Rescoring of docking poses under Occam's Razor: are there simpler solutions?

Michael Zhenin, Malkeet Singh Bahia, Gilles Marcou, Alexandre Varnek, Hanoch Senderowitz, Dragos Horvath

Journal of Computer-Aided Molecular Design

Rescoring of Docking Poses under Occam's Razor - Are there Simpler Solutions?

--Manuscript Draft--

Manuscript Number:	
Full Title:	Rescoring of Docking Poses under Occam's Razor - Are there Simpler Solutions?
Article Type:	Original Research Article
Keywords:	Force Field Calculations, Docking, Scoring
Corresponding Author:	Dragos Horvath, Ph.D. CNRS Strasbourg, FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	CNRS
Corresponding Author's Secondary Institution:	
First Author:	Michael Zhenin
First Author Secondary Information:	
Order of Authors:	Michael Zhenin
	Malkeet Singh Bahia, Ph.D.
	Gilles Marcou, Ph.D.
	Alexandre Varnek, Ph.D.
	Hanoch Senderowitz, Ph.D.
	Dragos Horvath, Ph.D.
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Ligand affinity prediction from docking simulations is usually performed by means of highly empirical and diverse protocols. These protocols often involve the re-scoring of poses generated by a Force Field (FF) based Hamiltonian to provide either estimated binding affinities - or alternatively, some empirical goodness score. Re-scoring is performed by so-called scoring functions - typically, a reweighted sum of FF terms augmented by additional terms (e.g., desolvation/entropic penalty, hydrophobicity, aromatic interactions etc.). Sometimes, the scoring function actually drives ligand positioning, but often it only operates on the best scoring poses ranked top by the initial ligand positioning tool. In either of these rather intricate scenarios, scoring functions are docking-specific models, and most require machine-learning-based calibration. Therefore, docking simulations are less straightforward when compared to "standard" molecular simulations in which the FF Hamiltonian defines the energy, and affinity emerges as an ensemble average property over pools of representative conformers (i.e. the trajectory). According to Occam's Razor principle, additional model complexity is only acceptable if demonstrated to bring a significant improvement of prediction quality. In this work we therefore examined whether the complexity inherent to scoring functions is indeed justified. For this purpose we compared S4MPLE (Sampler for Multiple Protein-Ligand Entities), a general purpose conformation sampler based on the AMBER/GAFF FF, complemented with continuum solvation terms with several state of the art docking tools that rely on calibrated scoring functions (Glide, Gold, Autodock-Vina) in terms of its ability to top-rank the actives from large and diverse ligand series associated with various proteins. There is no clear winner of this study, where each program performed well on most of the targets, but also failed with respect to at least one of them. Therefore, a well-parameterized force field with a simple, energy-based ligand ranking protocol appears to be as effective docking protocol as intricate rescoring strategies

	based on scoring functions. Such a tool that can sample the conformational space of the free ligand, the bound ligand and the protein binding site using the same force field can alleviate many of the approximations common to contemporary docking protocols and allow e.g., for docking into highly flexible active sites when current scoring functions are not well suited to estimate receptor strain energies.

Rescoring of Docking Poses under Occam's Razor – Are there Simpler Solutions?

Michael Zhenin[§], Malkeet Singh Bahia[§], Gilles Marcou[‡], Alexandre Varnek[‡], Hanoch Senderowitz[§] & Dragos Horvath[‡]*

‡ Université de Strasbourg, 1 rue B. Pascal, Strasbourg 67000, France,

§ Bar Ilan University, Ramat-Gan, 5290002, Israel

* Corresponding author: dhorvath@unistra.fr

1 Abstract

Ligand affinity prediction from docking simulations is usually performed by means of highly empirical and diverse protocols. These protocols often involve the re-scoring of poses generated by a Force Field (FF) based Hamiltonian to provide either estimated binding affinities – or alternatively, some empirical goodness score. Re-scoring is performed by so-called scoring functions - typically, a reweighted sum of FF terms augmented by additional terms (e.g., desolvation/entropic penalty, hydrophobicity, aromatic interactions etc.). Sometimes, the scoring function actually drives ligand positioning, but often it only operates on the best scoring poses ranked top by the initial ligand positioning tool. In either of these rather intricate scenarios, scoring functions are docking-specific models, and most require machine-learning-based calibration. Therefore, docking simulations are less straightforward when compared to "standard" molecular simulations in which the FF Hamiltonian defines the energy, and affinity emerges as an ensemble average property over pools of representative conformers (i.e. the trajectory).

According to Occam's Razor principle, additional model complexity is only acceptable if demonstrated to bring a significant improvement of prediction quality. In this work we therefore examined whether the complexity inherent to scoring functions is indeed justified. For this purpose we compared S4MPLE (Sampler for Multiple Protein-Ligand Entities), a general purpose conformation sampler based on the AMBER/GAFF FF, complemented with continuum solvation terms with several state of the art docking tools that rely on calibrated scoring functions (Glide, Gold, Autodock-Vina) in terms of its ability to top-rank the actives from large and diverse ligand series associated with various proteins. There is no clear winner of this study, where each program performed well on most of the targets, but also failed with respect to at least one of them. Therefore, a well-parameterized force field with a simple, energy-based ligand ranking protocol appears to be as effective docking protocol as intricate rescoring strategies based on scoring functions. Such a tool that can sample the conformational space of the free ligand, the bound ligand and the protein binding site using the same force field can alleviate many of the approximations common to contemporary docking protocols and allow e.g., for docking into highly flexible active sites when current scoring functions are not well suited to estimate receptor strain energies.

2 Introduction

Scoring functions[1-8] are nowadays a key component of virtually every *in Silico* docking protocol, being used to evaluate the "pertinence" of the various ligand poses that are typically obtained from Force Field (FF)-based Molecular Mechanics (MM) calculations[9-13]. In general, scores are functions of a given geometry (pose) of the ligand-site complex, their arguments being either the internal coordinates (interatomic distances) in the complex or, more typically, empirical terms that are directly calculable from the said geometry (such as the number of established hydrogen bonds, the buried surface area or individual force-field-based energy terms).

Traditionally[1], the taxonomy of scoring functions distinguishes between "force field-based", "knowledge-based" and "empirical" scoring functions – a rather vague and arguable classification. In principle, "FF-based" functions are, like the FF-based Hamiltonian of MM calculations, a *weighed* sum of the various terms (electrostatic, van der Waals, torsional, covalent, *etc*) provided by the employed FF. In standard force fields, the weights are by default set to one (at least formally – because in practice, the choice of an effective dielectric constant is just a means to weigh down the Coulomb term, *etc*.). By contrast, in scoring functions the weights are being fitted by multilinear regression, in order to have the scoring function output match experimental free energy/affinity values for training examples of binding site-ligand complexes. "Empirical" scoring functions follow exactly the same principle, except that they might include additional terms, not present in the initial FF (for example, entropy penalties, estimated by counting the number of rotatable bonds in the ligand which are assumed to be restricted upon binding the protein or terms related to buried surface area). These terms are generically merely "molecular descriptors" of the complex. In addition, terms may include potentials of mean force of an implicit solvent model (implicit desolvation, hydrophobic contact intensity estimators), which are not default FF terms – but might be added to the FF engine. However, FF energy components lack a rigorous physical meaning and are simply rather complex "descriptors" of the site-ligand complex. Thus, we do not see any real differences between "FF-based" and "empirical" scoring functions. In fact, *all* scoring functions are inherently empirical, raising some questions about the appropriateness of the often-seen expression "empirical scoring function". Like elsewhere in the field of Quantitative Structure-Activity Relationships, QSAR[14-17], there is, in principle, complete freedom in matters of the choice of the functional form and machine learning protocols. Therefore, it is of little practical interest to formally distinguish[1] between "classical" linear regression-based scoring functions *versus* non-linear "machine-learned" approaches: multilinear regression too is formally a machine learning technique, even though the simplest one.

