Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201300143

*e*FindSite: Enhanced Fingerprint-Based Virtual Screening Against Predicted Ligand Binding Sites in Protein Models

Wei P. Feinstein[a] and Michal Brylinski*[a, b]

Abstract: A standard practice for lead identification in drug discovery is ligand virtual screening, which utilizes computing technologies to detect small compounds that likely bind to target proteins prior to experimental screens. A high accuracy is often achieved when the target protein has a resolved crystal structure; however, using protein models still renders significant challenges. Towards this goal, we recently developed *e*FindSite that predicts ligand binding sites using a collection of effective algorithms, including meta-threading, machine learning and reliable confidence estimation systems. Here, we incorporate fingerprint-based virtual screening capabilities in *e*FindSite in addition to its flagship role as a ligand binding pocket predictor. Virtual screening benchmarks using the enhanced Directory of Useful Decoys demonstrate that *e*FindSite significantly outperforms AutoDock Vina as assessed by several evaluation metrics. Importantly, this holds true regardless of the quality of target protein structures. As a first genome-wide application of *e*FindSite, we conduct large-scale virtual screening of the entire proteome of *Escherichia coli* with encouraging results. In the new approach to fingerprint-based virtual screening using remote protein homology, *e*FindSite demonstrates its compelling proficiency offering a high ranking accuracy and low susceptibility to target structure deformations. The enhanced version of *e*FindSite is freely available to the academic community at http://www.brylinski.org/efindsite.

Keywords: Ligand virtual screening · Fingerprint-based virtual screening · Molecular docking · Protein threading · Data fusion · Machine learning · Support vector machines

1 Introduction

Ligand virtual screening is a computational methodology for selecting small molecules (ligands) that bind to target proteins (receptors). Of particular interest in modern drug discovery, this technique is cost-effective in predicting potential hit compounds before undertaking experimental drug screening. Therefore, ligand virtual screening has become a standard practice in pharmaceutical industry as well as in drug related research.[1] Currently, one of the most commonly used techniques in computer-aided drug design is ligand virtual screening by molecular docking, which predicts physical interactions between receptor proteins and drug candidates at the atomic level.[2,3] This process requires two keys elements: an effective search algorithm and a reliable scoring function. In order to identify an optimal conformation of a ligand-protein complex, a robust searching algorithm is pivotal; here, the major challenge is to efficiently explore the protein-ligand conformational space, which can be potentially very large. Equally important is an accurate scoring function to evaluate binding affinities of docked compounds, so that bioactive molecules are assigned higher ranks than inactive ligands. Over the past years, a significant progress has been made and a number of molecular docking algorithms and tools have been developed. Docking methods, e.g. AutoDock,[4,5] DOCK,[6] FlexX,[7] GOLD,[8,9] Glide,[10] and Surflex-Dock,[11] employ their own searching schemes and scoring functions, thus present individual strengths and weaknesses. Studies demonstrating successful experimental validation of many of these tools have also been reported;[12–15] however, significant challenges exist.[16,17] For example, high-resolution protein structures are typically required for reliable virtual screening and ligand ranking, which hinders the application of ligand virtual screening in large-scale projects at the proteome level.

As one of scientific breakthroughs, genome sequencing of hundreds of organisms including human has been completed. A constantly increasing pace of sequencing leads to the exponential accumulation of genomic data. Benefiting from this unique scientific advancement, systems biology has emerged to accelerate studies of complex interactions at the proteome-level.[18] Clearly, systems-level approaches require a comprehensive knowledge of the entire reper-

[a] *W. P. Feinstein, M. Brylinski*
Department of Biological Sciences, Louisiana State University
Baton Rouge, LA 70803, USA
**e-mail: michal@brylinski.org*

[b] *M. Brylinski*
Center for Computation & Technology, Louisiana State University
Baton Rouge, LA 70803, USA

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201300143.

toire of gene products within a given proteome, preferably including structural information. However, for the human genome as an example, experimentally solved protein structures only account for less than a quarter of the entire proteome. This dire situation calls for alternative methods to fill in the gap. Providentially, fast advancements in computing technologies have empowered computational approaches to protein structure modeling facilitating a broad range of research activities. In protein structure prediction thus far, two primary approaches are widely used. Comparative modeling is suitable for generating structures for proteins closely homologous to those with known structures (templates), whereas threading/fold recognition utilizes weakly homologous templates found by mining the "twilit zone" of sequence similarity.[19] The latter is based on the observation that protein structure is more conserved than sequence.[20,21] For either approach, the basic concept is to identify a template or a set of templates from the Protein Database Bank,[22] which are subsequently used to generate a structure model of the target protein. Incorporating computationally generated models of gene products significantly expands the structural coverage of the human proteome and, consequently, improves proteome-wide functional inference and annotation.[23–25] Across-genome protein structure modeling should also benefit drug design by accelerating the discovery of leads for polypharmacology[26,27] as well as facilitating drug repositioning.[28,29] Despite continuous improvements in modeling techniques, the quality of protein models remains lower than experimentally solved structures. Thus, the key question is whether protein models can be reliably used in structure-based function annotation.[30] In drug development, the question becomes whether protein models of varying quality can be routinely utilized in ligand virtual screening without compromising its ranking capabilities, which is particularly important for proteome-wide applications.

In this spirit, we extended *e*FindSite, a recently developed approach for ligand binding site prediction,[31] to perform virtual screening. *e*FindSite is an evolution/structure-based method that employs a collection of advanced techniques including highly sensitive meta-threading and unsupervised as well as supervised machine learning algorithms. It is especially powerful in the prediction of ligand binding pockets in weakly homologous protein models. *e*FindSite exploits a tendency of proteins to preserve the locations of ligand binding sites in certain folds.[32] Consequently, regions possessing functionally important features tend to be evolutionarily conserved. These include ligand attributes as well; for instance, compounds binding to evolutionarily remotely related homologues contain strongly conserved anchor functional groups.[33] Based on these observations, *e*FindSite extracts binding ligands and their chemical properties from holo-templates detected by sequence profile-driven meta-threading[34] for use in fingerprint-based virtual screening. Note that despite implementing similar techniques to these widely used in ligand-based virtual screening, *e*FindSite is conceptually more similar to structure-based methods, viz. it requires only target protein structures, but no *a priori* knowledge on sets of binding molecules. Using target crystal structures as well as weakly homologous protein models, we evaluate the performance of *e*FindSite in virtual screening to identify small organic molecules that likely bind to the predicted binding pockets using KEGG Compound[35] and ZINC12[36] libraries. Furthermore, in large-scale benchmarks using the enhanced version of the Directory of Useful Decoys (DUD-E),[37] we compare *e*FindSite to AutoDock Vina,[5] which is one of the most widely used tools for structure-based virtual screening. We show that *e*FindSite maintains its high ligand ranking accuracy at a fairly constant level regardless of the structure quality of target proteins. Finally, as the example of a genome-wide application, we perform virtual screening against the entire proteome of *Escherichia coli* with encouraging results. Data collected through the work described in this study as well as stand-alone software distribution and online services for *e*FindSite are freely available to the academic community at http://www.brylinski.org/efindsite.

2 Materials and Methods

2.1 Holo-Template Library and PDB-Bench Dataset

*e*FindSite requires a template library of protein-ligand complexes, which was compiled using ligand-bound proteins from the Protein Small Molecule Database.[38] Template redundancy was removed using PISCES[39] and a threshold of 40% pairwise sequence identity. However, proteins that bind multiple ligands at different locations separated by at least 8 Å, were included even if their global sequence similarity is $>40\%$. With respect to ligand selection, we kept only small organic compounds composed of 6–100 heavy atoms non-covalently bound to template proteins. This filtering process produced a non-redundant and representative holo-template library composed of 15,285 proteins complexed with 20,215 ligands.

The first benchmarking dataset, referred to as PDB-bench, was compiled from the template library using three additional selection criteria. First, proteins 50–600 residues in length were identified. Second, we kept only those proteins, for which at least three weakly homologous and structurally related ligand-bound templates were identified using meta-threading. Here, weak homology is demarcated by a maximum sequence identity of 40%, whereas the structural relationship is measured by a TM-score[40] reported by Fr-TM-align[41] with a significance threshold set to 0.4. The last criterion considers only those proteins that bind either a single ligand or multiple ligands, but in approximately the same location according to the Protein Data Bank (PDB).[22] Applying these criteria yields a non-redundant dataset of 3,659 protein-ligand complexes, PDB-bench. In addition to the default target-template sequence identity threshold of 40%, we also benchmark binding

pocket prediction and ligand ranking against the PDB-bench dataset using only those templates whose sequence identity is below 30%, 20% and 10%.

2.2 Directory of Useful Decoys, Enhanced (DUD-E) Dataset

DUD-E is a database specifically designed to perform rigorous tests of docking algorithms, scoring functions and virtual screening tools.[37] Compared to the original DUD dataset,[42] DUD-E comprises a more diverse set of 102 proteins including ion channels and G-protein coupled receptors. The total number of experimentally validated active compounds in DUD-E is 22,886, which gives an average number of 224 ligands per protein target. Furthermore, sets of property-matching decoy molecules are significantly expanded to 50 per one active compound.

2.3 Benchmarking Protein Structures

For both datasets, PDB-bench and DUD-E, we compiled three sets of target protein structures. The first set comprises crystal structures obtained from the PDB.[22] In addition, we generated two sets of protein models of high and moderate quality, which are used to assess the sensitivity of ligand virtual screening to structural deformations in target protein structures. Weakly homologous protein models were constructed by *e*Thread, a recently developed method for template-based protein structure modeling.[34,43] *e*Thread employs structure assembly using either Modeller[44] or TASSER-Lite;[45] we used both protocols to generate up to 20 models excluding those templates that share $>40\%$ sequence identity with the target protein. Next, one model with a TM-score to native of >0.7 was randomly selected for the high-quality dataset. Similarly, another model from a TM-score range of 0.4–0.7 was randomly chosen and included in the moderate-quality dataset. When the model construction procedure did not produce structures of preferred quality for either dataset, the crystal structure was artificially distorted to a desired resolution using a simple Monte Carlo procedure.[46]

2.4 Virtual Screening Using eFindSite

The flowchart for virtual screening by *e*FindSite is presented in Figure 1. *e*FindSite utilizes holo-templates with known structures to identify binding pockets in target proteins.[31] For each predicted binding site, e.g. the one shown in Figure 1A, template-bound ligands are extracted and converted to a fingerprint representation. Molecular fingerprints are bit strings that represent the structural and chemical features of organic compounds.[47] Here, we employ two fingerprints commonly used in cheminformatics: 166-bit MACCS[48] and 1024-bit Daylight (http://www.daylight.com/dayhtml/doc/theory/). The calculation of Daylight fingerprints is conducted by OpenBabel[49] and MACCS fingerprints by MayaChemTools (http://www.mayachemtools.org/).

