from the 75th percentile is over 1.5 times the interquartile range (distance between the 25th and 75th percentile).

experiments were conducted with Glide HTVS and SP docking protocols (XP was not considered here because our data suggest that this option does not offer advantages over the faster SP in virtual screening). From Figure 3, the following observations can be made:

- For Glide SP, selecting a specific protein structure for cross docking based on ligand similarity does not yield an inferior success rate with respect to performing cross-docking experiments on all the protein structures available and selecting the docking pose corresponding to the best docking score.
- For Glide SP, ligand-based methods for selecting the protein structure used in docking experiments appear to perform comparably well for this data set; for Glide HTVS, the selection guided by ECFP_4 fingerprints performs slightly better.
- In each case, the median RMSD of experiments based on a specific protein conformation, selected by ligand-based similarity methods, is lower than the median RMSD observed for cross docking; in most cases, the difference is greater than 1 Å.
- For Glide SP, the medians of native docking, ensemble cross docking, and docking on a single protein conformation selected using ligand-based similarity are all below 1.3 Å, which is less than half of the median RMSD observed in cross-docking experiments (2.66 Å).
- Ensemble cross docking using Glide SP is more accurate than native docking using Glide HTVS. Thus, suboptimal ligand conformational sampling, the only difference between Glide HTVS and SP, affects docking success rate more than the use of a suboptimal protein conformation (cross docking versus native docking). When using ensemble docking with Glide HTVS, the ligand conformations explored increase by a factor of 39 for this data set (number of non-native proteins conformations), while additionally including multiple relevant

dx.doi.org/10.1021/ci5001604 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

protein conformations. As a result, when using Glide HTVS, ensemble docking is more accurate than native docking (which is not true for Glide SP). Furthermore, ensemble docking using Glide HTVS is also significantly more accurate than cross docking using Glide SP.

In summary, if many protein conformations are available, the docking success rate can be drastically improved by running calculations on a single protein structure selected by means of fingerprint similarity to the respective crystallographic ligand. This is particularly true for Glide SP, as the algorithm performs a significantly more exhaustive ligand conformational search, which leads to a considerable difference in docking success rate compared to Glide HTVS. These observations are likely to be reproduced for different targets according to the extent of the ligand-dependent protein conformational change.

Application to External Data Sets. Ultimately, in virtual screening, we would like to discern active compounds from inactives. Many studies have been presented that compare different virtual screening methods.[6−8] However, there is no consensus on which method is the best. This is likely due to the many, often in controllable, variables affecting the outcome of the analysis:

- Studies should be prospective. Molecules similar to the crystallographic ligand will be synthesized after its discovery, including the structurally related compounds in the virtual screening data set that may inherently affect the objectivity of the study.
- Different compound data sets may lead to completely different results.
- Different protein targets may lead to different conclusions.
- The users performing the studies may be more expert with one method than another.
- The options of several methods could be pre-optimized, and this could boost the success rate.
- The metrics used for judging the performance of a virtual screening campaign is affected by the characteristics of the data set (e.g., ratio of ligands to decoys).
- The criteria used for selecting the inactives or decoys can affect the enrichment factors (e.g., pharmacophoric prescreening, physicochemical similarity, or random).

In this study, two data sets were used: the GSK data set composed of 294 compounds (24 confirmed actives and 270 confirmed inactives versus CDK2) and the DUD data set composed of 1829 compounds (50 confirmed actives against CDK2, the remaining compounds are decoys). We observed no comparability between the AUC (area under the receiver operator characteristic curve) in virtual screening using the three different in silico technologies (ECFP_4 fingerprint similarity, ROCS 3D similarity, and Glide docking) applied to the GSK and the DUD data sets, suggesting that the same virtual screening strategies applied to different data sets do not generally perform comparably well (Figure 4). Overall, the average AUC for 2D methods (0.51 ± 0.15) was lower than for 3D ligand-based methods (0.61 ± 0.15) and docking (0.71 ± 0.15); thus, in this case, the accuracy (quantified as AUC) was directly proportional to the computational expense. It is interesting to note that other studies comparing different screening technologies (and using different metrics) reached completely different conclusions from ours as well as from each other, highlighting the variability associated with different users, data sets, and software.[6−8] It is also noteworthy that our data set contained examples of ligands that could be used as templates for ECFP_4 similarity searches yielding higher AUC