Knowledge-based[7, 8, 18, 19] scoring functions are, however, inherently different, in as far as they imply no direct fitting of affinity values as the explained variable, nor do they require any knowledge of effective affinities for the "training" set of complexes. These functions are based on statistical analysis of relative occurrence rates of pairwise contacts in experimentally solved site-ligand complexes. The central working hypothesis behind the method is that the more energetically favorable the contact between two atoms of given types is, the more often it will occur in the experimentally solved structures (compared to some "baseline" probability of those atom types to touch "by chance" – a rather ill-defined concept, which is the weak point of the theory). Observed occurrence rates are thus converted to mean free energy contributions per contact. This is a methodologically distinct approach from the above-mentioned "fitted" scoring functions and suffers from specific drawbacks – such as its intrinsic inability to learn that geometries with bad contacts are not stable. In experimental crystal structures, there are no examples of bad contacts (else, the binding mode would not have been observed and used for training). Therefore, knowledge-based scores are typically provided with an additionally fitted repulsive van der Waals-like term. All this notwithstanding, the analogy to naïve Bayesian learning[20] from observed occurrence rates clearly suggests that the approach is nevertheless yet another state-of-the-art machine learning technique. Concludingly, we wish to emphasize that scoring functions are quintessentially QSAR models based on predicted or experimental site-ligand geometries, covering various QSAR model building strategies, and inheriting all the strength and weaknesses of QSAR models.

We would like to emphasize at this point that it is not always easy to draw a clear separation line between "force field engines" and "scoring functions" based on FF terms. The fact that "original" FF terms are reweighed in scoring functions is by no means a real difference – after all, the original FF terms are derived through empirical parameterization (which is conceptually equivalent to reweighting). Thus, the only clear distinction we can see between FF engines and scoring functions is their *usage:* the former are expected to work for a large range of sampling problems, while the latter are typically restricted to the *a posteriori* estimation of pose quality. However, the choice between these two formalisms should eventually only be based on the number of tasks (e.g., docking, conformational sampling, protein folding) in which each method excels The question we wish to address here is whether both these formalisms are needed as distinct entities, or whether a unified all-purpose "force field" (or alternatively "scoring"?) engine might do the work all alone. The fact that many classical docking programs (including some used here) actually use the same "scoring" function

for both posing and docking is already an important step towards "unification". Unfortunately, these scoring functions are not likely applicable to classes of molecular simulations other than docking.

Docking pose rescoring, while practically wide spread, is theoretically questionable, as the "pertinence" of the poses should be granted by the algorithm that is proposing them. Poses should represent local minima at best, or stochastically generated geometries residing in low-energy areas of the ligand-site interaction energy surface, at least. One may argue that the scoring function should *not* be the one to pilot the docking, because its role is that of a "predictor" of the *free* energy of binding. However, free energy is an ensemble property and as such is associated with an entire conformational space zone. Therefore, it is not trivial whether – and, if so, how – it could be predicted based on a single geometry. Taking the ensemble average over many poses to estimate the binding free energy, as prone by fundamental statistical physics, is not a widespread approach. The idea has been tried – all while knowing that the limited set of docking poses cannot match the theoretical Boltzmann ensemble expected for rigorous docking[21]. However, it did not make it into "mainstream" docking programs – with the notable exception of MedusaDock[22].

Hence, most of the current docking approaches implicitly assume that a free energy score can be obtained from a single pose, and fitting is needed to compensate for all systematic errors committed by focusing on a single geometry instead of a conformational space zone (and by neglecting many aspects of the "docking event"; see below). This leads to an essential question: how to unambiguously define the geometry to be used as "the" representative of its conformational space neighborhood? State-of-the-art docking often uses docked ligand poses – hence, local minima of the FF-driven Hamiltonian. These correspond to perfectly arbitrary points on the scoring function landscape, as local minima of the latter do not coincide with the ones of the docking energy landscape. Docking programs are in general[23] quite successful in retrieving ligand poses close (typically, within an RMSD<2 Å) to the "native" poses as published in the Protein Data Bank. However, having found one or a few pose(s) at RMSD<2 Å is by no means sufficient to ensure that the binding affinity is correctly predicted. The conformational space zone delimited by the RMSD<2 Å criterion is not a smooth region around the global energy minimum, but an extremely rugged landscape hosting a plethora of local energy minima. Ideally, a scoring function alleged to predict the property of an entire conformational space zone based on one of its representative geometries should return nearly-constant values for all geometries in that zone – thereby downscaling the impact of the actual choice of the representative geometry. In practice, there will be discrepancies within the scores assigned to the actual "successful" poses found by a (typically stochastic) sampling process if a scoring function different from the

docking function is used. This is expected even though the scoring function landscape is smoothed out by down-weighing the rugged non-bonded energy terms (typically, its weight in scoring functions is of the order of 0.1).

Therefore using the actual scoring function to drive the poses *i.e.*, let it serve as the objective function being minimized at the pose sampling step is likely to improve the reproducibility of finding poses corresponding to favorable scores. This is current practice in a few docking programs, such as MedusaDock[24] or Gold[25, 26], albeit in the latter the choice of using a same objective function for docking and scoring or not is left to the user (and adopted in the present work). The most widely spread strategy is however a "hybrid" one, with a classical (or on-purposed simplified) FF engine (for example, in Schrödinger's Glide[27-29]), or another empirical approach such as ligand overlaying atop of a co-crystallized binder (in OpenEye's FRED[30-32]) being used to rapidly generate many preliminary poses. These poses are then evaluated by the scoring function, and the pose(s) optimizing the scoring function value being returned as final docking pose(s). In this sense, it can be argued that such docking protocols are indeed using the scoring function to actually 'guide' the docking. However, such scoring function optimization is restricted to the problem space zones already representing (near) optimal docking solutions according to the objective function used for preliminary posing. For example, optimization of the ChemGauss4 scoring function in FRED is restricted to picking its locally best value over 729 poses generated by small-step rigid-body roto-translations of the initial pose representing the overlay atop of a co-crystallized reference ligand.

It is somewhat difficult to assess where exactly state-of-the-art docking tools are positioned in the range between the two extreme paradigms: (a) complete sampling of the scoring function landscape over the entire docking problem space, *versus* (b) rescoring of poses generated by approaches that are different from the scoring function. The protocol for selecting "the" representative pose returning the "correct" (i.e., reflecting the free energy) score is often the result of an interplay between FF Hamiltonian/other posing generating protocols and scoring function optimization. This suggests that scoring functions may be rather docking protocol specific – not necessarily transferable. Moreover, stochastic docking approaches (which happen to be the most widely used), do not guarantee a complete reproducibility of all the potentially important details of the representative geometry to score. In addition, most of the scoring functions ignore the energetic and entropic costs of the conformational adaptation of the ligand to the site. Sometimes[33], the ligand strain energy is evaluated, and subtracted from the total energy, on the basis of the conformers sampled in the presence of the

site. This however is a poor approximation of the ligand strain component if in absence of the site a ligand adopts other conformations.

The role of empirical scoring functions is to compensate for all the systematic errors of the docking protocols, by means of an additional, machine-learned model, exploiting the output of the primary docking calculations as molecular descriptors. However, treating scoring as a machine learning problem comes with a price tag in the form of a limited Applicability Domain[34, 35] (AD). Development of interaction fingerprint[36-38] monitoring helped to focus virtual screening on compounds featuring already known ligand-site interaction patterns – the AD within which scoring functions are most likely to return accurate predictions. This helps to improve prediction accuracy, while sacrificing the unique theoretical ability of a docking tool to discover completely novel binding paradigms – so far unknown binding pharmacophores carried by novel molecular scaffolds.