Next, using an average linkage clustering with the Tanimoto coefficient threshold of 0.7, template-bound compounds are clustered into *n* groups. This procedure results in two types of clusters: Daylight and MACCS, denoted in Figure 1B as C^D and C^M, respectively. Each cluster has a weight,

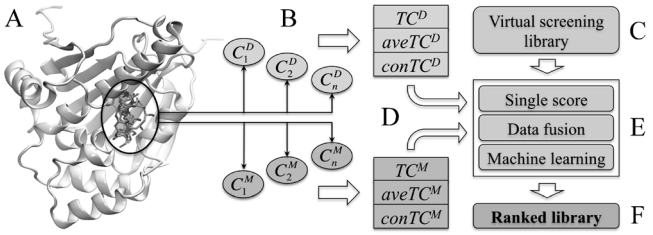

Figure 1. Flowchart of virtual screening using *e*FindSite. (A) *e*FindSite starts with the prediction of binding sites in the target structure and collects ligands bound to the template proteins at similar locations. (B) Template-bound ligands are partitioned into *n* clusters using two molecular fingerprints: Daylight (C^D) and MACCS (C^M). Compounds from a virtual screening library (C) are scored using template ligand clusters and three versions of the Tanimoto coefficient, *TC*, *aveTC* and *conTC*, calculated for Daylight (superscript *D*) and MACCS (superscript *M*) fingerprints (D). (E) A variety of scoring functions are available to rank the query compounds including single fingerprint-based scores, data fusion techniques as well as machine learning, leading to the final ranked library (F).

which corresponds to the fraction of compounds that belong to this cluster. In addition, from individual fingerprints of template-bound compounds, we calculate two fingerprint profiles: Daylight and MACCS. Molecular fingerprints are binary, i.e. each bit position is set either on or off, whereas in a fingerprint profile, it is replaced by a fraction of compounds that have this bit position set on.

In order to maximize compound ranking accuracy, we incorporate 3 different measures of fingerprint overlap between a query compound and template-bound molecules: traditional (*TC*), average (*aveTC*) and continuous (*conTC*) Tanimoto coefficient. These scores are calculated separately for Daylight and MACCS fingerprints, see Figure 1D. Tanimoto coefficient, *TC*, is one of the most popular measures to quantify the similarity of two sets of bits and it is traditionally defined as:[50]

$$TC = \frac{c}{a+b+c} \quad (1)$$

where *a* is the count of bits on in the 1st string but not in the 2nd string, *b* is the count of bits on in the 2nd string but not in the 1st string, and *c* is the count of the bits on in both strings. In addition, the overlap between two molecular fingerprints can be measured by the average Tanimoto coefficient, *aveTC*:[51]

$$aveTC = \frac{TC + TC'}{2} \quad (2)$$

where TC' is the Tanimoto coefficient calculated for bit positions set off rather than set on. Furthermore, we use a version of the Tanimoto coefficient for continuous variables:[52]

$$conTC = \frac{\sum x_{pi}x_{ci}}{\sum x_{pi}^2 + \sum x_{ci}^2 - \sum x_{pi}x_{ci}} \quad (3)$$

where x_{pi} is the i-th descriptor of a fingerprint profile and x_{ci} is the i-th descriptor of a query compound. Tanimoto coefficient for continuous variables, *conTC*, measures a consensus score between a query compound and all template ligands, which are represented by a fingerprint profile. Traditional, *TC*, and average Tanimoto coefficient, *aveTC*, scores are calculated using a weighted average over the template ligand clusters:

$$TC^D = \sum_j^n w_j TC_j \quad (4)$$

where the superscript D stands for Daylight fingerprints, n is the number of template ligand clusters, w_j is the weight of j-th cluster as defined above, and TC_j is the traditional Tanimoto coefficient between a query compound and a representative template ligand (cluster centroid) from j-th cluster. The remaining single scores, $aveTC^D$, $conTC^D$, TC^M, $aveTC^M$ and $conTC^M$, are calculated in a similar fashion.

Figure 1E lists scoring functions available for virtual screening using *e*FindSite. In addition to the 6 individual scoring functions: *TC*, *aveTC* and *conTC* calculated using 1024-bit Daylight and 166-bit MACCS molecular fingerprints, we developed 3 composite scoring functions using data fusion techniques, in which information on the same dataset is integrated for a more coherent representation.[53] Data fusion-based scoring functions combine 6 individual fingerprint scores and apply SUM, MIN and MAX rules. That is, library compounds are re-ranked by the sum of their individual scores, the minimal and the maximal values, respectively. Other than data fusion, we also designed a machine learning approach to ligand ranking using Support Vector Machines (SVM) for classification problems (SVC). Here, we use an SVC implementation from libSVM[54] and a feature vector for machine learning consisting of 6 individual scoring functions: TC^D, $aveTC^D$, $conTC^D$, TC^M, $aveTC^M$, and $conTC^M$. A two-class (binding/non-binding) SVC model is used to estimate the probability that a given ligand binds to the predicted pocket. The implemented machine learning model is cross-validated against the DUD-E dataset using a leave-one-out protocol. Specifically, one protein is removed from the dataset before constructing an SVC model and the performance of the model is evaluated by the excluded case; this procedure is repeated for the entire dataset.

2.5 C++ Implementation of Fingerprints

*e*FindSite stores molecular fingerprints using the class template bitset of fixed-size sequences of N bits, where N is 1024 for Daylight and 166 for MACCS fingerprints (C++ syntax is std::bitset<1024> and std::bitset<166>, respectively). This particular implementation allows a rapid comparison of two fingerprints using standard logic operators: XOR, AND, OR (C++ operators are ^=, &= and |=, respectively) and a public member function std::bitset::count. Using these operations, the traditional Tanimoto coefficient, *TC*, can be expressed as:

$$TC = \frac{OR - AND}{OR - AND + XOR} \quad (5)$$

The calculation of *aveTC* can be done in a similar fashion, additionally including a public member function std::bitset::flip to calculate the Tanimoto coefficient for bit positions set off rather than set on. We note that this algorithm eliminates expensive iterations through containers, which are required when using standard array-like implementations of fingerprints. Finally, each element of the class template bitset occupies only one bit, thus this design is also highly optimized for space allocation.

2.6 Confidence Index

Irrespective of the scoring function used, virtual screening confidence is assessed using a *Z*-score calculated for the top ranked compound. For instance, the *Z*-score when using TC^D is defined as:

$$Z\text{-score} = \frac{TC^D_{top} - \langle TC^D \rangle}{\sigma^{TC^D}} \quad (6)$$

where TC^D_{top} is the TC^D for the top-ranked compound, and $\langle TC^D \rangle$ and σ^{TCD} are the average TC^D and the standard deviation calculated over all library compounds. *Z*-score confidence estimates for $aveTC^D$, $conTC^D$, TC^M, $aveTC^M$ and $conTC^M$, are calculated in a similar fashion.

We also developed a machine learning-based confidence index for virtual screening using *e*FindSite and composite (data fusion) scoring functions. Specifically, we assign ligand ranking with either "low" or "high" confidence by an SVM classification model, which uses *Z*-score values calculated for the top-ranked compound by 6 individual fingerprint-based scoring functions. Similar to compound scoring using SVC, we also use a machine learning implementation from libSVM;[54] the model is cross-validated against the PDB-bench dataset using a leave-one-out protocol. A two-class ("low"/"high") classifier estimates a probability that

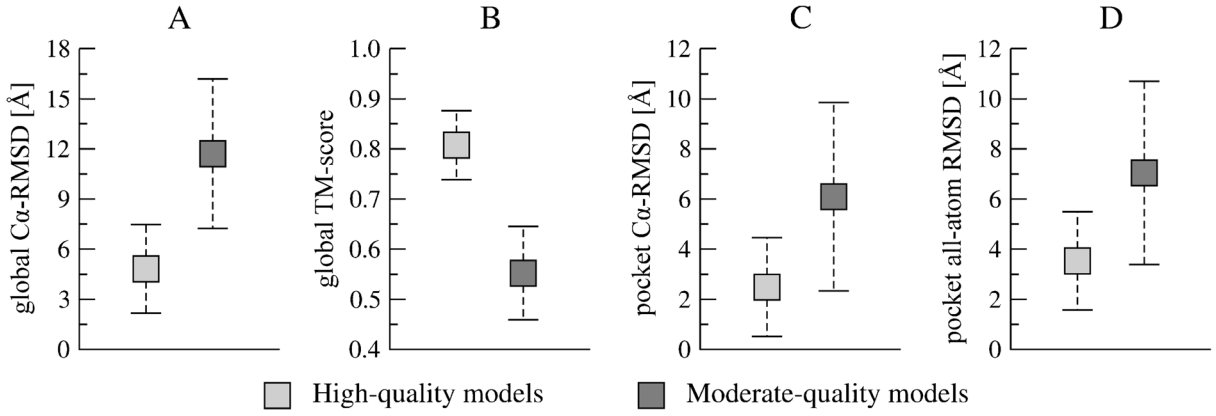

Figure 2. Structure quality of two datasets of protein models used in addition to crystal structures as targets for ligand virtual screening. Global (A) Cα-*RMSD* and (B) TM-score, (C) Cα-*RMSD* and (D) all-atom (non-hydrogen) *RMSD* of ligand binding sites.

the native ligand is ranked within the top 1% and 10% of the screening library.

2.7 Virtual Screening Using AutoDock Vina

The performance of *e*FindSite is compared to AutoDock Vina, version 1.1.2.[5] Target protein structures are converted to the required PDBQT format using MGL Tools, version 1.5.4.[55] The addition of polar hydrogens and partial charges as well as format conversion to PDBQT of ligand molecules is carried out using Open Babel, version 2.3.1.[56] In Vina, the default protocol is used with the docking box center set to the predicted pocket center reported by *e*FindSite.