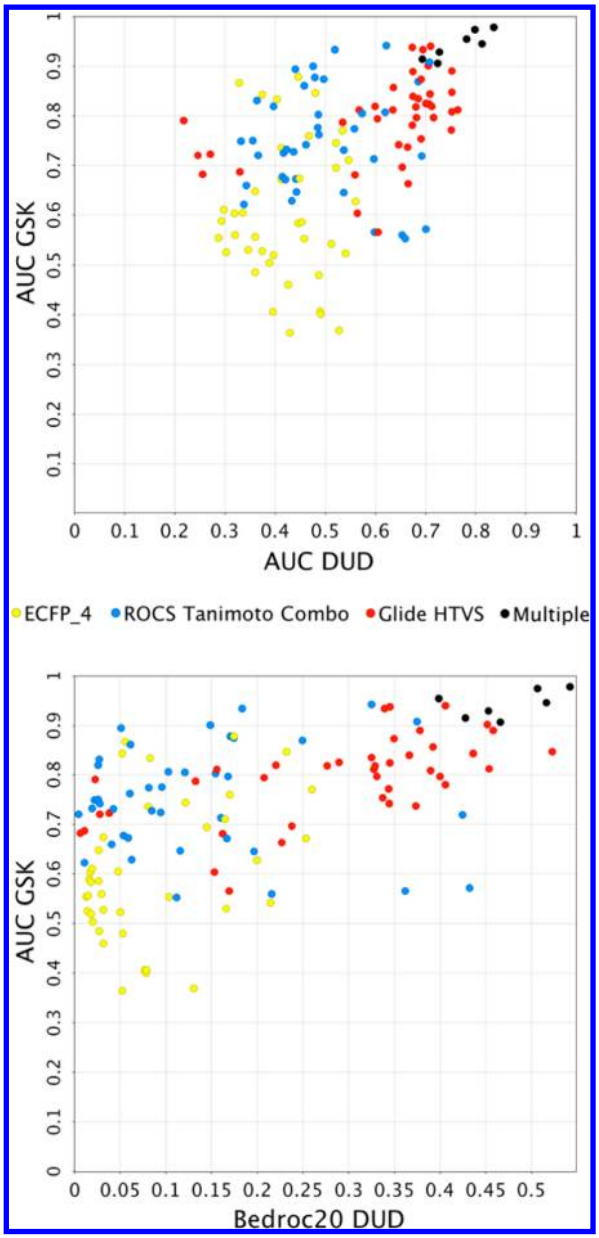


Figure 4. Screening success for different virtual screening campaigns using the DUD and the GSK data sets. For the DUD data sets, both the AUC (representing the ability to distinguish actives from inactives) and the BEDROC at $\alpha = 20$ (representing the ability to associate the highest scores to active molecules) are considered due to the high ligand-to-decoy ratio. Methods other than "Multiple" are only aware of either a single protein structure or a single crystallographic ligand.

with respect to the ROCS 3D similarity search and HTVS Glide docking on the corresponding protein structure (e.g., fingerprint similarity based on PDB 4EZ3 produced AUC = 0.87 when applied to the GSK data set, see Supporting Information). In summary, starting from a single protein−ligand complex, there is no way a priori to know which method will perform best. For the DUD data set, performances are reported both in terms of AUC and BEDROC at $\alpha = 20$, which is often reported as a metric suitable for evaluating the "early enrichment".[35] From the two plots, the same conclusion can be derived; using methods aware of multiple protein−ligand complexes consistently increases the chances of virtual screening success rates.

dx.doi.org/10.1021/ci5001604 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Performances of Virtual Screening Strategies Aware of 40 CDK2 PDB Structures[a]

row	score	AUC GSK	CCR GSK	AUC DUD	BEDROC20 DUD
1	Ens HTVS	0.95	0.92	0.81	0.53
2	Ens ROCS_TC	0.95	0.89	0.78	0.43
3	Ens ECFP_4	0.91	0.88	0.69	0.43
4	HTVS (ROCS_TC)	0.93	0.87	0.73	0.45
5	HTVS (ECFP_4)	0.91	0.89	0.72	0.47
6	Z2 HTVS-ROCS_TC-ECFP_4	0.98	0.93	0.84	0.54
7	Z2 HTVS(ECFP_4)-ROCS_TC-ECFP_4	0.97	0.92	0.80	0.51

[a]AUC is reported for the DUD and the GSK data sets. The average between sensitivity and specificity (CCR) is reported for the GSK data sets, for which inactives are available. The BEDROC20 is reported for the DUD data set due to the high ligand to decoy ratio. "Ens" is the abbreviation of ensemble, where the best score is selected after screening against all the crystallographic structures available. HTVS ROCS_TC and HTVS ECFP_4 identify HTVS docking experiments based on a protein structure selected by means of ligand-based similarity. Z2 identify data fusion approaches based on ensemble HTVS docking, ensemble ROCS Tanimoto Combo, and ensemble ECFP_4 similarity. HTVS docking on a single protein structure was selected using ECFP_4 similarity, ensemble ROCS Tanimoto Combo, and ensemble ECFP_4.