All these problems prompted us to explore the performances of the most simple and unambiguous approach – using the "default" FF-based Hamiltonian as implemented for example in S4MPLE (including a continuum desolvation term and hydrophobic/hydrogen bonding contact bonuses based on differentiable contact fingerprints) for both docking and scoring. If this approach would lead to high-quality Receiver Operating Characteristic (ROC) curves, similar to those obtained from scoring function, it would benefit from the advantage of simplicity and avoid all the previously outlined pitfalls and methodological incongruence. A "parameter-free" score should be understood as free of fitable terms that are specific for scoring.

The straightforward candidate for calculating ROC curves would be the binding energy difference ΔE, defined in equation (1) as the difference between the lowest energy level of the most stable ligand-site geometry, minus the lowest energy level of the most stable unbound ligand geometry, minus – with flexible docking – the lowest energy level of the most stable empty "*apo*" active site geometry.

$$\Delta E = min_i\langle E_i\rangle^{ligand@site} - min_i\langle E_i\rangle^{ligand} - min_i\langle E_i\rangle^{site} \qquad (1)$$

Ensembles $\langle i \rangle$ of the bound states "ligand@site", free ligand and *apo* site geometries are to be generated by any arbitrary conformational sampling procedure, expected to converge (*i.e.* reproducibly rediscover the same lowest-energy geometries when initiated, *e.g.*, from different starting points). Since it is known that solvent effects in general and, in particular, hydrophobic interactions, of entropic nature[39, 40], are of utmost importance in ligand binding, the FF engine should

likely require the inclusion of a continuum solvation potential of mean force accounting for these key contributions. Even with this provision, a binding energy score accounting for solvent-related entropic effects is not yet a binding *free* energy – but the two values might display a sufficient degree of correlation, especially for targets in which enthalpy-entropy compensation[41, 42] comes into play. If this is the case, then a docking protocol with no need for rescoring may prove successful in virtual screening. It solely requires (a) a consistent FF engine, accurately estimating intra- and intermolecular energies (solvent effects included), and (b) a conformational sampling protocol, able to reproducibly discover the relevant minima of the above-mentioned energy landscape. Such a docking protocol is thus perfectly compliant with the requirement of maximal simplicity, as prone by Occam's razor principle – but can it compete with scoring function-endowed approaches?

The conformational sampling program S4MPLE[43-45] is well-suited for the candidate role of such a "rescoring-free" docking tool. First, being designed as a general tool for arbitrary conformational sampling problem, it can enumerate conformers for both free ligands, active protein sites – at user-defined degree of flexibility, going from automated readjustment of rotatable directional hydrogen bonds to flexible residue side chains, to flexible protein loops – and protein-ligand complexes. Second, it implements both implicit desolvation and hydrophobic contact terms, as additional terms to its AMBER[46]/Generalized Amber[47] (GAFF) FF engine, as mentioned in the preceding paragraph. Note that addition of the latter terms, in order to obtain the final "solvent-aware" FF engine, referred as "FitFF" in the original publication[43] required the fitting of associated empirical parameters, the objective function being the classical "redocking success" (RMSD of pose with respect to ligand geometry in the experimental PDB structure). ΔE as defined in equation (1) can be straightforwardly obtained with S4MPLE. This is based on the same energy and PMF term enabling S4MPLE to fold small peptides such as the Trp cage (1L2Y), or to address "covalent" docking applied to fragment growing protocols – tasks that are out of the applicability domain of "classical" scoring functions.

In order to compare the performance of "rescoring-free" S4MPLE to state-of-art docking tools, specific, non-trivial docking challenges were selected. Seven very different biological targets (GPCRs, kinases, other enzymes) of known 3D structure, and for which large sets of putative ligands are available (containing both binders and non-binders), were selected and standardized. The smallest set features 747 compounds, and the largest 6843. For the three most data-rich of the seven targets, compounds were randomly split into two equal-size sets, for deployment on different machines. Part of the resulting eleven sets featured both real-life experimentally validated actives and inactives as reported in the ChEMBL[48] database, while the others were sets of actives *versus* artificial decoys,

from the enhanced Directory of Useful Decoys, DUD-E[49]. S4MPLE (using FitFF as previously reported[43], with no additional tuning of its parameters) was employed to both dock and sample the free candidate ligands within each set, then rank them by calculated ΔE values according to equation (1). After sorting the set by increasing ΔE value, the priority ranking of binders over non-binders was assessed by taking the area (AUC) under the ROC curve[50]. Eventually, the same sets of compounds were subjected to docking, rescoring and therewith associated ranking, with three popular, state-of-the-art docking programs: Glide[51], Gold[52] and Autodock-Vina[53]. There is no absolute winner in this benchmarking study – all programs reached near-perfect results on some sets (ROC AUC $\geq$ 0.9) but experienced significant problems with at least one of the sets (ROC AUC$\approx$0.5, meaning perfectly random ranking of candidate ligands). If the four docking tools were assigned gold/silver/bronze medals for each of the 11 individual challenges (sets), in decreasing order of their set-specific ROC AUC values, all the programs would have won "gold" at least once, and all except Gold would have also failed to obtain a medal at least once.

S4MPLE does not stand out of the pool of the four benchmarked approaches – neither as the best, nor as the worst performer. However, it uses a FF-based Hamiltonian to drive the docking (by contrast to Gold and Autodock-Vina, using dedicated scoring functions as drivers), and does *not* feature any pose rescoring function, in contrast with Glide. It is true that the FF engine in S4MPLE is endowed with potentials of mean force for desolvation/hydrophobic effect that are typical to a scoring function. Yet, the current parameterization[43] of these additional terms never relied on affinity-ranking simulations. Finally, S4MPLE is also useful for single-species (including peptide) sampling, *i.e.* it is not limited to a docking protocol requiring a protein site and an organic ligand. This allows for a consistent introduction of the energy penalty typically "paid" by the ligand when attaining the bioactive conformation. The results presented in this work therefore suggest that a direct, proper parameterization of the FF engine of docking tools may represent a more elegant and more parsimonious solution to the scoring problem.

3 Methods

3.1 Datasets

Structure-activity sets automatically extracted from ChEMBL (v.20) for external validation of a drug space mapping project[54] (see cited publication for the data curation protocol) were one key source of benchmark compounds of this work. Each set is associated to a protein target (enzyme or

receptor) and contains only compounds with experimentally known activities for that target. Compounds with the highest activity levels are marked as "active", the others count as "inactive" (see cited paper for the activity label assignment procedure).

Next, the QSAR modelability of each set was assessed, in order to demonstrate that assigned actives and inactives, unavoidable errors of the active/inactive labeling procedure notwithstanding, are *separable* based on the structural information contained in 2D ISIDA descriptors[55]. If successfully cross-validating QSAR models could be developed, then any docking failure to achieve such separation cannot be attributed to incoherent active/inactive labeling. SVM classification model building[56] with evolutionary optimization of model parameters (including the choice of the ISIDA descriptor space) was undertaken for each set, following the default aggressive procedure of 12 times-repeated three-fold cross-validation. The objective function of the model builder was cross-validated balanced accuracy (XV-BA). However, in order to enable the direct comparison to docking results, the performance of the best model (of maximal XV-BA) was expressed as a ROC AUC value, as follows: During each of the 12 repeated three-fold cross-validation procedure cycles, each compound is assigned exactly once to the left-out tier of items serving as external prediction set. Each compound harvests thus exactly 12 independent "votes" in favor or against the hypothesis that it is active. The total number of votes in favor was used as the scoring criterion with respect to which the set was sorted, in descending order, generating the cross-validated ROC curve and reporting its AUC.