2.8 Compound Libraries for Virtual Screening

*e*FindSite virtual screening can be used with the following screening libraries (number of compounds is given in parentheses): BindingDB[57] (338,662), DrugBank[58] (6,126), KEGG Compound[35] (11,265), KEGG Drug[35] (5,992), RCSB PDB[59] (12,879), NCI-Open[60] (239,870), ChEMBL[61] (248,344) and ZINC12[36] (244,659). Due to the large number of compounds in ChEMBL and ZINC12, we compiled their non-redundant subsets using the SUBSET program[62] and a pairwise Tanimoto coefficient threshold of 0.8.

2.9 Genome-Scale Ligand-Based Virtual Screening

For genome-scale virtual screening using *e*FindSite, we selected *Escherichia coli* K12 strain,[63] which is widely used in molecular biology and bioengineering. Structure models of 4,552 *E. coli* gene products 50–600 residues in length have been constructed using *e*Thread as described previously.[31] Briefly, 3D models were assembled using Modeller;[44] however, when an estimated TM-score was <0.5 indicating difficult modeling, TASSER-Lite[64] was used to construct additional models. In these cases, the final model of a target protein was selected based on a higher TM-score estimated by *e*Rank. Using structure models, ligand binding sites were predicted in gene products in *E. coli* proteome by *e*FindSite. In the present study, each putative binding pocket is further subject to ligand virtual screening against ZINC12 and KEGG Compound libraries in order to identify potential binding molecules.

3 Results and Discussion

3.1 Virtual Screening Against PDB-Bench Dataset

Initial virtual screening benchmarking calculations are carried out for PDB-bench proteins with an objective to identify native ligands within a non-redundant background library of 244,659 compounds from ZINC12.[36] In these benchmarks, we use three sets of target structures: crystal structures as well as high- and moderate-quality protein models. The characteristics of non-native, modeled structures are presented in Figure 2. Figure 2A (2B) shows that the average global Cα-*RMSD* from native (TM-score) for high- and moderate-quality models is 4.8 Å (0.81) and 11.7 Å (0.55), respectively. These values are also well correlated with the local structure quality of ligand binding sites, whose Cα (all-atom) *RMSD* is 2.5 Å (3.3 Å) and 6.1 Å (7.0 Å), respectively; see Figures 2C and 2D. Certainly, these deviations from experimental conformations pose a significant challenge for using protein models as targets in virtual screening.

In Table 1, the ranking accuracy of *e*FindSite is assessed by the median rank of native ligands normalized by the total number of compounds in the screening library. First, we evaluate 6 individual scoring functions based on 2 types of molecular fingerprints, Daylight and MACCS, and 3 versions of Tanimoto coefficient: *TC*, *aveTC* and *conTC*. In addition to the entire benchmarking dataset, we assess the results separately for the subset of targets for which binding sites are accurately predicted, i.e. Matthew's correlation coefficient (MCC) for binding residues is ≥ 0.5. Independently on the target structure quality, *aveTC* is the most ef-

Table 1. Median rank of the native ligand from the PDB-bench dataset expressed as the percentage of the screening library.

Dataset	Daylight fingerprints[a]			MACCS fingerprints[a]			Data fusion		
	TC^D	$aveTC^D$	$conTC^D$	TC^M	$aveTC^M$	$conTC^M$	SUM	MAX	MIN
Crystal structures	4.02%	1.46%	3.10%	7.51%	2.97%	5.06%	3.88%	1.11%	7.27%
Crystal structures *MCC* ≥ 0.5[b]	0.12%	0.10%	0.20%	0.32%	0.14%	0.23%	0.21%	0.04%	0.33%
High-quality models	4.04%	1.47%	3.12%	8.27%	3.10%	5.24%	4.33%	1.21%	7.86%
High-quality models *MCC* ≥ 0.5[b]	0.12%	0.10%	0.21%	0.32%	0.15%	0.23%	0.22%	0.04%	0.34%
Moderate-quality models	4.08%	1.44%	3.02%	7.47%	3.03%	5.18%	3.99%	1.20%	6.68%
Moderate-quality models *MCC* ≥ 0.5[b]	0.09%	0.09%	0.16%	0.26%	0.14%	0.19%	0.19%	0.03%	0.32%

[a] TC, aveTC and conTC is the traditional, average and continuous Tanimoto coefficient, respectively. [b] Only correctly predicted pockets for which MCC calculated over the binding residues is $\geq$0.5 are used.

fective individual scoring function; using Daylight and MACCS fingerprints, a native ligand is typically ranked within the top 1.46% and 2.97% of the screening library for all predicted pockets, respectively. Not surprisingly, when only accurately predicted pockets are considered, the ranking accuracy increases to 0.10% and 0.14%, respectively. Moreover, Daylight fingerprints are more accurate than MACCS in these benchmarking calculations. Further improvement is observed when data fusion is applied to combine compound ranks obtained by individual scoring functions. Depending on the quality of target structures, the native ligand is now ranked within the top 1.1–1.2% and 0.03–0.04% of the library for all and the subset of accurately predicted pockets, respectively. These results are in line with previous studies reporting the enhanced performance of binary similarity searching by data fusion techniques.[65,66] Importantly, this analysis also demonstrates that *e*FindSite to large extent tolerates distortions in target protein structures, thus it is applicable not only to crystal structures, but also to high- as well as moderate-quality models.

3.2 Effects of Protein Homology on Virtual Screening

Many novel protein targets may be evolutionarily only weakly related to structures currently available in the PDB. In that regard, we evaluate the impact of low protein homology on ligand binding site prediction and virtual screening using *e*FindSite. In Figure 3, in addition to the default sequence identity threshold of 40% used in this study, we predict binding sites and conduct virtual screening for the PDB-bench dataset using only those templates whose sequence identity is $\leq$30% and $\leq$20%. The accuracy of binding site prediction at 40% and 30% sequence identity thresholds is comparably high; for instance the percentage of proteins for which at least one pocket is detected is 98% and 95%, respectively. Moreover, binding sites are predicted within 8 Å (4 Å) from the geometric center of a native ligand for 71% (57%) and 67% (53%) of the targets, respectively. The performance of *e*FindSite starts deteriorating at very low sequence identity thresholds; excluding templates with >20% sequence identity to the target results in at least one binding site predicted and these predicted within 8 Å and 4 Å for 81%, 45% and 33% of the target proteins, respectively. We note that at the threshold of 10%, binding sites are detected for less than 1% of the targets, thus these results are not included in Figure 3.

Next, we calculate the fraction of targets, for which the native ligand is ranked within the top 1% and 10% of the ZINC12 screening library. Moreover, we consider only these targets, for which the binding site is predicted within a distance of 8 Å and 4 Å from the experimental pocket center; this is because virtual screening is unreliable for incorrectly predicted pockets as shown in Table 1. Under these conditions, the accuracy of virtual screening using *e*FindSite is fairly independent on protein homology. Figure 3 shows that at the sequence identity thresholds of 20–40%, for ~70% (~60%) and ~65% (~55%) of the targets, the native ligand is ranked within the top 10% (1%) when using pockets predicted within 4 Å and 8 Å, respectively. Thus, very remote protein homology (less than 20% sequence identity) affects the accuracy of pocket prediction; however, virtual screening is still successful when correct pockets are detected.

3.3 Virtual Screening Against DUD-E Dataset

In addition to the PDB-bench dataset, we evaluate the performance of *e*FindSite against the DUD-E dataset,[37] the enhanced version of the Directory of Useful Decoys[42] that is widely used in virtual screening benchmarking as a gold standard dataset. A key feature of these compound sets is that decoy molecules are carefully selected to match physicochemical properties of active compounds; however, they have different topologies and, consequently, bioactivity profiles. Similar to the PDB-bench, we compare virtual screening results using individual scoring functions as well as data fusion techniques. The diversity of target proteins and compound sets also allows for the construction and cross-validation of a non-linear, machine learning-based scoring function. We assess the performance by several

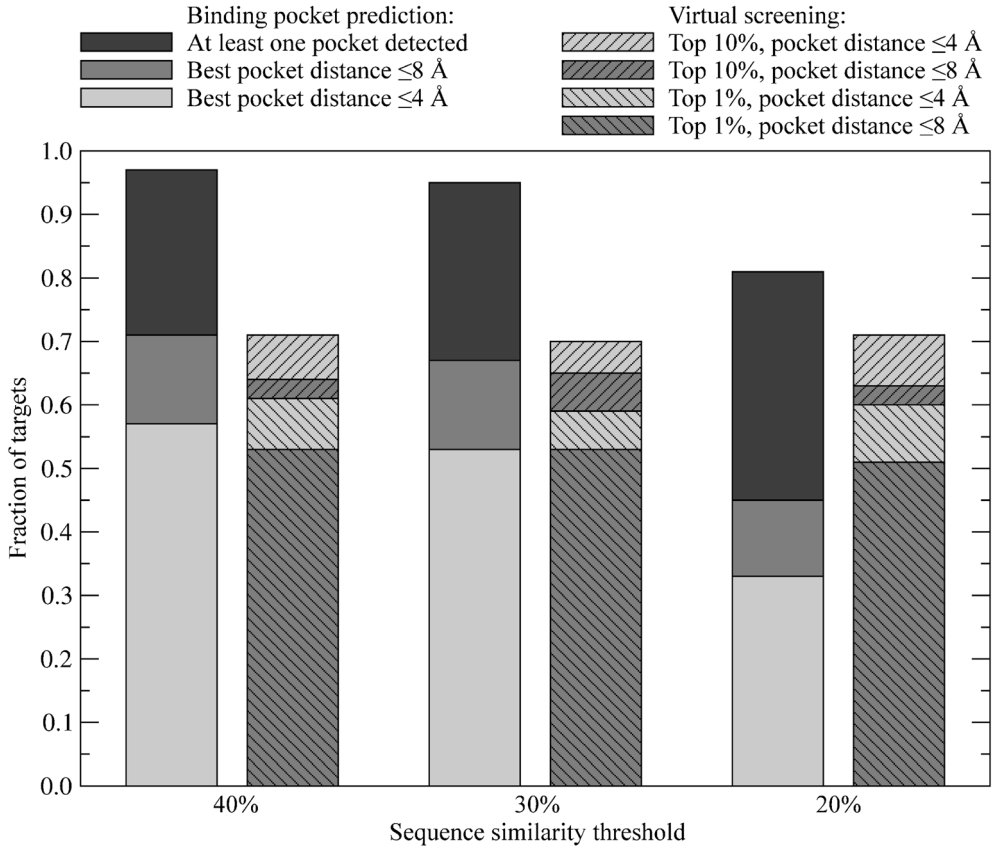

Figure 3. Accuracy of binding pocket prediction and virtual screening by *e*FindSite at different sequence identity thresholds for the crystal structures of PDB-bench proteins. Pocket prediction is assessed by the fraction of targets for which *e*FindSite detects at least one binding site, as well as these for which the best pocket is predicted within 8 Å and 4 Å from the geometric center of a bound ligand in the experimental structure. The accuracy of virtual screening is accessed by the fraction of pockets for which the native ligand is ranked within the top 1% and 10% of the ZINC12 screening library.