Methods using information from multiple protein−ligand complexes are reported in Table 1. For the GSK data set, the correct classification rate (CCR, which is the average between specificity and sensitivity) is reported due to the availability of experimentally confirmed inactives. CCR values in Table 1 meet all of the requirements typically expected by predictive QSAR classification models.[36] Each of the methods in Table 1 outputs only one score, which is not weighted or elaborated by any statistical tool, and therefore could be more generalizable with respect to QSAR models, which are inevitably influenced by the size of the data set, chemotypes contained in the training set, number of actives versus inactives, and likelihood of overfitting.

Ensemble methods in Table 1 (rows 1−3) use the best score across the panel of 40 protein−ligand complexes for each screened molecule. While for ligand-based methods, ensemble scoring could easily be performed in virtual screening campaigns, for docking the number of processors, licenses and time available will necessarily be limiting factors. Rows 4 and 5 are related to the use of Glide HTVS docking on a single protein conformation selected by means of ECFP_4 or ROCS Tanimoto Combo similarity to the crystallographic ligand pose. These methods use the information from every protein complex available, produce binding hypotheses as output, and can be completed in a reasonable computational time. The last two methods reported in Table 1 are based on the data fusion metric Z_2, which is derived by the Fisher z-transformation of the distributions of each one of the considered scores (ECFP_4, ROCS Tanimoto Combo, Glide HTVS). Such transformation provides a unified framework to numerically compare data belonging to the different distributions. The caveat in this case is that the sign of the Z values associated with Glide HTVS is inverted, as negative docking values identify better solutions, while for ECFP_4 and ROCS scores, the opposite is true. Z_2 is the average of the value of the two highest Z-scores among the three considered.[8] In Table 1, Z_2 values are derived either from ensemble docking (on the 40 CDK2 protein structures) or from docking on a single protein conformation selected by means of ECFP_4 Tanimoto similarity to the native ligand. For the GSK and the DUD data sets, each of the methods in Table 1 performs excellently. The best AUC and BEDROC values were observed when using all of the available in silico technologies and all of the available protein complexes (row 7, Table 1). Not surprisingly, this screening strategy is also the most computationally intensive, both in terms of time and software licenses required. Overall, ensemble docking and data fusion methods consistently deliver the best results.

Each method in Table 1 has advantages and limitations. While ligand-based ensemble methods offer the advantage of exploiting a larger amount of experimental information in a reasonable computational time, docking on a selected structure has the advantage of offering binding poses as a result. The best identified poses can be further optimized using more accurate docking protocols (e.g., Glide SP), visual inspection, torsion analysis, tautomer/protomer analysis, and if a hypothesis is supported by experimental data (e.g., crystallography, NMR, SAR), binding-mode testing. The latter refinement steps have a considerably lower throughput and a higher degree of chemical knowledge and user input required but could be crucial for further enriching the docking hit selection with experimental hits.

Exploration versus Exploitation. The plots in Figure 5 compare the relative number of actives for the two external data sets in different ranges of docking scores and ECFP_4 Tanimoto similarity to the native ligand. In this scenario, fingerprint similarity serves as a reliability index for docking experiments; for docking poses with scores better than −8, the percentage of actives increases with the fingerprint similarity. While the relative percentage of actives is different in the two data sets, the enrichment trend is the same. Practical implications of these plots are that approximately 90% of the screening library could be filtered by means of ECFP_4 similarity (below 0.2), while for the remaining compounds, Glide SP docking (as opposed to HTVS) could be used in a comparable computational time. While compounds with high fingerprint similarity (e.g., greater than 0.4) and good docking score (e.g., better than −9) would most likely be virtual screening hits if a Z_2 approach were used, there is a significant number of compounds with intermediate ligand similarity (0.2−0.4) that would not be identified by data fusion. These compounds would by definition be different from the native ligand and could therefore bring novelty to medicinal chemistry projects. The docking success rate for compounds with intermediate similarity to the native ligand is lower with respect to docking applied to compounds with high similarity; however, this is still significantly higher than random. Depending on the available resources together with the objectives of the medicinal chemistry project, computational and medicinal chemists can choose the extent to which they intend to favor success rate over novelty or vice versa.

dx.doi.org/10.1021/ci5001604 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

