All in One: Cavity Detection, Druggability Estimate, Cavity-Based Pharmacophore Perception, and Virtual Screening

Viet-Khoa Tran-Nguyen, Franck da Silva, Guillaume Bret, Didier Rognan

All in One: Cavity Detection, Druggability Estimate, Cavity-Based Pharmacophore Perception and Virtual Screening.

Viet-Khoa Tran-Nguyen[†#], Franck Da Silva[†#], Guillaume Bret[†] and Didier Rognan[†*]

[†] Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 67400 Illkirch, France.

[#] Both authors equally contributed to this work.

* To whom correspondence should be addressed (phone: +33 3 68 85 42 35, fax: +33 3 68 85 43 10, email: rognan@unistra.fr)

ABSTRACT

Discovering the very first ligands of pharmacologically important targets in a fast and cost-efficient manner is an important issue in drug discovery. In the absence of structural information on either endogenous or synthetic ligands, computational chemists classically identify the very first hits by docking compound libraries to a binding site of interest, with well-known biases arising from the usage of scoring functions. We herewith propose a novel computational method tailored to ligand-free protein structures and consisting in the generation of simple cavity-based pharmacophores to which potential ligands could be aligned by the use of a smooth Gaussian function. The method, embedded in the IChem toolkit, automatically detects ligand-binding cavities, then predicts their structural druggability, and last creates a structure-based pharmacophore for predicted druggable binding sites. A companion tool (Shaper2) was designed to align ligands to cavity-derived pharmacophoric features. The proposed method is as efficient as state-of-the-art virtual screening methods (ROCS, Surflex-Dock) in both posing and virtual screening challenges. Interestingly, IChem-Shaper2 is clearly orthogonal to these latter methods in retrieving unique chemotypes from high-throughput virtual screening data.

INTRODUCTION

Computer-aided drug design[1] has become a standard tool to assist medicinal chemists in identifying and/or optimizing hits for targets of pharmaceutical interest. Corresponding computational methods are classically divided into ligand-based[2] or structure-based approaches[3] as to whether preexisting knowledge of ligands or target structures is taken into account. Among ligand-centric methods, pharmacophore searches[4] are extremely popular for many reasons: (i) the concept of pharmacophore is very intuitive and easily understandable for both computational and medicinal chemists, (ii) it does not require the *a priori* knowledge of the target's three-dimensional (3D) structure, (iii) it does not suffer from the main drawbacks[5] of structure-based approaches (e.g. inaccurate binding free energy estimates) since topological scoring functions[6] are used to rank ligand adequacy (fitness) to a pharmacophore query, (iv) aligning a ligand to a pharmacophore intuitively guides its further optimization in order to gain or lose additional features.

Typical ligand-based pharmacophore searches first require aligning template ligands sharing the same functional effect, then extract common features of these aligned ligands to derive a pharmacophore hypothesis, and last, search compound libraries for hits satisfying this hypothesis. If the X-ray structures of protein-ligand complexes are available, protein-ligand-based pharmacophores[7-10] may be derived as well by mapping features onto protein-interacting ligand atoms, and therefore complement purely ligand-based pharmacophores. However, there are still many protein structures and/or novel cavities for which not a single ligand has ever been identified. In order to avoid problems associated with structure-based approaches (e.g. absolute or relative ranking of compounds of interest, target flexibility) for such orphan targets, several methods have been proposed over the last decade to fill the gap between structure-based methods and pharmacophore searches.

Structure-based pharmacophore perception methods classically use a set of molecular probes (atoms, fragments) to locate energetically preferred probe locations. Grid-based methods (e.g. GRID,[11] SuperStar,[12] FTMap,[13] VolSite,[14] T2F,[15] GRAIL[16]) locate these preferred positions on a three-dimensional

lattice encompassing either the full protein or at least a user-defined binding cavity. Energy minima on the contour maps[17-19] are then saved for every probe and used as guides to define structure-based pharmacophoric features. Fragment-based methods rely on the prediction of hotspots from molecular dynamics simulations of the target (e.g. MCSS,[20] SILCS,[21] HSRP[22]) with multiple copies of fragments bearing well-defined pharmacophoric properties. Again, the most energetically favorable positions of every fragment are later converted into pharmacophoric features. Last, the position of pharmacophoric features can be topologically predicted by scanning the cavity-lining and accessible amino acids, in order to generate topologically ideal interaction vectors pointing at 3D space (spheres, cones) where potential ligand atoms should be located to optimally interact with the protein surface. The pioneering method LUDI[23] has inspired many structure-based pharmacophore perception methods (e.g. Virtual ligand,[24] SBP,[25] HS-Pharm,[26] Snooker,[27] Examplar[28]) to position ideal pharmacophoric moieties from the 3D structure of a binding cavity.