metrics widely used in cheminformatics: enrichment factor (EF) for the top 1% and 10% of the ranked library, Boltzmann-enhanced discrimination of receiver operating characteristics (BEDROC), area under the accumulation curve (AUAC) and ACT-50%. EF measures the enrichment of the top fraction of the ranked library with active compounds compared to that obtained purely by a random chance; larger EF indicates better ranking capabilities. BEDROC addresses the so-called "early recognition problem"; it was designed to assess the overall performance of an algorithm by assigning privileged weights to active compounds enriched in the top fraction of the ranked library.[67] We use BEDROC20 in our analysis, which means that 80% of final BEDROC scores are based on the first 8% of the ranked dataset. AUAC measures the distribution of active compounds over the whole screening library and ACT-50% corresponds to the top fraction of the ranked library that contains half of the active molecules.

First, we identify these DUD-E proteins, for which *e*FindSite predicted binding sites within a distance of $\leq$8 Å with MCC calculated for binding residues of $\geq$0.4. Figure 4 shows the distribution of distances between predicted and experimental pockets. Consistent with our previous results,[31] the performance of pocket prediction drops off with the decreasing quality of target structures from crystal structures to high- and moderate-quality protein models. Based on the accuracy of predicted binding sites, we selected from the DUD-E dataset 81 crystal structures, 68 high- and 57 moderate-quality models for virtual screening benchmarks.

Table 2 evaluates different scoring functions implemented in *e*FindSite on the DUD-E dataset. Depending on the quality of target structures, using Daylight and MACCS bit strings yields BEDROC20 values of 0.23–0.27 and 0.28–0.29, respectively, thus MACCS fingerprints are slightly more accurate here than Daylight fingerprints. Individual scoring functions are outperformed by combined ranking methods; for instance, machine learning using SVC gives BEDROC20 of 0.30–0.31. Data fusion, particularly using the SUM rule, is the most accurate with BEDROC20 values up to 0.33. As assessed by AUAC, data fusion yields scores of 0.72–0.76, which are higher than those calculated using individual Tanimoto-based scoring functions falling in the range of 0.69–0.75. Here, SVC machine learning is notably less accu-

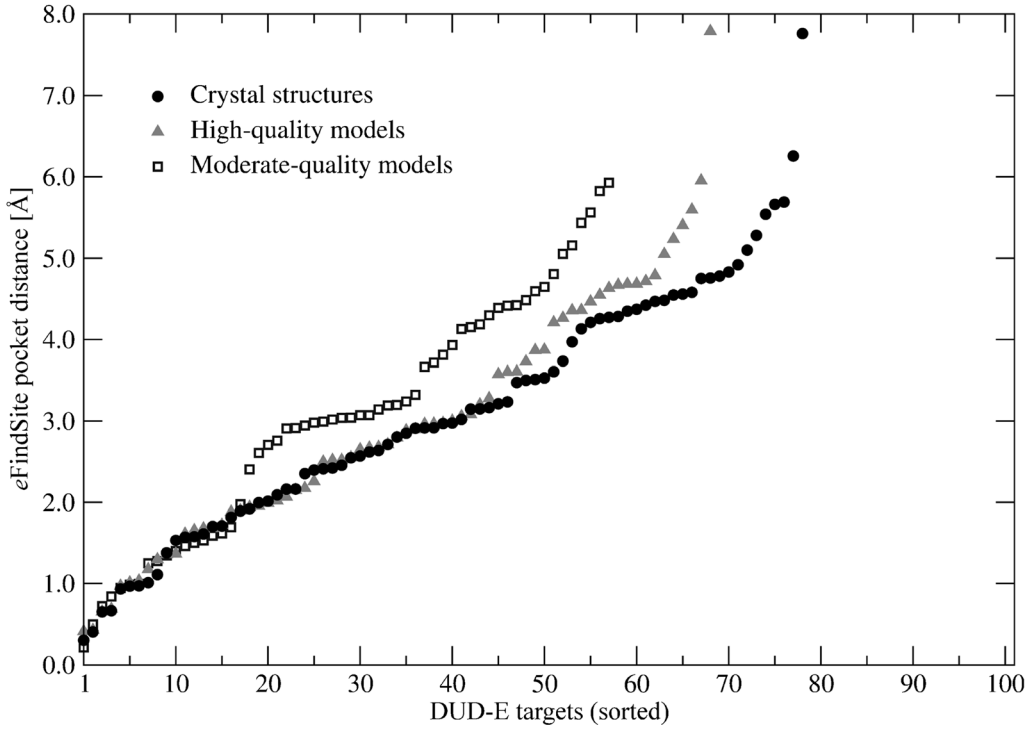

Figure 4. Distance between the center of the best binding pocket predicted by *e*FindSite and the geometric center of a native ligand for DUD-E proteins. Crystal structures, high- and moderate-quality models are sorted on the *x*-axis.

Table 2. Evaluation of different scoring functions for fingerprint-based virtual screening by *e*FindSite using crystal structures as well as different quality protein models constructed for the DUD-E dataset. Ranking accuracy is assessed by several evaluation metrics: EF, BEDROC20, AUAC and ACT-50%; reported values are averages over the dataset. Tested functions include 6 single fingerprint-based scores using 1024-bit Daylight and 166-bit MACCS bit strings, 3 data fusion techniques (SUM, MAX and MIN), and a machine learning-based approach (SVC).

Dataset	Metric	Daylight fingerprints[a]			MACCS fingerprints[a]			Data fusion			SVC[b]
		TC^D	$aveTC^D$	$conTC^D$	TC^M	$aveTC^M$	$conTC^M$	SUM	MAX	MIN	
Crystal structures	$EF^{1\%}$	9.29	9.44	9.53	11.62	11.62	11.63	9.36	8.47	9.16	12.95
	$EF^{10\%}$	3.41	3.75	3.69	3.85	3.85	3.92	4.03	4.04	3.72	3.32
	BEDROC20	0.24	0.26	0.26	0.28	0.28	0.28	0.32	0.32	0.30	0.31
	AUAC	0.70	0.71	0.71	0.73	0.73	0.74	0.75	0.74	0.72	0.61
	ACT-50%	0.26	0.24	0.24	0.22	0.22	0.21	0.21	0.21	0.23	0.34
High-quality models	$EF^{1\%}$	9.37	9.33	9.56	11.77	11.77	11.73	9.96	8.80	9.73	13.41
	$EF^{10\%}$	3.44	3.81	3.75	4.00	4.00	4.07	4.20	4.22	3.82	3.22
	BEDROC20	0.24	0.27	0.26	0.29	0.29	0.29	0.33	0.33	0.31	0.30
	AUAC	0.70	0.72	0.71	0.74	0.74	0.75	0.76	0.75	0.73	0.61
	ACT-50%	0.25	0.23	0.24	0.21	0.21	0.20	0.20	0.20	0.22	0.33
Moderate-quality models	$EF^{1\%}$	8.77	9.12	9.16	11.97	11.97	11.69	9.98	9.08	9.80	13.99
	$EF^{10\%}$	3.29	3.65	3.64	4.04	4.04	4.07	4.11	4.25	3.77	3.32
	BEDROC20	0.23	0.25	0.25	0.29	0.29	0.29	0.32	0.33	0.30	0.30
	AUAC	0.69	0.71	0.71	0.74	0.74	0.75	0.76	0.75	0.73	0.62
	ACT-50%	0.26	0.23	0.24	0.20	0.20	0.19	0.20	0.19	0.22	0.31

[a] *TC*, *aveTC* and *conTC* is the traditional, average and continuous Tanimoto coefficient, respectively. [b] Support Vector Machines for classification.

rate with AUAC of 0.61–0.62. A similar trend is observed using ACT-50% as the evaluation metric; smaller ACT-50% values in Table 2 indicate more sensitive scoring functions. Interestingly, SVC yields the highest $EF^{1\%}$, which corresponds to the percentage of active compounds detected within the top 1% of the ranked library; this shows that machine learning most effectively recognizes a small subset of bioactive molecules. Nevertheless, the overall per-

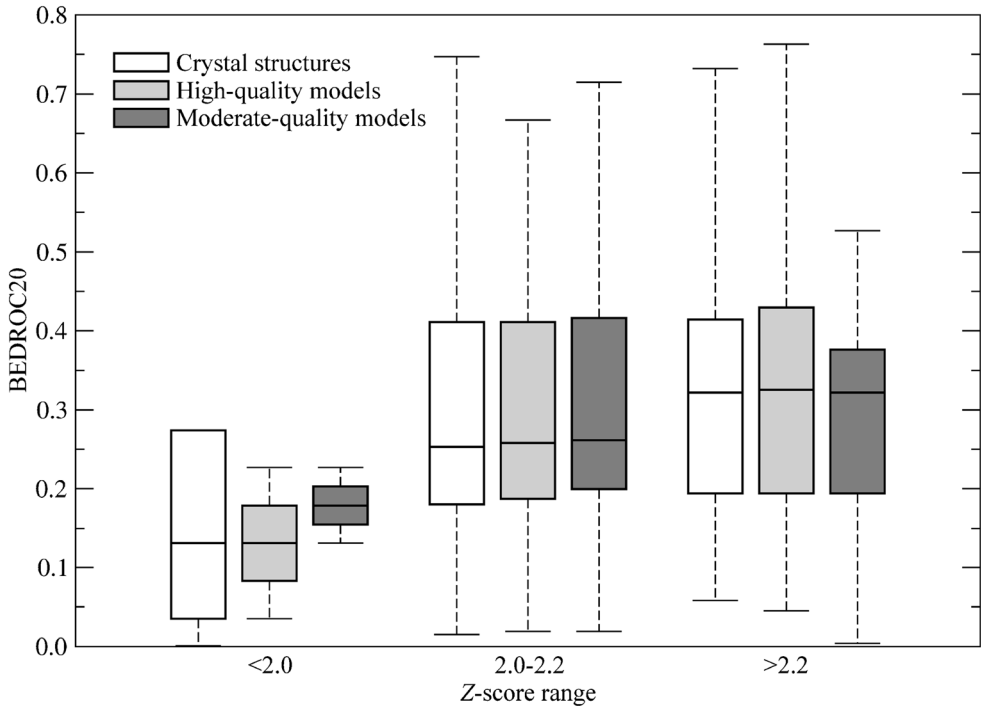

Figure 5. Confidence of virtual screening assessed by a *Z*-score of the top-ranked compound for the DUD-E dataset. For each set of target structures (crystal, high- and moderate-quality models), proteins are assigned to three groups based on the *Z*-score: <2.0, 2.0–2.2 and >2.2. The distribution of BEDROC20 scores within each group is shown as box-and-whisker graphs. Boxes end at the quartiles Q_1 and Q_3; a horizontal line in a box is the median. Whiskers point at the farthest points that are within 3/2 times the interquartile range.