Seven targets associated with large, QSAR-modelable sets and with experimentally solved crystal structures were selected for this benchmarking study. As can be seen from Table 1, they include one GPCR (the Angiotensin receptor ATII), two kinase receptors and key enzymes of various families (cyclooxygenase, phosphodiesterase, protease, histone deacetylase). The Protein Data Bank structures of the targets were looked up, and if multiple structures were present, a convenient high-resolution representative, cocrystallized with a ligand was selected (see Table 1). As this benchmark considers only rigid-site docking, the problem of the site flexibility was ignored, neither has it been attempted to pick the site geometry potentially maximizing docking success over the considered sets. All crystal waters were deleted, and hydrogen atoms were added, following side chain protonation rules corresponding to physiological pH, by the VegaZZ[57] software. Further active site preparation steps are docking software-specific and will be described below.

Table 1: Targets considered for the benchmarking studies, with the ChEMBL IDs and the PDB code of the protein structure used for docking.

Target ChEMBL ID	PDB Code	Target Name
CHEMBL1827	4OEX	Phosphodiesterase V
CHEMBL1865	3GV4	Histone Deacetylase
CHEMBL203	1XKK	Epidermal Growth Factor Receptor
CHEMBL204	1BHX	Thrombin
CHEMBL227	4YAY	Angiontensin receptor II
CHEMBL230	3LN1	Cyclooxygenase-2
CHEMBL279	1YWN	Vascular Endothelial Growth Factor Receptor 2

For three of these targets (Thrombin, Cyclooxygenase-2 and Phosphodiesterase V), present in the extended Directory of Useful Decoys (E-DUD), the associated active/decoy sets were also co-opted into this study. As the DUD set of Cyclooxygenase-2 is very large (~14K compounds), it was randomly split into two sets, to be docked in parallel on different hardware. The sets, as two-column (SMILES, activity class) text files are provided as Supplementary Material.

3.2 S4MPLE

S4MPLE (Sampler For Multiple Protein-Ligand Entities), a molecular modeling program based on a Lamarckian genetic algorithm, has been described previously[43-45]. This conformational tool, allowing the selection of the degrees of freedom of the system to be considered during search, can be employed for a wide variety of simulation types: conformational sampling of ligands or small peptides, and docking of both fragment-sized and drug-sized compounds. There is no explicit limit with respect to the number of considered entities – simultaneous docking of multiple ligands is supported. The energy function relies on the force field (FF) formalism, and uses AMBER [46] and GAFF [47] to respectively simulate peptide and small organic moieties of the considered system. Here, all simulations are performed with the "Fit FF" energy scheme described, calibrated and validated previously[43]. The control of conformational similarity is performed by a symmetry-compliant pair-based interaction fingerprint (PIF) which monitors two interaction types: close carbons contacts (based on C-C distance) and hydrogen-bonds. Contacts monitored in the fingerprints may contribute to the hydrophobic or hydrogen bonding energy terms if they are assigned non-zero weights. Two configurations of the system are considered equivalent if the Hamming[58] distance between their fingerprints is lower than a user-defined threshold. The program is written in object-Pascal, and used in command-line mode.

3.3 S4MPLE Docking Protocol

3.3.1 Active Site Preparation

All protein atoms were fixed, by enumerating their sequence numbers into the dedicated *fixed_atoms* file. S4MPLE uses a predefined cutoff of 12 Å for non-bonded interactions. Protein atoms that are too far from the active site in order to ever come within 12 Å to any ligand atom would merely slow down calculations by requesting the continuous update of their distances to ligand atoms. Therefore, docking was not run on the entire protein, but on the selection of relevant residues that have at least one atom at less than 10 Å from any of the co-crystallized ligand, herewith used to define the active site region. Moreover, S4MPLE requires the user-specified input of "hot spots" – key solvent-accessible atoms, chosen preferentially at the bottom of the site cavity, which are used for random prepositioning of the ligand into the active site. These may, but do not have to include site atoms seen to make contacts to the cocrystallized PDB ligand. Their choice has no impact on the docking energy function (they are not used to tether the ligand).

3.3.2 Ligand Preparation

Ligands, initially provided as standardized SMILES, preprocessed by the standardization tool of the Strasbourg virtual screening web server http://infochim.u-strasbg.fr/webserv/VSEngine.html, underwent an automated conversion, by means of an in-house tool developed on the basis of the ChemAxon API, to a fully protonated initial 3D structure. The tool relies on the tautomer[59] and respectively pK_a plugin[60] to generate the most probable microspecies of the expected main tautomeric form (alternatively, users might request several tautomeric/protonation states to be generated, and each to be docked as an independent candidate – but the option was not used here). Explicit hydrogens are assigned, and a single conformer is then generated, by the conformer plugin. Eventually, the charge plugin[61] is used to assign Gasteiger charges to this structure. Last, the tool detects flexible rings and proposes, for each, the single bond to be formally "broken" in order to enable intra-cyclic torsional axes to be driven by S4MPLE. Next, antechamber[62] and other utilities, as called by GAFF pilot scripts, are used to assign GAFF ligand types, and to automatically generate associated FF parameters for the internal coordinates found in the ligand, if such did not yet exist. FF types and Gasteiger charges are added as data S4MPLE-readable fields to the MDL sd file used to store the proposed initial conformer of the ligand. The fully parameterized sd file and – if applicable – the file

with the aforementioned intra-cyclic bonds are added to the tar archive of the ligand set, the final product of the ligand preprocessing script.

As GAFF parameterization is bound to generate new FF parameters, all ligands were processed on a same Linux machine, in a sequential way (parallelization of the task is difficult, because of the risk of concurrent writing access of new parameters to the updatable FF files). Nevertheless, the process is conveniently fast (hundreds to thousands of ligands/hour, depending on their complexity). Note that updated FF files need to be exported to the hardware platforms used for docking, if they are different from the machine used for ligand preprocessing.

3.3.3 S4MPLE Docking

S4MPLE docking scripts were written for parallel processing of ligand sets on various types of hardware: local multicore workstations, SLURM-driven clusters or gLite-driven computer grids. Irrespective of the environment, the procedure begins by extracting all the data pertaining to a given ligand into a dedicated directory, then running a 200-generation evolutionary conformational search with S4MPLE, on the free ligand, at default settings. This basic setup is considered to be sufficient for rather rigid, drug-like ligands. Most stable free ligand conformers are stored on disk, together with their intramolecular energies $\langle E_i \rangle^{ligand}$. Next, active site data (molecular file as Tripos mol2, plus the required fixed*_atoms* and *hot_spots* files) are added to this directory. Upon restart, S4MPLE will thus detect the presence of two partner molecules, and seamlessly switch into "docking" mode.

A first brief simulation is run with the S4MPLE *testDiff* option, in order to calibrate the optimal cutoff for the interaction fingerprint dissimilarity value *minfpdiff*, representing the threshold at which two conformers are considered as redundant, and thus pruned during the evolutionary process. The proper management of population diversity has been found to be of paramount importance with respect to ensuring the convergence/reproducibility of evolutionary simulations. As ligands vary in sizes, so does their interaction fingerprint, making it difficult to come up with a universally applicable *minfpdiff* value – hence, the need to calibrate it for each system. The population initialization procedure, normally serving as the first step for the evolutionary simulation, is called repeatedly (10 times). After each call, the interaction fingerprints of the randomly generated population members are compared to each other, generating the complete Hamming distance matrix for all pairs of conformers in the population. The lowest, mean and maximal Hamming distances for each population are stored. The average of these lowest, mean and maximal Hamming distances over the 10 visited random

populations are output to the disk. The *minfpdiff* threshold is defined as 90% of the average of the ten lowest intra-population Hamming distances.