Whatever the method, the number of generated features (a few hundreds) exceeds by far the upper complexity tolerated by pharmacophore searching algorithms. The number of features must be therefore considerably lowered to an acceptable value, usually below 10. A preselection can be done based on energetic[15, 16, 20-22] and buriedness criteria,[15, 19] overlap with hydration sites,[22] or location with respect to knowledge-based predicted anchoring hotspots[26] to prune pharmacophoric features. Most methods finish the filtering step by a hierarchical clustering based on feature properties and inter-features distances.

Receptor-based pharmacophore searches have proven to perform at least as well as molecular docking, with respect to enrichment in true actives in retrospective virtual screening experiments.[21, 22, 26, 28] They however suffer, with a few exceptions,[21, 28] from a lack of automation since many of the above-cited post-processing steps are tedious, leaving therefore the user with subjective decisions to make with respect to e.g. the nature of probes to use, acceptable energy minima, or the number of

clusters. Moreover, the true value of receptor-based pharmacophore searches in posing a ligand has rarely been examined[29] and compared to molecular docking.

To address the above limitations, we herewith modified a previously-described cavity detection method (VolSite[14]) in order to automatize many steps between cavity detection and workable pharmacophore query definition. VolSite has notably been embedded in the IChem[30] toolkit to perform the following operations: (i) on-the-fly detection of all cavities at the surface of the target of interest, (ii) prediction of their structural druggability, (iii) perception of potential pharmacophores from the 3D structures of predicted druggable cavities. We next modified the previously reported Shaper[14] method to align ligand atoms to cavity features by shape matching and tested several topological and energy scoring functions in posing and virtual screening challenges.

COMPUTATIONAL METHODS

Datasets

sc-PDB Diverse Set: 213 diverse protein-ligand complexes (**Table S1**) were retrieved from the sc-PDB database of protein-ligand complexes[31] according to the diversity of their protein-ligand interaction patterns, measured by a previously-reported graph matching procedure (GRIM).[32] Starting from a full GRIM similarity matrix calculated on 9,283 entries of the sc-PDB archive, clusters were defined using a simple agglomerative clustering, a minimal pair-wise similarity (GrimScore) of 0.70 between its representatives, a minimal size of 6 entries, and a single linkage criterion. For every cluster, representative X-ray structures of the bound ligand and its cognate target (cluster center) were downloaded from the sc-PDB website.[33]

Astex Diverse Set: The 85 entries of the Astex Diverse Set[34] (**Table S2**) were downloaded from the CCDC website[35] and processed as follows. For each entry, the protein-ligand complex was reconstructed in SYBYL-X.2.1.1[36] by merging the ligand (mol2 file format) into the protein (mol2 file format). Bound water molecules were imported from the corresponding RCSB Protein DataBank (PDB)[37] file, all hydrogen atoms were deleted, and the fully hydrated complex (heavy atoms only) was protonated using Protoss.[38] Ions and co-factors having no heavy atoms located in a 4.5 Å-radius sphere centered on the ligand's center of mass were deleted. Water molecules were kept if two conditions were satisfied: (i) the oxygen atom was located in the above-described sphere; (ii) the bound water was engaged in at least two hydrogen bonds (donor-acceptor distance ≤ 3.5 Å, donor-hydrogen-acceptor angle ≥ 120 deg.) with the protein. The ligand, as defined in the original Astex data, and the hydrated protein (including ions and co-factors), were separately saved in mol2 file format.

DUD-E subset: 10 entries (**Table S3**), selected from a previous benchmarking study[32] and representing 5 important target families (G protein-coupled receptors, nuclear receptors, protein kinases, proteases, other enzymes) were retrieved from the DUD-E dataset[39] and further processed as reported above for the Astex Diverse Set.