formance of data fusion techniques, particularly using the SUM rule, is superior compared to other methods. This is consistent with previous studies on multiple search methods showing a systematic improvement of compound ranking by applying data fusion techniques.[68,69] Comparing results obtained for crystal structures to those for different quality protein models demonstrates a fairly high insensitivity of *e*FindSite to the structure deformations of target receptors. This is an important feature of our approach that addresses the detrimental impact of non-native receptor structures on virtual screening outcome.[17,70] On the whole, *e*FindSite implements sensitive scoring functions and exhibits a high tolerance to structural imperfections of target proteins, thus holds a significant promise for large-scale virtual screening applications.

3.4 Confidence Index System

A reliable confidence index for virtual screening is a useful feature that can help to identify these targets, for which ligand ranking is likely accurate. Here, we use a *Z*-score calculated for the top-ranked compound that measures its remoteness in standard deviation units from the average score obtained across the entire screening library. Using data collected for the DUD-E dataset, we show in Figure 5 that the *Z*-score is correlated with the accuracy of ligand ranking as measured by BEDROC20. For crystal target structures as well as high- and moderate-quality protein models, the median BEDROC20 is ~0.15 at a low Z-score of <2.0. *Z*-score values of 2.0–2.2 and >2.2 indicate more confident predictions, for which the median BEDROC20 scores are ~0.26 and ~0.32, respectively.

We also developed a machine learning-based approach for estimating the confidence of virtual screening using *e*FindSite. It employs *Z*-score values obtained for six individual fingerprint-based scoring functions to assign ligand ranking with either a "low" or "high" confidence. This classifier is cross-validated on the PDB-bench dataset; its accuracy in detecting these predictions, in which the native ligand is ranked within the top 1% and 10% of the screening library, is 0.56 and 0.75, respectively. Although not perfect, these confidence estimation systems may provide valuable information on the reliability of virtual screening in practical applications.

3.5 Potential for Identifying Novel Compounds

A weak point of ligand-based virtual screening is its relatively lower potential for discovering novel compounds compared to e.g. structure-based virtual screening by molecular docking. In a traditional ligand-based approach, library compounds are ranked based on their chemical similarity to already known binders. In *e*FindSite, small organic molecules extracted from evolutionarily related protein-

ligand complexes are used instead of known binders. Modeling techniques such as fingerprint profiling and clustering are designed to improve the sensitivity of detecting more diverse molecules that are not simply variants of already known compounds. The potential for identifying novel molecules emerges from the ability to rank them early in an ordered list using molecular fingerprints constructed from those compounds that are at most chemically weakly related. We analyze the potential of *e*FindSite for identifying "novel" compounds using a simulated DUD-E dataset. In this experiment, we benchmark the scoring engine of *e*FindSite using active molecules associated with a given target protein instead of the template-bound ligands. This strategy allows us to precisely control the amount of chemical information used to perform virtual screening. Specifically, for each active molecule, we exclude those compounds that have chemical similarity above some threshold and construct fingerprints from the remaining ligands; this procedure is repeated for all active molecules. Thus, query compounds are ranked within a screening library using these molecules that are to some extent chemically dissimilar.

Figure 6 shows the results obtained for the simulated DUD-E dataset using *e*FindSite and data fusion with the SUM rule. Using a chemical similarity threshold represented by the Tanimoto coefficient progressively decreasing from 0.8 to 0.2, ranking accuracy is assessed by $EF^{1\%}$, BEDROC20 and AUAC (Figures 6A, 6B and 6C, respectively). Allowing chemically similar compounds at a high Tanimoto coefficient threshold of 0.8 to be included as ligand templates yields the median $EF^{1\%}$, BEDROC20 and AUAC of 37.0, 0.61 and 0.88, respectively. In general, *e*FindSite maintains its high ranking capability even when the Tanimoto coefficient drops to 0.4; here, the median $EF^{1\%}$, BEDROC20 and AUAC are 15.1, 0.36 and 0.83, respectively. We note that this accuracy is slightly above that reported in Table 2, where ligand templates extracted from evolutionarily remotely related proteins are used. Altogether, these results suggest that the performance of *e*FindSite in virtual screening is fairly high even when template ligands are chemically weakly related, thus it holds a significant promise for identifying novel compounds.

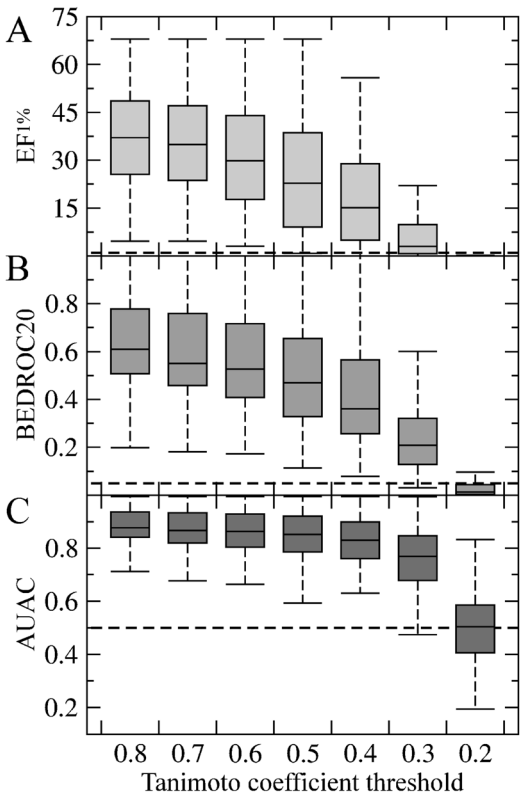

Figure 6. Performance of virtual screening using data fusion on the simulated DUD-E dataset. The results are assessed by (A) enrichment factor for the top 1% of the ranked library, (B) BEDROC20, and (C) AUAC, excluding those compounds whose Tanimoto coefficient to the query compound is above the threshold shown on the *x*-axis. Boxes end at the quartiles Q_1 and Q_3; a horizontal line in a box is the median. Whiskers point at the farthest points that are within 3/2 times the interquartile range. For each metric, a horizontal dashed line represents the accuracy of random ligand ranking.

3.6 Comparison with AutoDock Vina

For any new methodology it is obligatory to analyze its performance with respect to widely used state-of-the-art algorithms. In that regard, we compare *e*FindSite to AutoDock Vina[5] in representative virtual screening benchmarks against the DUD-E dataset.[37] Table 3 reports the results assessed by EF, BEDROC20, AUAC and ACT-50%. Depending on the evaluation criteria, *e*FindSite using data fusion outperforms Vina for crystal structures; for example the average $EF^{1\%}$/BEDROC20/AUAC is 9.36/0.32/0.75 and 6.17/0.28/0.68, respectively. The performance difference is clearly more dramatic for weakly homologous protein models; here, $EF^{1\%}$, $EF^{10\%}$ and BEDROC20 show a significant, two-fold drop-off in ranking accuracy by Vina, whereas the performance of *e*FindSite remains at a constant level. The performance of *e*FindSite for both high- and moderate-quality models seems to be slightly better than that for crystal structures; a similar observation also applies to Vina, for which moderate-quality models give better performance than high-quality models. This can be explained using Figure 4, which shows that in benchmarking calculations against the DUD-E dataset, we use 81, 68 and 57 crystal structures, high- and moderate-quality protein models, respectively. Pocket prediction accuracy for these additional crystal structures and high-quality models is on average lower, which in turn decreases the performance of ligand ranking as we demonstrate in Table 1. Therefore virtual screening against moderate-quality models using both *e*FindSite and Vina starts with fewer lower quality binding pockets, yielding a slightly better performance.

Table 3. Performance comparison between *e*FindSite and AutoDock Vina using crystal structures as well as different quality protein models constructed for the DUD-E dataset. Ranking accuracy is assessed by several evaluation metrics: EF, BEDROC20, AUAC and ACT-50% for confidently predicted pockets only ($MCC \geq 0.5$). Reported values are averages over the dataset. For *e*FindSite, data fusion with the SUM rule is used.

Dataset	Metric	*e*FindSite	AutoDock Vina
Crystal structures	$EF^{1\%}$	9.36	6.17
	$EF^{10\%}$	4.03	3.11
	BEDROC20	0.318	0.283
	AUAC	0.747	0.681
	ACT-50%	0.212	0.261
High-quality models	$EF^{1\%}$	9.96	2.45
	$EF^{10\%}$	4.20	1.82
	BEDROC20	0.333	0.128
	AUAC	0.758	0.593
	ACT-50%	0.200	0.377
Moderate-quality models	$EF^{1\%}$	9.98	2.86
	$EF^{10\%}$	4.11	1.95
	BEDROC20	0.322	0.135
	AUAC	0.756	0.595
	ACT-50%	0.196	0.380

The same results are analyzed further by breaking down the dataset into individual proteins in Figure 7 with the corresponding numerical data included as Supplementary Tables 1–3. Light green areas in Figure 7 highlight targets, for which *e*FindSite outperforms Vina. It is apparent that *e*FindSite is more accurate for the majority of cases regardless of evaluation metric. Furthermore, in those cases for which Vina performs better than *e*FindSite, the ranking is primarily based on the crystal structures of target proteins (red circles). Significantly fewer high- (blue squares) and moderate-quality (yellow triangles) are located within green areas. For instance, considering BEDROC20 (AUAC) scores (Figures 7C and 7D), *e*FindSite yields better ranking than Vina for 49% (63%), 82% (88%) and 80% (80%) of the target receptors when crystal structures, high- and moderate-quality models are used, respectively. Table 3 and Figure 7 clearly demonstrate that particularly for modeled protein structures, the improvement of *e*FindSite over Vina is not only quantitative with better average scores, but also qualitative, i.e. reliable ligand ranking is obtained for notably more targets.