Eventually, the main docking simulation is started as a G=1000-generation evolutionary optimization, with the above-determined *minfpdiff* value as a population diversity control parameter. Top poses are generated and stored together with their energy values $\langle E_i \rangle^{ligand@site}$. The energy of the ground state of the *apo* protein is not of direct interest in this study, being constant throughout a set of ligands bound to a same target. Therefore, the docking index ΔE for the current ligand can be directly estimated as $\langle E_i \rangle^{ligand@site} - \langle E_i \rangle^{ligand}$. After completion of docking calculations for all ligands, these can be rank ordered by increasing ΔE, and the "final" ROC curve can be generated in order to determine the area under it, the final benchmarking criterion. However, because S4MPLE will report the so-far best energy value achieved at every generation, it is also possible to retrospectively trace the ROC curve of "premature" results that would have been obtained if the docking process would have been stopped at G<1000 generations, by taking the so-far best energy value reached at generation G instead of the final $min_i \langle E_i \rangle^{ligand@site}$. The variation of the ROC AUC as a function of performed number of generations may be informative about the minimal required computational effort needed in typical S4MPLE docking simulations.

3.4 Docking with Glide, Gold and Autodock-Vina

Prior to docking with Glide, Gold and Autodock-Vina, ligands and proteins (except for Gold; see below) were prepared using Schrödinger Maestro's (Version 10.5, 2016-1 release) LigPrep and Prep Wiz applications (respectively) with default settings, except the following:

LigPrep: Protonation states were calculated at pH = 7±0.5 and specific chiralities were retained.

Prep Wiz: Water molecules were deleted, missing side chains were added (if needed), pKa values were calculated (at pH = 7) to assign the correct protonation states for all titratable residues and finally, restrained minimization was carried out.

3.4.1 Glide[27-29]

Glide docking was performed via Schrodinger's Maestro (Version 10.5, 2016-1 release) using default settings. The grid location and size were set automatically defined, based on the crystallographic ligands. Docking calculations were performed using the standard precision protocol (Glide – SP), and only a single pose having the best Glide score was retained.

3.4.2 Gold[25, 26]

Gold docking was performed with version 5.5. Prior to docking, all protein-ligand complexes were prepared following a standard 'wizard' workflow as implemented in the interactive 3D visualization program Hermes (version 1.8.2). As part of this process all crystallographic water molecules were removed. In all cases the ligand binding site was defined based on the coordinates of the crystallographic ligand. Docking was performed with default parameters (10 docking runs for each ligand, 10^5 generations for each run, flip all planar R-NR-1R2, flip protonated carboxylic acids, default torsion angle distributions). Both docking and scoring used the (default) CHEMPLP scoring function. Following docking, a single best docking conformation of each input ligand was selected based on the CHEMPLP score.

3.4.3 Autodock-Vina[53]

Protein and ligand structures were converted to pdbqt format, using Auto Dock Tools 1.5.6 and the appropriate scripts (as implemented in MGLtools). All rotatable bonds of the ligands were allowed to freely rotate, while the protein was held rigid. Partial charges calculated by LigPrep were retained. Docking was performed with Autodock Vina 1.1.2 using default parameters. In all cases, the grid box was centered on the active site of the protein and the spacing between the grid points was set to 1 Å. Docking was streamlined using in-house written scripts. A single pose with the best score was retained for each ligand.

4 Results & Discussion

4.1 QSAR Modelability of Benchmarking Data Sets

Table 2 below provides an overview of the eleven benchmarking sets, reporting their size, number of compounds being labeled as "active" and, eventually, the cross-validated ROC-AUC of the SVM classification models generated as described in §3.1, in order to assess the modelability of the sets. Here, good modelability – taken as success in achieving robustly cross-validating classification models – is a demonstration that the sets are actually rich in structure-activity information. The compounds labeled as active are structurally distinct from the tested inactives, as well as from DUD decoys – in a sense that can be captured by molecular descriptors such as the ISIDA fragment counts used here. This also means that activity labels are being reliably assigned, experimental error and

empiricism of the label assignment protocol notwithstanding – or else the classes would not be separable in a cross-validated machine learning attempt. This is true throughout the eleven sets, even though they widely differ in terms of the degree of imbalance between the actives and inactives: the fraction of actives may range from almost 50% to merely 3%.

Actually, all the sets are almost perfectly separable by machine learning. This is perhaps not surprising, noting that on one hand these targets are amongst the best known in drug design, thus benefitting from large and coherent series of tested ligands. Furthermore, the evolutionary model optimizer is being given the option to choose the most suitable ISIDA fragmentation scheme, out of one hundred considered possibilities that were preselected because of their recurrent success in QSAR model building. They not only cover various strategies to define fragments (atom sequences, circular fragments, Carhart-style[63] atom pairs, *etc*) but also propose different coloring schemes (by atom type, by FF type, by pharmacophore type), capturing distinct chemical information[64, 65]. Some of these schemes are seen to precisely encode the structural patterns that best correlate with activity on a given target. It is also important to note that ChEMBL "real-life" series of tested actives and inactives are very well separable, while DUD sets are *perfectly* separable. The distinction between actives and human-selected decoys is structurally much more obvious than the ones between actives and inactives which were actually designed to be actives, and are therefore quite often forming genuine "activity cliffs".

Table 2: The eleven benchmarking sets, denoted as (data source)_(target PDB code), where "data source" indicates either the ChEMBL ID of the target corresponding to the PBD code (for ChEMBL sets of actually tested actives and inactives) or "DUD" for sets of actives and decoys. SVM ROC-AUC is reported for the SVM classification models built in order to assess the modelability of these sets.

SET	Set size	Number of Actives	SVM ROC-AUC
CHEMBL1827_4OEX	1427	683	0.978
DUD_4OEX	2065	88	0.999
CHEMBL1865_3GV4	747	231	0.966
CHEMBL203_1XKK	5019	1212	0.985
CHEMBL204_1BHX	2915	1272	0.970
DUD_1BHX	2528	72	0.999

CHEMBL227_4YAY	948	403	0.997
CHEMBL230_3LN1	3129	626	0.967
DUD.1_3LN1	6843	213	0.999
DUD.2_3LN1	6842	213	0.999
CHEMBL279_1YWN	5235	1532	0.974

4.2 S4MPLE Docking Performance as a Function of Sampling Effort

How would premature stopping of the docking procedure impact the relevance of ΔE scores as a ligand prioritization criterion? Since ΔE is based on the lowest FF energy found by the evolutionary simulation, this question can be easily answered by monitoring the decrease of energy of the so-far most stable pose after each generation G. Using these so-far best pose energies instead of the final energies following simulation completion (at G=1000) allows a direct monitoring of the ROC AUC values as a function of conformational sampling effort at docking stage. This study, the results of which are illustrated in Figure 1, has a two-fold interest:

- First, it represents an internal consistency check of the docking procedure: the docking score ΔE is expected to *gain* in proficiency as sampling improves. Such a behavior would demonstrate that docking works because of the specific ligand-site contacts being discovered, and their importance to the binding energy being highlighted. Should ΔE scores turn out to be "relevant" and separate actives from inactives in spite of obviously insufficient sampling (say, after only 10 evolutionary generations), this would likely hint to some artefactual behavior – presumably, size artifacts in a set where "actives" are larger, thus prone to show more non-specific contacts. Fortunately, this is not observed: ROC AUC is seen to steadily increase – even if final results may be deceiving for the few sets that are not successfully docked by S4MPLE. Some better-than-random results are nevertheless obtained even at 10 generations. Unsurprisingly, they all concern DUD sets, which were already demonstrated to be "too easy" to separate, thus most likely to suffer from biases as above-mentioned. The presence of the "twin" sets DUD_3LN1, equal-size random halves of the Cycloxygenase-2 DUD set furthermore provides an idea of the reproducibility of the ROC AUC trends. They are expected and indeed found to behave similarly, in spite of their processing on different computer systems.
- Last but not least, the study provides an estimated value for the maximal number of generations, *i.e.* a termination criterion for S4MPLE runs: a few hundreds of generations appear to be sufficient

for most of the targets. This is significantly less than the 1000 used here and consistent with the previous study[43] aimed at properly reproducing native ligand poses. It is debatable whether some of the shown curves might have pursued their growth beyond 1000 generations – in particular in the case of Thrombin CHEMBL204_1BHX known for its rather flexible ligands, and which fails to rise beyond the randomness threshold of ROC AUC=0.5.