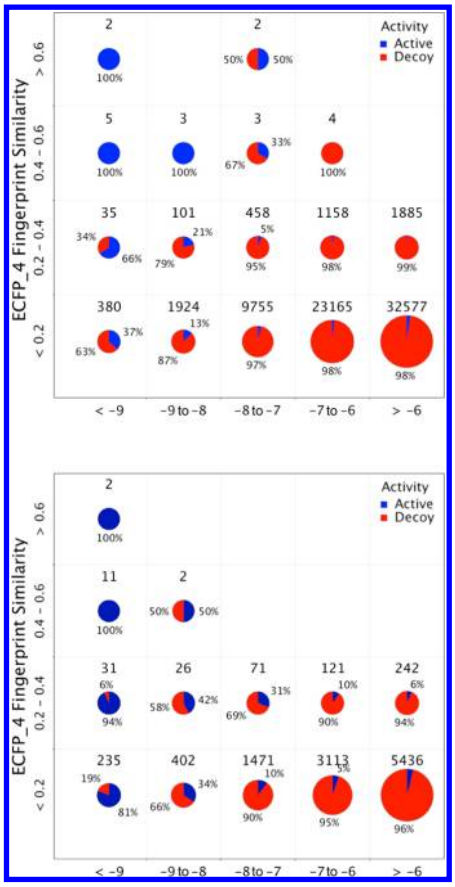


Figure 5. Relative percentage of actives and decoys (DUD data set, upper plot) and actives and inactives (GSK data set, lower plot) in different ECFP_4 and HTVS Glide docking score ranges.

CONCLUSIONS

Many different computational methods are available to modelers. However, the complementarity of these methods and the extent to which they may be combined to improve true positive hit rates has had little exploration. Here, ligand-based methods (molecular similarity) have been applied to provide guidance on probabilities of success of molecular docking strategies. These experiments provide a framework from which experiments may be designed to incorporate learning from both ligand-based and structure-based modeling methodologies. The value of the experimental data is maximized, which will result in solutions that will more likely have interpretable hypotheses with the ability to more robustly interpret negative data that may be important for medicinal chemistry design.

It is well documented that it is impossible to know a priori which computational methodology will be most appropriate, in terms of providing true positives, when starting a virtual screening campaign.[6−8] The incorporation of multiple protein−ligand complexes, where available, and the application of multiple computational methodologies using data fusion is preferable to using a single protein conformation in a virtual screening approach. When multiple protein structures are available, we recommend the application of ligand-based methods to select an appropriate protein conformation to maximize the ligand similarity of compound sets to be docked. This approach will also provide a probable confidence indicator for the compounds to be docked. In particular, cross docking on the protein structure bound to the most similar crystallographic ligand available significantly increases the success rate of the docking method, regardless of the degree of precision of the algorithm.

Although it is important to have a reliable and robust method, to the extent that the methods and data permit, the structural novelty of solutions is also of significant importance. The applicability domain of any model can tested to its limits in the search for novel hit matter, but caution must be exhibited to ensure that the resulting predictions are appropriate and robust. The discovery of a single, yet structurally novel, hit ligand can be much more beneficial and informative in the early stages of a drug design project than the identification of many close analogues of a known ligand.

According to the results of this study, if for a project there is a high number of crystallographic protein−ligand complexes available, data fusion and ensemble docking can provide classification accuracy that is comparable or superior to classification accuracy and AUC values identifying predictive QSPR models; for the methods analyzed, the correct classification rate (average between specificity and sensitivity) was always greater than 85% (and the AUC was greater than 0.9) when predicting a set of 294 experimentally tested CDK2 inhibitors and non-inhibitors. Furthermore, these models have the potential to be more generalizable allowing the identification of novel and diverse ligands.

The framework presented here offers an approach to applying ligand docking in drug design projects using the most appropriate co-crystal structure according to ligand similarity. This leads to higher confidence in docking results using structures with similar ligands but also a more thorough understanding of the confidence of docking results with structures that have less similar ligands. The results of this study offer a new approach to allowing modelers greater control of the utility of ligand docking in drug design projects.

ASSOCIATED CONTENT

Supporting Information

The PDB codes of the structures used in this study, as well as the results for the calculations described. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +44 (0) 20 8722 4033. E-mail: Nathan.Brown@icr.ac.uk.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. F.B. and N.B. are supported by the Cancer Research UK Grant C309/A11566. We also acknowledge helpful discussions with Prof. Julian Blagg.

Notes

The authors declare no competing financial interest.

ABBREVIATIONS

AUC, area under the receiver operator characteristic curve; CCR, correct classification rate; CDK2, cyclin-dependent kinase 2; DUD, directory of useful decoys; extended-connectivity fingerprints, ECFP; GSK, GlaxoSmithKline; HTVS, high-throughput virtual screening; PDB, protein data bank; SP, standard precision; Tanimoto combo, TC; VS, virtual screening; XP, extra-precision

REFERENCES

(1) Sotriffer, C.; Mannhold, R.; Kubinyi, H.; Folkers, G. *Virtual Screening: Principles, Challenges, and Practical Guidelines*; John Wiley & Sons: New York, 2011.