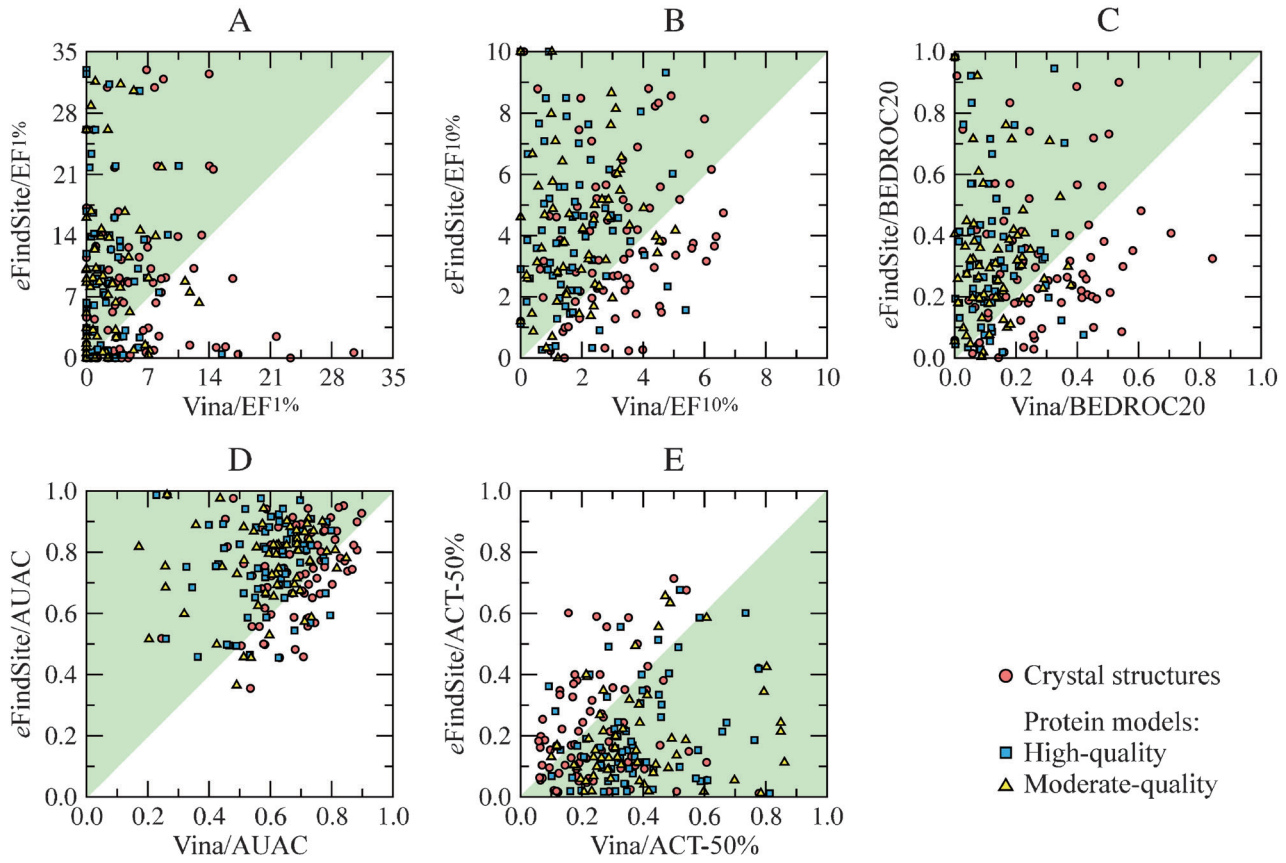

Figure 7. Performance comparison between *e*FindSite and AutoDock Vina in virtual screening against the DUD-E dataset. Compound ranking accuracy is assessed by: (A) $EF^{1\%}$, (B) $EF^{10\%}$, (C) BEDROC20, (D) AUAC, and (E) ACT-50% for target crystal structures (red circles) as well as high- (blue squares) and moderate-quality (yellow triangles) protein models. Light green areas highlight the improved performance of *e*FindSite over Vina.

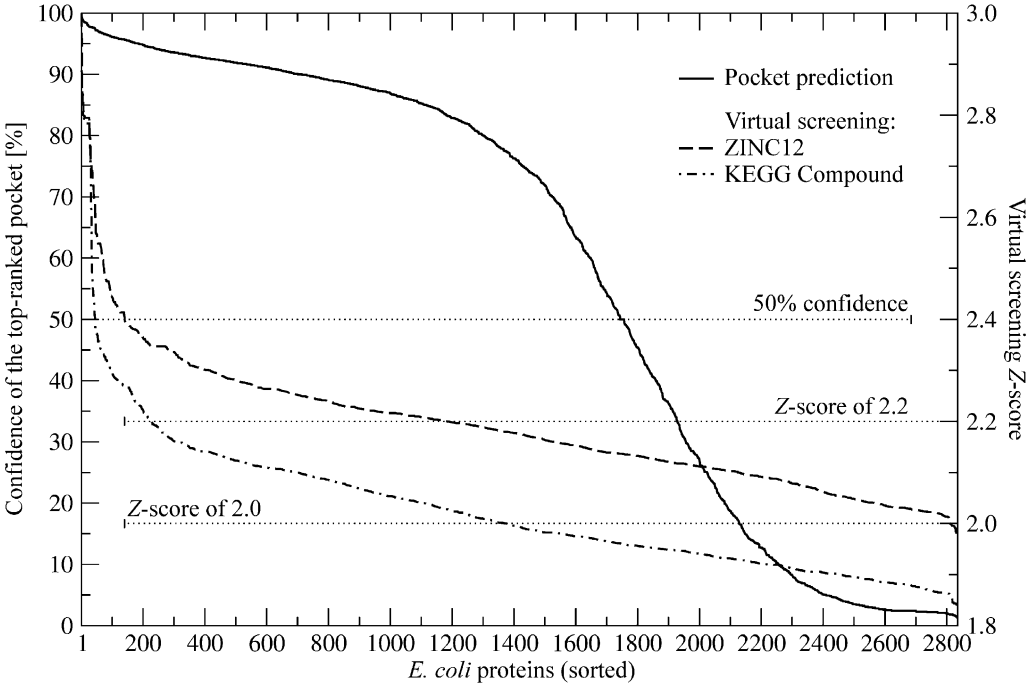

Figure 8. Confidence of binding pocket prediction (left ordinate) and ligand virtual screening (right ordinate) for *E. coli* proteome. Ligand ranking confidence is expressed as a *Z*-score for the top-ranked compound selected from two screening libraries: ZINC12 and KEGG Compound. Horizontal dotted lines delineate 50% confidence for pocket prediction, and a virtual screening *Z*-score of 2.0 and 2.2.

3.7 Proteome-Wide Virtual Screening for *E. coli*

Encouraging results obtained in comprehensive benchmarks motivated us to apply *e*FindSite in across-genome virtual screening. Specifically, we conduct large-scale virtual screening for the entire proteome of *Escherichia coli*. First, using *e*Thread,[34] we constructed protein models for 4,552 gene products; 85% of these models have an estimated TM-score of $\geq$0.4, thus provide reliable targets for further ligand binding annotation.[31] Next, we predicted ligand-binding pockets using *e*FindSite.[31] At least one ligand binding pocket is predicted for 2,828 gene products, which comprise 62% of *E. coli* proteome. Figure 8 shows that approximately 63% of the top-ranked binding pockets are assigned a confidence of $>$50%; we note that confidence estimates provided by *e*FindSite correlate well with the actual prediction accuracy.[31] Finally, each putative binding pocket is subject to virtual screening against KEGG Compound and ZINC12 libraries using enhanced *e*FindSite and data fusion with the SUM rule, which provides the most reliable compound ranking for modeled protein structures. The former library contains 11,265 small organic molecules known to bind to proteins,[35] whereas the latter comprises 244,659 mostly synthetic compounds for drug development and design.[36]

The reliability of virtual screening can be evaluated by a *Z*-score of the top-ranked compound since *Z*-score values correlate with ligand ranking accuracy; higher scores typically indicate a higher accuracy of virtual screening using *e*FindSite (see Figure 5). Figure 8 shows that the top-ranked compound selected by *e*FindSite from KEGG Compound and ZINC12 libraries has a *Z*-score of $\geq$2.2 for 7.9% and 41.7% of binding sites in *E. coli*, respectively. The top-ranked compound is within the *Z*-score range of 2.0–2.2 for additional 40.6% and 57.5% of binding sites, respectively. Furthermore, we also estimate the ranking accuracy using a machine learning classifier calibrated on the PDB-bench dataset. We expect that virtual screening against the KEGG Compound and ZINC12 libraries ranks the native compound in the top 1% for 2,446 and 2,810 binding sites, accounting for 86% and 99% of all putative pockets identified in *E. coli* proteome, respectively. Thus for the majority of gene products in *E. coli*, not only binding site locations, but also binding ligands can be confidently predicted.