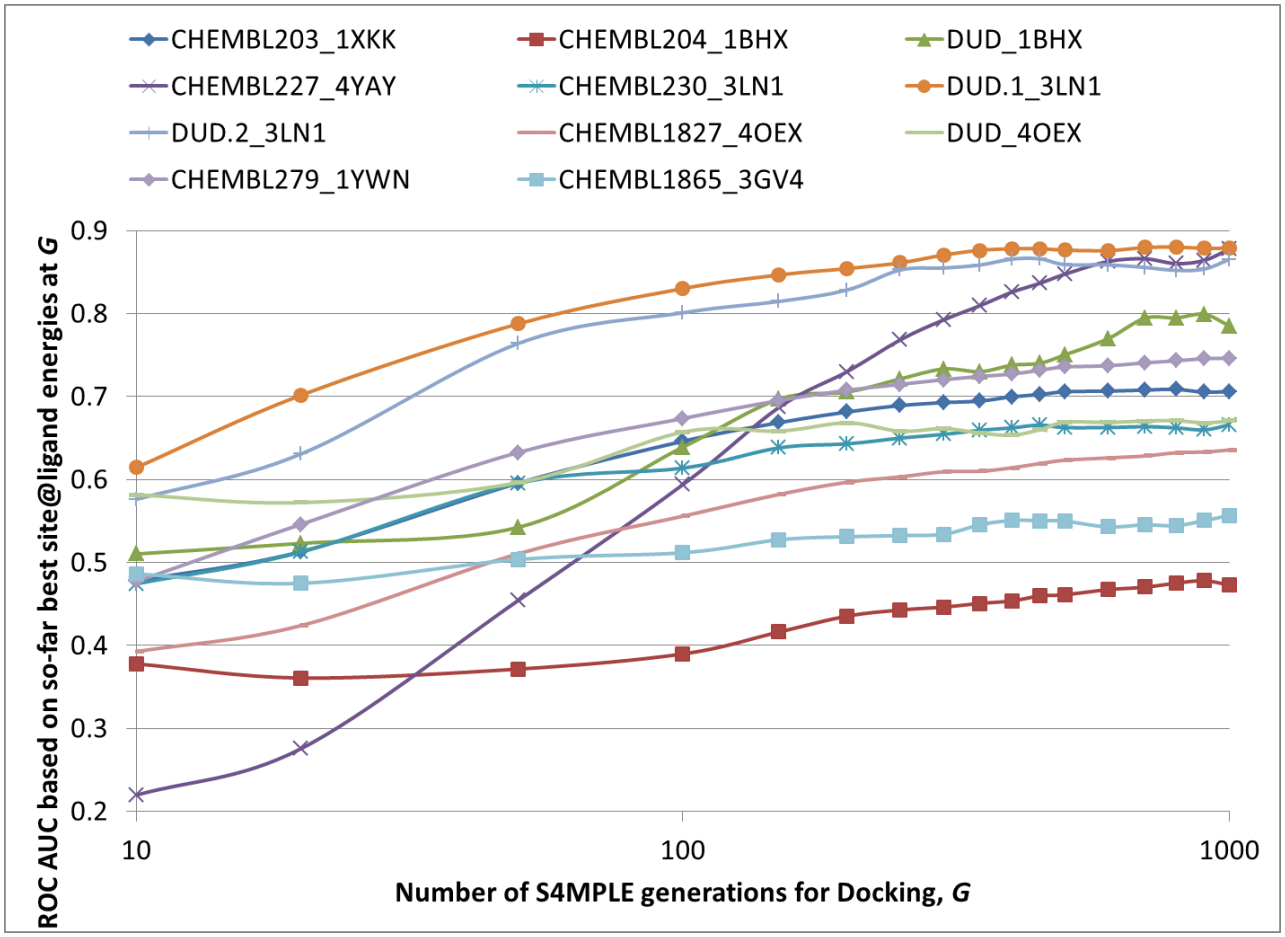

Figure 1: Dependence of S4MPLE ROC AUC scores on the docking effort, expressed in terms of evolutionary generations.

4.3 Benchmarking of S4MPLE versus Classical Docking/Rescoring Approaches

First, it is important to point out that S4MPLE is not a dedicated docking tool, but a broad applicability range conformational sampler. As such, it is significantly slower than commercial docking programs: one full ligand docking procedure may take several hours on a x86_64 core. Therefore, parallelization of the S4MPLE docking effort on clusters and computer grids is mandatory.

Last but not least, this benchmarking should be regarded as a global comparison of the four docking protocols. If one docking protocol were to systematically stand out as the best (or worst), it would unfortunately not be possible, on the basis of herein reported results, to link this outstanding behavior to any specific feature of the docking protocol – it might be due to differences in the handling of the protein site, in the handling of ligand protonation states, in the pose sampling procedure or eventually in the pose rescoring – or absence thereof. Systematic underperformance of a tool may also be a warning of a systematic malpractice in the hands of an inexperienced user. Fortunately, no such trends were noticed. All docking protocols were kept to default choices of operational parameters (i.e. no tuning of the latter was undertaken in order to maximize the performance of any given tool) in order to provide a fair comparison.

As can be seen from Table 3 below (columns N_{OK}), not all the compound set members could be successfully docked by all the tools. On one hand, S4MPLE was "privileged" by the fact that compound import and standardization already relied on ChemAxon tool, so that chemotypes causing problems with the ChemAxon API were tacitly discarded before the final compilation of docking sets. Thus, all the compounds successfully passed the ChemAxon-driven ligand preprocessing phase, and the very versatile GAFF parameter generation/assignment tool. On the other hand, S4MPLE was disadvantaged because it was operated on failure-prone computer grids and public clusters. Indeed, docking of several ligands failed due to grid or cluster malfunctions. Because of practical constraints, those simulations were not restarted. Commercial software, running on stable multi-CPU workstations, often registered failures, presumably caused by internal parameterization problems for specific chemotypes.

Table 3: Benchmarking results, reporting the numbers of successfully processed set members (N_{OK} – highlighted in red when outstandingly low; compare to set sizes in Table 2) and the resulting ROC AUC scores, for each of the benchmarked docking tools. AUC cell coloring reflects "medals" given for each challenge: "gold", "silver", "bronze" and "taillight", in decreasing ROC AUC order. If AUC was below 0.55, a "taillight" status is assigned by default.