(2) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2001**, *1*, 882−894.

(3) Villoutreix, B. O.; Eudes, R.; Miteva, M. A. Structure-based virtual ligand screening: Recent success stories. *Comb. Chem. High Throughput Screening* **2009**, *12*, 1000−1016.

(4) Muratore, G.; Goracci, L.; Mercorelli, B.; Foeglein, A.; Digard, P.; Cruciani, G.; Palú, G.; Loregian, A. Small molecule inhibitors of influenza A and B viruses that act by disrupting subunit interactions of the viral polymerase. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 6247−6252.

(5) Muegge, I. Synergies of virtual screening approaches. *Mini-Rev. Med. Chem.* **2008**, *8*, 927−933.

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

(7) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(8) Sastry, G. M.; Inakollu, V. S. S.; Sherman, W. Boosting virtual screening enrichments with data fusion: Coalescing hits from two-dimensional fingerprints, shape, and docking. *J. Chem. Inf. Model.* **2013**, *53*, 1531−1542.

(9) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(10) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771−784.

(11) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* **2014**, *57*, 3186−3204.

(12) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein−ligand docking against non-native protein conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214−2225.

(13) Tuccinardi, T.; Botta, M.; Giordano, A.; Martinelli, A. Protein kinases: Docking and homology modeling reliability. *J. Chem. Inf. Model.* **2010**, *50*, 1432−1441.

(14) Sutherland, J. L.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J. Chem. Inf. Model.* **2007**, *47*, 2293−2302.

(15) Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337*, 209−225.

(16) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47*, 45−55.

(17) Duca, J. S.; Madison, V. S.; Voigt, J. H. Cross-docking of inhibitors into CDK2 structures. 1. *J. Chem. Inf. Model.* **2008**, *48*, 659−668.

(18) Dranchak, P.; MacArthur, R.; Guha, R.; Zuercher, W. J.; Drewry, D. H.; Douglas, S. A.; Inglese, J. Profile of the GSK published protein kinase inhibitor set across ATP-dependent and-independent luciferases: Implications for reporter-gene assays. *PLoS One* **2013**, *8* (3), e57888.

(19) Voigt, J. H.; Elkin, C.; Madison, V. S.; Duca, J. S. Cross-docking of inhibitors into CDK2 structures. 2. *J. Chem. Inf. Model.* **2008**, *48*, 669−678.

(20) Pipeline Pilot v8.0.1.500. http://accelrys.com/products/pipeline-pilot/ (accessed February 2014).

(21) Grant, J.; Gallardo, M.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.

(22) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shankin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(23) KNIME v2.6.4. http://www.knime.com (accessed February 2014).

(24) RDKit v2.1.0: Cheminformatics and Machine Learning Software. http://www.rdkit.org (accessed February 2014).

(25) Ashton, M.; Barnard, J.; Casset, F.; Charlton, M.; Downs, G.; Gorse, D.; Holliday, J.; Lahana, R.; Willett, P. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quant. Structure-Act. Relat.* **2002**, *21*, 598−604.

(26) The ChEMBL Database. https://www.ebi.ac.uk/chembldb/ (accessed February 2014).

(27) DUD: A Directory of Useful Decoys. http://dud.docking.org (accessed February 2014).

(28) Molecular Operating Enviroment 2012.10. http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm (accessed February 2014).

(29) Schrödinger. http://www.schrodinger.com (accessed February 2014).

(30) Hawkins, P. C.; Nicholls, A. Conformer generation with OMEGA: Learning from the data set and analysis of the failures. *J. Chem. Inf. Model.* **2012**, *50*, 572.

(31) Microsoft Excel 14.3.9 (Office 2011). http://office.microsoft.com/en-gb/excel/ (accessed February 2014).

(32) Vortex v2013.08.24252. http://www.dotmatics.com/products/vortex/ (accessed February 2014).

(33) RStudio v0.98.484. http://www.rstudio.com (accessed February 2014).

(34) Yabuuchi, H.; Niijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T.; Minowa, Y.; Tsujimoto, G.; Okuno, Y. Analysis of multiple compound−protein interactions reveals novel bioactive molecules. *Mol. Sys. Biol.* **2011**, *7*, 472.

(35) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.

(36) Golbraikh, A.; Muratove, E.; Fources, D.; Tropsha, A. Dataset modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 1−4.