3.8 A Case Study for Proteome-Wide Virtual Screening

To conclude this study, we discuss a representative example that demonstrates the potential of enhanced *e*FindSite for proteome-wide ligand virtual screening. We selected a 241aa *E. coli* protein, LptB (Ensembl ID: EBESCP00000218125), whose experimental structure is not available. Moreover, it represents a non-trivial case, since the highest sequence identity to a protein in PDB (branched chain amino acid ABC transporter from *Thermotoga maritime*, PDB-ID: 1ji0) is only 33%. However, exploring remote homology using *e*Thread, a confident structure model for this target is constructed with an estimated TM-

score of 0.82; see gray cartoon model in Figure 9. *e*FindSite identified 6 putative ligand binding sites in the modeled structure; the top-ranked pocket highlighted in Figure 9, involving residues V18, P37, N38, G39, A40, G41, K42, T43

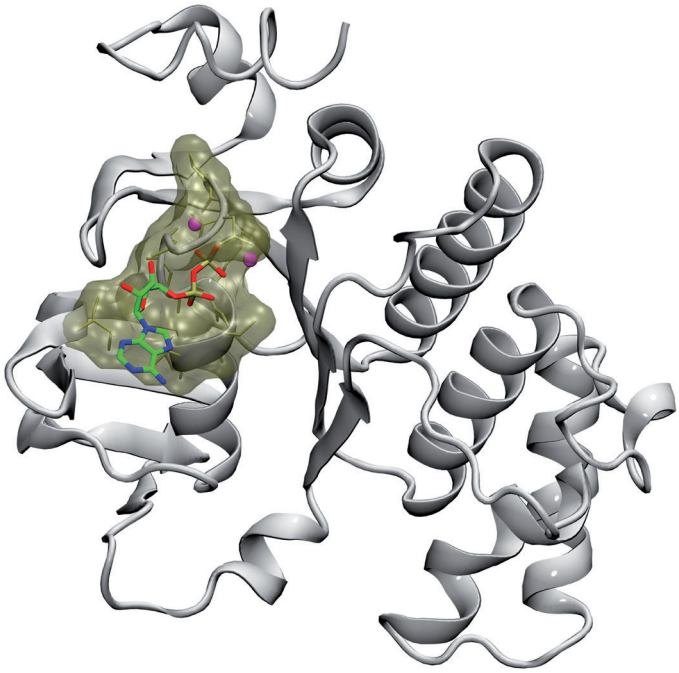

Figure 9. Structure model (gray cartoon) constructed for LptB gene from *E. coli*. Binding pocket residues predicted by *e*FindSite are shown as golden sticks and a transparent surface. A putative binding ligand identified by virtual screening, adenosine-5′-diphosphate (sticks colored by atom type) as well as two magnesium ions (pink balls) are transferred from a template protein, SufC from *Thermus thermophilus* (PDB-ID: 2d2f) upon its global superposition on LptB.

and T44, is assigned a high, 87.9% confidence. Most templates that share this consensus binding site belong to the ABC-ATPase family of ATP-binding cassette transporters. These proteins are responsible for the translocation of various molecules across membranes, where the ATPase component provides energy for the cross-membrane movement.[71,72] If this hypothesis holds true, the target protein is expected to bind ATP-like nucleotides. Indeed, top five compounds picked up by virtual screening using *e*FindSite against the KEGG Compound library are 3′-keto-3′-deoxy-ATP, 3′-keto-3′-deoxy-AMP, deoxyadenosine 5′-triphosphate, 2′-deoxyadenosine 5′-diphosphate and 2′-deoxyadenosine 5′-phosphate (KEGG ID: C07024, C07025, C00131, C00206 and C00360, respectively). Strikingly similar, the top-ranked compounds in the ZINC library, ZINC06585262, ZINC01235954, ZINC01579998, ZINC05004678 and ZINC16939847 are adenosine 1-oxide, *N*-benzoyladenosine, 2-amino-8-[(2*S*,3*S*,4*S*,5*S*)-3,4-dihydroxy-5-methylol-tetrahydrofuran-2-yl]imidazo[1,2-a][1,3,5]triazin-4-yl, 2-(2-amino-6,8-dichloro-purin-9-yl)-5-(hydroxymethyl)tetrahydrofuran-3,4-diol and 9-[(2*S*,3*R*,4*R*,5*S*)-3,4-dihydroxy-5-(hydroxymethyl)tetrahydrofuran-2-yl]purine-6-carboxamide, respectively. All these top-ranked compounds are ATP/ADP/AMP-related nucleotides suggesting that the predicted binding site in LptB indeed binds ATP-like molecules. These results support our earlier prediction that the target protein belongs to the family of ABC-ATPase.

Available experimental data provides evidence that the *E. coli* essential gene LptB is directly involved in lipopolysaccharide transport across the periplasm.[73] It was suggested that LptB, described therein as a soluble protein possessing the ATP binding fold but not transmembrane domain, could provide the energy from ATP hydrolysis to extract lipopolysaccharides from the periplasmic surface of the inner membrane and deliver it to the LptD/LptE complex in the outer membrane.[74] Our modeling results not only support these experimental findings, but also shed light on molecular structure of LptB and its putative interactions with small molecules. Most importantly, *e*FindSite screening can identify promising lead compounds providing a good starting point for the structure-based development of pharmaceuticals and, in the case of LptB, possibly new antibiotic agents.

3.9 Computational Efficiency

Virtual screening calculations typically involve processing large datasets of query compounds, thus computational efficiency is essential. Most algorithms implement molecular fingerprints as sequential containers that encapsulate either fixed or dynamic size arrays. For instance, widely used OpenBabel employs vectors of unsigned integers to store fingerprint data.[56] In contrast, *e*FindSite implements a bitset container of fixed-size sequences of bits. Bitsets can be manipulated by standard logic operators (XOR, AND, OR), which significantly improves computational efficiency. This is shown in Figure 10, which compares the performance of an implementation using vectors of unsigned integers to that of fixed-size bitsets in virtual screening of 1×10^6 library compounds. The performance of both algorithms decreases with the increasing number of template ligand clusters due to the larger number of individual Tanimoto coefficient calculations; see Equation 4. For instance, the throughput of vectors of unsigned integers and bitsets at 10 template ligand clusters is ~12 k and 23 k query compounds per second, respectively. Consequently, the higher performance of bitset implementation significantly shortens the total simulation time, which is shown as an inset plot in Figure 10. In addition, bitsets are much more memory efficient. For example, storing a screening library of 1×10^6 compounds as both Daylight and MACCS fingerprints requires 2.38 GB of RAM using vectors of unsigned integers vs. 0.15 GB for bitsets, thus using bitsets requires 16× less bits than integers to store the fingerprint data.

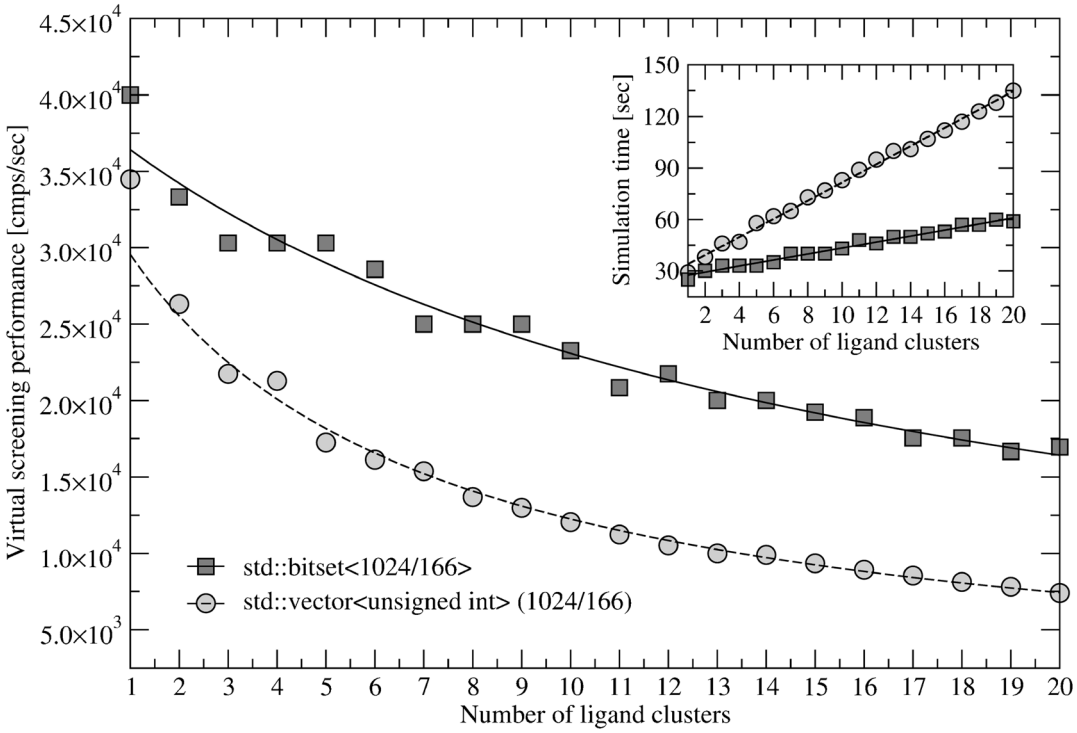

Figure 10. Performance of fingerprint-based virtual screening by *e*FindSite using different C++ data structures. Computational throughput is assessed by the number of compounds per second as a function of the number of template ligand clusters. Inset: throughput is replaced by the total time required to virtually screen a library of 1×10^6 compounds.

4 Conclusions

High-throughput screening is widely used in drug discovery; however, it frequently can be lengthy and expensive. In contrast, virtual screening utilizes computing techniques to process a large dataset of chemical compounds in a relatively short time and at low costs. Thus, it typically precedes experimental screens limiting compound libraries to those compounds that have the highest chance to exhibit a desired activity. As such, it has become a standard practice in pharmaceutical industry for lead compound identification. Nevertheless, for structure-based approaches to virtual screening, the quality of target protein structures is still a salient issue. Experimentally solved structures are unavailable for many important drug targets, which necessitates using protein models. Because of major developments in genome sequencing technologies, the latter can be routinely generated for the majority of gene products in numerous organisms. This presents appealing opportunities for conducting across-proteome virtual screening, which can be used in the lead development for polypharmacology or in systems level applications such as drug repositioning. Despite the continuous progress in improving the prediction reliability and compound ranking accuracy to meet the challenges of modern pharmacology, limitations exist, thus the development of new and more effective virtual screening methods is required.

In this spirit, we extended *e*FindSite, a recently developed evolution/structure-based ligand binding site predictor, to perform ligand virtual screening as well. *e*FindSite implements accurate scoring functions, machine learning and data fusion techniques to predict binding ligands with a high accuracy and offers a reliable system for the confidence estimation. Compared to widely used AutoDock Vina in comprehensive benchmarks, *e*FindSite provides improved compound ranking, as assessed by a variety of evaluation metrics. Importantly, this high performance is achieved not only for target crystal structures, but also for weakly homologous protein models whose structure quality can vary. We also show that it is effective when using only weakly related protein templates selected from the "twilight zone" of sequence similarity, as well as holds a promise for identifying "novel" compounds. Finally, we demonstrate the potential of *e*FindSite for proteome-wide applications and identify putative binding molecules for the majority of gene products in *E. coli* proteome. Because of its high tolerance to structural distortions in receptor proteins, *e*FindSite should provide a useful approach to virtual screening when only target protein sequences are available.