SET/METHOD	S4MPLE		GLIDE		GOLD		AutoDock-Vina	
	N_{OK}	AUC	N_{OK}	AUC	N_{OK}	AUC	N_{OK}	AUC
CHEMBL1827_4OEX	1423	0.636	1370	0.576	1427	0.786	1427	0.590
DUD_4OEX	2031	0.671	2051	0.773	2063	0.757	2063	0.591
CHEMBL1865_3GV4	731	0.552	744	0.453	747	0.549	742	0.603

CHEMBL203_1XKK	4676	0.706	4985	0.713	5012	0.686	5011	0.550
CHEMBL204_1BHX	2662	0.473	2896	0.526	2907	0.624	2759	0.609
DUD_1BHX	2355	0.785	2523	0.898	2528	0.813	2528	0.628
CHEMBL227_4YAY	927	0.878	948	0.892	948	0.907	948	0.767
CHEMBL230_3LN1	3084	0.666	2662	0.757	3126	0.725	3124	0.743
DUD.1_3LN1	6748	0.879	5382	0.969	6835	0.952	6835	0.925
DUD.2_3LN1	6793	0.865	5417	0.990	6837	0.933	6837	0.924
CHEMBL279_1YWN	5148	0.745	5212	0.618	5232	0.701	5231	0.679

The analysis of ROC AUC scores above firstly shows that all sophisticated, time-consuming 3D docking calculations lag far behind ultrafast 2D machine-learned models (see Table 2), in terms of active/inactive separation. The comparison is however not fair – the machine-learned models were specifically trained on these sets. Even though the employed three-fold cross-validation scheme was as "aggressive" as feasible, the left-out compounds likely had reasonably near neighbors in the learning sets used to calibrate the models predicting them.

It is very likely that training set for the scoring functions in Glide, Gold and Autodock-Vina are insignificantly or not at all overlapping with the herein employed compounds. This notwithstanding, results in Table 3 nevertheless confirm the "nearly pathological" ability of separation of DUD actives *versus* decoys: DUD sets consistently have the best ROC AUC scores, all software confounded (including 4OEX, but to a lesser extent). This raises – once more – the question of the usefulness of artificial benchmarking sets, in the context where public and experimentally validated structure-activity data is increasingly available.

As for S4MPLE, it was set up using a completely different paradigm (native ligand redocking success) on completely unrelated compounds and targets – meaning that the additional FF parameter fitting only focused on the position of the docking energy minima, but completely ignored the question of their actual *depth.*

The key conclusion that emerges from Table 3 is that all four docking protocols are seen to behave remarkably similarly in this test. "Difficult" docking sets – notably the ChEMBL-extracted Thrombin set CHEMBL204_1BHX and the Histone Deacetylase set CHEMBL1865_3GV4 – are a real challenge for all methodologies. Easy sets – of DUD provenience – are well-docked by most tools. Even if one would exacerbate the observed ROC AUC value differences by assigning "medals" to the docking protocols in terms of strict performance ranking, as was done in Table 3, this would still not

highlight any obvious winner of the benchmarking challenge. Autodock-Vina appears most often in the "taillight" position, especially when considering that S4MPLE earned two of its taillight scores with excellent ROC AUC values close to 0.9. Glide is often a winner, but often rather disappointing, while Gold and S4MPLE overall steady, reasonably successful approaches. In terms of means of ROC AUC values over all compound sets, Gold reaches a value of 0.77±0.12, followed by Glide with 0.74±0.17, S4MPLE with 0.71±0.13 and Autodock-Vina, with 0.69±0.13. Means of the best and of the worst performer are within less than one standard deviation, which is another way to highlight the absence of any significant differences in the overall proficiency of these approaches.

Finally, we wish to point out that other metrics for evaluating the performances of docking tools are available, for example, the ability of the tool to reproduce experimentally observed binding modes. However, most docking tools are able to find at least one pose which is close (RMSD-wise) to the experimental binding mode. The problem is that these poses are seldom ranked in the first place. Thus, the so-called scoring problem is more complicated than the so-called docking problem and therefore in this work we focus on the former.

5 Conclusion

Having an additional layer for ligand pose rescoring therefore does not seem to bring any direct competitive advantage for docking. Implementing a very simple continuum solvent model in the FF engine is an option shown to be equally effective, but conceptually more parsimonious – in the sense of Occam's Razor principle – over the to-date privileged procedure of scoring function fitting. Moreover, improving the FF is an intrinsic necessity of molecular modeling and should benefit all the possible applications of the method, not only docking. By contrast, scoring function fitting is a docking-specific problem. Although S4MPLE energy well depth was never subjected to explicit fine-tuning, the method performed just as well as dedicated docking techniques. This notwithstanding, it cannot be ignored that all methods did outright or nearly fail for certain compound sets – meaning that explicit fine-tuning of S4MPLE energy well depth should be undertaken, as a potential way to contribute to the much-needed improvement of docking methodology.

6 Acknowledgements

The authors wish to thank the staff of the two computer centers which hosted the simulations: HPC (High-performance computing) of the University of Strasbourg and HPC of the chemistry faculty of Cluj-Napoca.

7 Supplementary Information

The program S4MPLE (x86_64) version can be dowloaded from our laboratory web site http://infochim.u-strasbg.fr (see Downloads). A tar file with compound series being used in this work is provided as Supplementary Information.

8 References

1. Pason, L. P.; Sotriffer, C. A., Empirical Scoring Functions for Affinity Prediction of Protein-ligand Complexes. *Molecular Informatics* **2016,** *35* (11-12), 541-548.
2. Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J., Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2015,** *5* (6), 405-424.
3. Cleves, A. E.; Jain, A. N., Knowledge-guided docking: accurate prospective prediction of bound configurations of novel ligands using Surflex-Dock. *Journal of Computer-Aided Molecular Design* **2015,** *29* (6), 485-509.
4. Lindh, M.; Svensson, F.; Schaal, W.; Zhang, J.; Skold, C.; Brandt, P.; Karlen, A., Toward a Benchmarking Data Set Able to Evaluate Ligand- and Structure-based Virtual Screening Using Public HTS Data. *Journal of Chemical Information and Modeling* **2015,** *55* (2), 343-353.
5. Xu, W. J.; Lucke, A. J.; Fairlie, D. P., Comparing sixteen scoring functions for predicting biological activities of ligands for protein targets. *Journal of Molecular Graphics & Modelling* **2015,** *57*, 76-88.
6. Parenti, M. D.; Rastelli, G., Advances and applications of binding affinity prediction methods in drug discovery. *Biotechnology Advances* **2012,** *30* (1), 244-250.
7. Neudert, G.; Klebe, G., DSX: A Knowledge-Based Scoring Function for the Assessment of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling* **2011,** *51* (10), 2731-2745.
8. Shen, Q. C.; Xiong, B.; Zheng, M. Y.; Luo, X. M.; Luo, C.; Liu, X. A.; Du, Y.; Li, J.; Zhu, W. L.; Shen, J. K.; Jiang, H. L., Knowledge-Based Scoring Functions in Drug Design: 2. Can the Knowledge Base Be Enriched? *Journal of Chemical Information and Modeling* **2011,** *51* (2), 386-397.
9. Guvench, O.; MacKerell, A. D., Jr., Comparison of protein force fields for molecular dynamics simulations. *Methods in Molecular Biology* **2008,** *443*, 63-88.
10. Damm, W.; Van Gunsteren, W. E., Reversible peptide folding: Dependence on molecular force field used. *Journal of Computational Chemistry* **2000,** *21* (9), 774-787.
11. Rasmussen, K., Consistent force fields for saccharides. *Journal of Carbohydrate Chemistry* **1999,** *18* (7), 789-805.
12. Halgren, T. A., Merck molecular force field. *Journal of Computational Chemistry* **1996**.
13. Halgren, T. A., Potential Energy Functions. *Current Opinion in Structural Biology* **1995,** *5* (2), 205-210.
14. Doweyko, A. M., QSAR: dead or alive? *Journal of Computer-Aided Molecular Design* **2008,** *22* (2), 81-89.

15. Gonzalez, M. P.; Teran, C.; Saiz-Urra, L.; Teijeira, M., Variable Selection Methods in QSAR: An Overview. *Current Topics in Medicinal Chemistry* **2008,** *8*, 1606-1627.
16. Klebe, G. Understanding QSAR: Do we always use the correct structural models to establish affinity correlation? http://www.qsar2008.org/home/FA04-10-12-42_h6vpw99c3zxmfq28f4e9/qsar2008.org/public_html/File/abstract%20session%207/Klebe_QSAR_Uppsala_2008.pdf (accessed 2009).
17. Maggiora, G. M., On outliers and activity cliffs - Why QSAR often disappoints. *J. Chem. Inf. Model.* **2006,** *46*, 1535-1535.
18. Mullinax, J. W.; Noid, W. G., Recovering physical potentials from a model protein databank. *Proceedings of the National Academy of Sciences of the United States of America* **2010,** *107* (46), 19867-19872.
19. Blundell, T. L.; Sibanda, B. L.; Sternberg, M. J.; Thornton, J. M., Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **1987,** *326* (6111), 347-352.
20. Watson, P., Naive Bayes classification using 2D pharmacophore feature triplet vectors. *Journal of Chemical Information and Modeling* **2008,** *48* (1), 166-178.
21. Horvath, D., A Virtual Screening Approach Applied to the Search of Trypanothione Reductase Inhibitors. *J. Med. Chem.* **1997,** *15*, 2412-2423.
22. Ding, F.; Dokholyan, N. V., Incorporating Backbone Flexibility in MedusaDock Improves Ligand-Binding Pose Prediction in the CSAR2011 Docking Benchmark. *Journal of Chemical Information and Modeling* **2013,** *53* (8), 1871-1879.
23. Krüger, D. M.; Jessen, G.; Gohlke, H., How Good Are State-of-the-Art Docking Tools in Predicting Ligand Binding Modes in Protein–Protein Interfaces? *Journal of Chemical Information and Modeling* **2012,** *52* (11), 2807-2811.
24. Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V., MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening. *Journal of Chemical Information and Modeling* **2008,** *48* (8), 1656-1662.
25. Jones, G.; Willett, P.; Glen, R. C., Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology* **1995,** *245* (1), 43-53.
26. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **1997,** *267* (3), 727-748.
27. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *Journal of Medicinal Chemistry* **2006,** *49* (21), 6177-6196.
28. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *Journal of Medicinal Chemistry* **2004,** *47* (7), 1750-1759.
29. 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.; Shenkin, P. S., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry* **2004,** *47* (7), 1739-1749.
30. McGann, M., FRED Pose Prediction and Virtual Screening Accuracy. *Journal of Chemical Information and Modeling* **2011,** *51* (3), 578-596.
31. McGann, M., Hybrid docking with FRED. *Abstracts of Papers of the American Chemical Society* **2011,** *241*.
32. McGann, M., FRED and HYBRID docking performance on standardized datasets. *Journal of Computer-Aided Molecular Design* **2012,** *26* (8), 897-906.
33. Morris, G. M. AutoDock. http://autodock.scripps.edu/.

34. Horvath, D.; Marcou, G.; Varnek, A., Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J. Chem Inf. Model.* **2009,** *49* (7), 1762-1776.
35. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T., QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA Alternatives to Laboratory Animals* **2005,** *33* (5), 445-459.
36. Brewerton, S. C., The use of protein-ligand interaction fingerprints in docking. *Current Opinion in Drug Discovery & Development* **2008,** *11* (3), 356-364.
37. Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S., A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): Theory and application. *J. Chem. Inf. Model.* **2007,** *47* (2), 279-294.
38. Marcou, G.; Rognan, D., Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007,** *47* (1), 195-207.
39. Choudhury, N., Montgomery-Pettitt, B., The Dewetting Transition and The Hydrophobic Effect. *J. Am. Chem. Soc.* **2007,** *129*, 4847 -4852.
40. Chandler, D., Interfaces and the driving force of hydrophobic assembly. *Nature* **2005,** *437* (7059), 640-647.
41. Bohm, H. J.; Stahl, M., The use of scoring functions in drug discovery applications. In *Reviews in Computational Chemistry, Vol 18*, Wiley-Vch, Inc: New York, 2002; Vol. 18, pp 41-87.
42. Liu, L.; Yang, C.; Guo, Q. X., A study on the enthalpy-entropy compensation in protein unfolding. *Biophys. Chem.* **2000,** *84*, 239-251.
43. Hoffer, L.; Chira, C.; Marcou, G.; Varnek, A.; Horvath, D., S4MPLE-Sampler for Multiple Protein-Ligand Entities: Methodology and Rigid-Site Docking Benchmarking. *Molecules (Basel, Switzerland)* **2015,** *20* (5), 8997-9028.
44. Hoffer, L.; Renaud, J.-P.; Horvath, D., In Silico Fragment-Based Drug Discovery: Setup and Validation of a Fragment-to-Lead Computational Protocol Using S4MPLE. *J. Chem. Inf. Model.* **2013,** *53* (4), 836-51.
45. Hoffer, L.; Horvath, D., S4MPLE - Sampler For Multiple Protein-Ligand Entities: Simultaneous docking of several entities. *J Chem Inf Model* **2012,** *53* (1), 88-102.
46. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P., AMBER a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995,** *91* (1-3), 1-41.
47. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004,** *25* (9), 1157-1174.
48. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2011,** *40* (D1), D1100-D1107.
49. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K., Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry* **2012,** *55* (14), 6582-6594.
50. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. In *Beyond Accuracy, F-score and ROC: A Family of Discriminant Measures for Performance Evaluation*, AAAI Workshop - Technical Report, 2006; pp 24-29.
51. Schrödinger, L. *Glide*, New York, 2005.
52. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved Protein-Ligand Docking Using GOLD. *Proteins: Struct., Funct., Genet.* **2003,** *52*, 609-623.

53. Trott, O.; Olson Arthur, J., AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2009,** *31* (2), 455-461.
54. Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D., Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *Journal of Computer-Aided Molecular Design* **2015,** *29* (12), 1087-1108.
55. Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. v.; Marcou, G., Isida - Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008,** *4* (3), 191-198.
56. Horvath, D.; Brown, J.; Marcou, G.; Varnek, A., An Evolutionary Optimizer of libsvm Models. *Challenges* **2014,** *5* (2), 450-472.
57. Pedretti, A.; Villa, L.; Vistoli, G., VEGA – An open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming. *Journal of Computer-Aided Molecular Design* **2004,** *18* (3), 167-173.
58. Willett, P., Barnard, J. M., Downs, G. M. , Chemical Similarity Searching. *J. Chem. Inf. Model.* **1998,** *38*, 983-996.
59. ChemAxon Tautomer Plugin. http://www.chemaxon.com/marvin-archive/4.1.3/marvin/chemaxon/marvin/help/calculator-plugins.html#tautomer (accessed Oct. 2011).
60. ChemAxon pKa Calculator Plugin. https://www.chemaxon.com/products/calculator-plugins/property-predictors/ (accessed Feb. 2013).
61. ChemAxon Calculation of Partial Charge Distributions. http://www.chemaxon.com/marvin/help/calculations/charge.html (accessed Feb. 2009).
62. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A., AMBER 12. *University of California* **2012**.
63. Carhart, E. R., Smith, D.H., Venkataraghavan, R., Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985,** *25*, 64-73.
64. Laboratoire de Chemoinformatique Strasbourg. Nomenclature of ISIDA Fragments 2012. http://infochim.u-strasbg.fr/recherche/Download/Fragmentor/Nomenclature_of_ISIDA_fragments_2011.pdf.
65. Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D., Isida Property-labelled Fragment Descriptors. *Molecular Informatics* **2010,** *29* (12), 855-868.

Click here to access/download

Compressed File (do not unpack)

ligand_series.tar.gz