The enhanced version of *e*FindSite is freely available to academic community as a user-friendly web-server and a well-documented standalone software distribution at http://www.brylinski.org/efindsite; this website also pro-

vides all benchmarking results reported in this paper. Furthermore, the results of large-scale virtual screening for *E. coli* proteome are freely available at http://www.brylinski.org/content/databases.

Acknowledgements

This study was supported by the *Louisiana Board of Regents* through the Board of Regents Support Fund [Contract LEQSF(2012-15)-RD-A-05] and *Oak Ridge Associated Universities* (*ORAU*) through the 2012 Ralph E. Powe Junior Faculty Enhancement Award. Portions of this research were conducted with high performance computational resources provided by *Louisiana State University* (http://www.hpc.lsu.edu) and the *Louisiana Optical Network Institute* (*LONI*, http://www.loni.org).

References

[1] E. Bielska, X. Lucas, A. Czerwoniec, J. Kasprzak, K. Kaminska, J. Bujnicki, *BioTechnologia, J. Biotechnol, Comput. Biol. Bionanotechnol.* **2011**, *92*, 249–264.
[2] H. Chen, P. D. Lyne, F. Giordanetto, T. Lovell, J. Li, *J. Chem. Inf. Model.* **2005**, *46*, 401–415.
[3] K. Onodera, K. Satou, H. Hirota, *J. Chem. Inf. Model.* **2007**, *47*, 1609–1618.
[4] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comput. Chem.* **1998**, *19*, 1639–1662.
[5] O. Trott, A. Olson, *J. Comput. Chem.* **2010**, *31*, 455–461.
[6] T. J. Ewing, S. Makino, A. G. Skillman, I. D. Kuntz, *J. Comput. Aided Mol. Des.* **2001**, *15*, 411–428.
[7] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470–489.
[8] M. L. Verdonk, G. Chessari, J. C. Cole, M. J. Hartshorn, C. W. Murray, J. W. M. Nissink, R. D. Taylor, R. Taylor, *J. Med. Chem.* **2005**, *48*, 6504–6515.
[9] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, R. D. Taylor, *Proteins* **2003**, *52*, 609–623.
[10] 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, *J. Med. Chem.* **2004**, *47*, 1739–1749.
[11] A. Jain, *J. Comput. Aided Mol. Des.* **2007**, *21*, 281–306.
[12] W. Xu, G. Chen, W. Zhu, Z. Zuo, *Bioorg. Med. Chem. Lett.* **2010**, *20*, 5763–5766.
[13] P. A. Holt, P. Ragazzon, L. Strekowski, J. B. Chaires, J. O. Trent, *Nucleic Acids Res.* **2009**, *37*, 1280–1287.
[14] J. Liu, D. Dyer, J. Wang, S. Wang, X. Du, B. Xu, H. Zhang, X. Wang, W. Hu, *PLoS ONE*, **2013**, *8*, e64984.
[15] X. Lucas, S. Simon, R. Schubert, S. Gunther, *PLoS ONE* **2013**, *8*, e60679.
[16] P. Ferrara, H. Gohlke, D. J. Price, G. Klebe, C. L. Brooks, *J. Med. Chem.* **2004**, *47*, 3032–3047.
[17] S. McGovern, B. Shoichet, *J. Med. Chem.* **2003**, *46*, 2895–2907.
[18] L. Hood, J. R. Heath, M. E. Phelps, B. Lin, *Science* **2004**, *306*, 640–643.
[19] B. Rost, *Protein Eng.* **1999**, *12*, 85–94.
[20] A. Andreeva, D. Howorth, J.-M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, A. G. Murzin, *Nucleic Acids Res.* **2008**, *36*, D419–D425.
[21] L. H. Greene, T. E. Lewis, S. Addou, A. Cuff, T. Dallman, M. Dibley, O. Redfern, F. Pearl, R. Nambudiry, A. Reid, I. Sillitoe, C. Yeats, J. M. Thornton, C. A. Orengo, *Nucleic Acids Res.* **2007**, *35*, D291–D297.
[22] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.
[23] C. A. Wilson, J. Kreychman, M. Gerstein, *J. Mol. Biol.* **2000**, *297*, 233–249.
[24] A. Stark, R. B. Russell, *Nucleic Acids Res.* **2003**, *31*, 3341–3344.
[25] G. J. Bartlett, C. T. Porter, N. Borkakoti, J. M. Thornton, *J. Mol. Biol.* **2002**, *324*, 105–121.
[26] A. Carrieri, V. I. Perez-Nueno, G. Lentini, D. W. Ritchie, *Curr. Top Med. Chem.* **2013**, *13*, 1069–1697.
[27] U. Koch, M. Hamacher, P. Nussbaumer, *Biochim. Biophys. Acta* **2013**, *1844*, 156–161.
[28] L. Xie, P. E. Bourne, *Curr. Opin. Struct. Biol.* **2011**, *21*, 189–199.
[29] A. Schrattenholz, V. Soskic, *Curr. Med. Chem.* **2008**, *15*, 1520–1528.
[30] Y. Zhang, *Curr. Opin. Struct. Biol.* **2009**, *19*, 145–155.
[31] M. Brylinski, W. Feinstein, *J. Comput. Aided Mol. Des.* **2013**, *27*, 551–567.
[32] R. B. Russell, P. D. Sasieni, M. J. E. Sternberg, *J. Mol. Biol.* **1998**, *282*, 903–918.
[33] M. Brylinski, J. Skolnick, *PLoS Comput. Biol.* **2009**, *5*, e1000405.
[34] M. Brylinski, D. Lingam, *PLoS ONE* **2012**, *7*, e50200.
[35] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, *Nucleic Acids Res.* **1999**, *27*, 29–34.
[36] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
[37] M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet, *J. Med. Chem.* **2012**, *55*, 6582–6594.
[38] I. Wallach, R. Lilien, *Bioinformatics* **2009**, *25*, 615–620.
[39] G. Wang, R. L. Dunbrack, Jr., *Bioinformatics* **2003**, *19*, 1589–1591.
[40] Y. Zhang, J. Skolnick, *Proteins* **2004**, *57*, 702–710.
[41] S. B. Pandit, J. Skolnick, *BMC Bioinform.* **2008**, *9*, 531.
[42] N. Huang, B. K. Shoichet, J. J. Irwin, *J. Med. Chem.* **2006**, *49*, 6789–6801.
[43] M. Brylinski, W. P. Feinstein, *J. Comput. Sci. Syst. Biol.* **2012**, *6*, 001–010.
[44] A. Sali, T. L. Blundell, *J. Mol. Biol.* **1993**, *234*, 779–815.
[45] S. Pandit, J. Skolnick, *BMC Bioinform.* **2008**, *9*, 531.
[46] E. Bindewald, J. Skolnick, *J. Comput. Chem.* **2005**, *26*, 374–383.
[47] N. Nikolova, J. Jaworska, *QSAR Comb. Sci.* **2003**, *22*, 1006–1026.
[48] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
[49] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, E. L. Willighagen, *J. Chem. Inf. Model.* **2006**, *46*, 991–998.
[50] T. T. Tanimoto, in *IBM Internal Report*, **1958**.
[51] L. Xue, J. W. Godden, F. L. Stahura, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151–1157.
[52] P. Willett, *J. Chem. Inf. Model.* **1998**, *38*, 983–996.
[53] L. Klein, *SPIE Press* **2004**, *PM138SC*.
[54] C. C. Chang, C. J. Lin, *ACM Transact. Intell. Syst. Technol.* **2011**, *2*, 27.
[55] M. F. Sanner, *J. Mol. Graph. Model.* **1999**, *17*, 57–61.

[56] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, E. L. Willighagen, *J. Chem. Inf. Model.* **2006**, *46*, 991 – 998.
[57] X. Chen, M. Liu, M. K. Gilson, *Comb. Chem. High Throughput Screen.* **2001**, *4*, 719 – 725.
[58] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, *Nucleic Acids Res.* **2006**, *34*, D668 – D672.
[59] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235 – 42.
[60] G. W. Milne, M. C. Nicklaus, J. S. Driscoll, S. Wang, D. Zaharevitz, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219 – 24.
[61] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100 – D1107.
[62] J. H. Voigt, B. Bienfait, S. Wang, M. C. Nicklaus, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702 – 712.
[63] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, Y. Shao, *Science* **1997**, *277*, 1453 – 1462.
[64] S. B. Pandit, Y. Zhang, J. Skolnick, *Biophys. J.* **2006**, *91*, 4180 – 4190.
[65] N. Salim, J. Holliday, P. Willett, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435 – 442.
[66] M. Whittle, V. J. Gillet, P. Willett, J. Loesel, *J. Chem. Inf. Model.* **2006**, *46*, 2206 – 2219.
[67] J. Truchon, C. Bayly, *J. Chem. Inf. Model.* **2007**, *47*, 488 – 508.
[68] J. D. Holliday, E. Kanoulas, N. Malim, P. Willett, *J. Cheminform.* **2011**, *3*, 29.
[69] F. Svensson, A. Karlen, C. Skold, *J. Chem. Inf. Model.* **2012**, *52*, 225 – 232.
[70] M. L. Verdonk, P. N. Mortenson, R. J. Hall, M. J. Hartshorn, C. W. Murray, *J. Chem. Inf. Model.* **2008**, *48*, 2214 – 2225.
[71] P. M. Jones, A. M. George, *Cell Mol. Life. Sci.* **2004**, *61*, 682 – 699.
[72] E. Schneider, S. Hunke, *FEMS Microbiol. Rev.* **1998**, *22*, 1 – 20.
[73] P. Sperandeo, R. Cescutti, R. Villa, C. Di Benedetto, D. Candia, G. Deho, A. Polissi, *J Bacteriol.* **2007**, *189*, 244 – 253.
[74] P. Sperandeo, F. K. Lau, A. Carpentieri, C. De Castro, A. Molinaro, G. Deho, T. J. Silhavy, A. Polissi, *J. Bacteriol.* **2008**, *190*, 4460 – 4469.

Received: September 13, 2013
Accepted: December 6, 2013
Published online: February 12, 2014