ROCK2 screening set: 59,805 compounds assayed for Rho kinase 2 (ROCK2) inhibition were downloaded from the PubChem archive in 2D sd file format. Primary screening data (% of inhibition at a single concentration of 6 μM, Bioassay AID 604)[40] for all compounds and dose-response inhibitory concentrations for primary hits (IC_{50} values, Bioassay AID 644)[41] were downloaded from the PubChem bioassay repository. Compounds with IC_{50} values equal to or lower than 10 μM (n = 67) were considered active, all other compounds were considered inactive. The X-ray structure of human ROCK2 kinase in complex with inhibitor 1426382-07-1 was retrieved from the PDB (PDB identifier 4WOT) and further processed as previously reported for the Astex Diverse Set. The starting 3D coordinates of

PubChem ligands (mol2 file format) were generated with Corina v.3.4[42] and all compounds were ionized at physiological pH with Filter v.2.5.1.4.[43] The fully processed dataset comprises 59,781 compounds (67 actives and 59,714 inactives).

ESR1 screening set: 10,486 compounds assayed for estrogen receptor α (ESR1) inhibition were downloaded from the PubChem archive in 2D sd file format. Dose-response inhibitory concentrations for confirmed hits (IC_{50} values, Bioassay AID 743080)[44] were downloaded from the PubChem bioassay repository. Compounds with IC_{50} values equal to or lower than 25 μM, exhibiting full inhibition curves and devoid of Sn and P atoms (n = 59) were kept as actives. To avoid biasing the inactive set, inactive compounds were selected among molecules free of Sn and P atoms, with molecular weights in the same range (310-750) as that observed for true actives. 1,530 inactive compounds were finally selected. The X-ray structure of human estrogen receptor α in complex with the selective antagonist 4-hydroxytamoxifen was retrieved from the PDB (PDB identifier 3ERT) and further processed as previously reported for the Astex Diverse Set. The starting 3D coordinates of PubChem ligands (mol2 file format) were generated with Corina v.3.4[42] and all compounds were ionized at physiological pH with Filter v.2.5.1.4.[43] The fully processed dataset comprises 1,589 compounds (59 actives and 1,530 inactives).

OPRK1 screening set: 284,220 compounds assayed as kappa opioid receptor (OPRK1) agonists were downloaded from the PubChem archive in 2D sd file format. Dose-response activity data (EC_{50} values, Bioassay AID 1777)[45] were downloaded from the PubChem bioassay repository. Compounds with EC_{50} values equal to or lower than 20 μM (n = 35) were considered active. All other compounds were considered as inactive, out of which a randomly selected set of 64,048 compounds was retrieved. The X-ray structure of the active state-stabilized human kappa opioid receptor in complex with the full agonist MP1104 was retrieved from the PDB (PDB identifier 6B73) and further processed as previously

reported for the Astex Diverse Set. The starting 3D coordinates of PubChem ligands (mol2 file format) were generated with Corina v.3.4[42] and all compounds were ionized at physiological pH with Filter v.2.5.1.4.[43] The fully processed dataset comprises 34,083 compounds (35 actives and 34,048 inactives).

Cavity-based pharmacophore perception (IChem)

The previously described VolSite algorithm[14] was embedded in the IChem toolkit v.5.2.9[32] with small modifications compared to the original description. First, hydrogen atoms were added to the input target PDB structure using Protoss,[38] therefore optimizing the intra and intermolecular hydrogen bond network for all molecules in the input PDB file. The pharmacophoric properties of protein atoms (hydrophobic, aromatic, hydrogen-bond donor, hydrogen-bond acceptor, positive ionizable, negative ionizable, metal) were detected on-the-fly from their atom types (mol2 input) thereby enabling us to consider additional molecules (ions, cofactors, water, prosthetic groups, nucleic acids) as parts of the protein. Second, hydrophobic protein atoms were redefined using tighter rules with respect to our seminal report.[14] Hydrophobic atoms were restricted to carbon or sulfur atoms not bonded to heteroatoms or halogen atoms. Cavity-based pharmacophores were defined using a 4-step protocol (**Figure 1**) as follows:

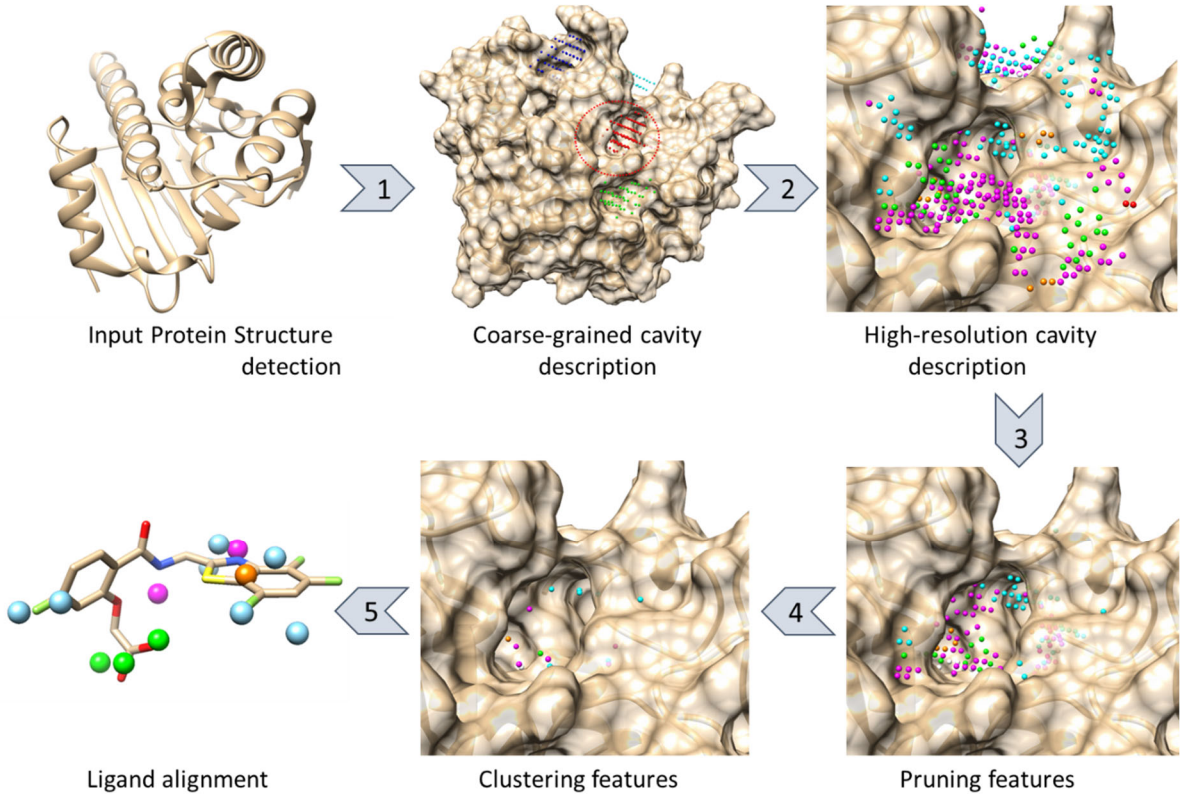

Figure 1. Overall flowchart of the method. **1)** Starting from a hydrogens-containing protein input structure, cavities were automatically detected using standard VolSite parameters and described as a collection of pharmacophoric features (blue, cyan, red and green dots); **2)** Predicted druggable cavities (enclosed by a red circle) were submitted to a second structure-based pharmacophoric description using a tighter grid resolution (1.0 Å). Pharmacophoric features (hydrophobic, cyan; aromatic, orange; hydrogen bond acceptor and negative ionizable, green; hydrogen bond donor and positive ionizable, magenta) were assigned according to the pharmacophoric properties of the nearest acceptable protein atom (see Computational methods); **3)** Pruning pharmacophoric features according to knowledge-based rules (buriedness, distance to cavity center, PLP interaction energy); **4)** Hierarchical clustering of pharmacophoric features; **5)** Shape-based alignment of ligand atoms to the cavity-based features (same color coding as in step 2) by optimizing the overlap of the corresponding molecular shapes.

Step 1 – Coarse-grained cavity detection: The general procedure for detecting cavities has already been described in a previous report[14] and will just be briefly summarized here. Starting from atomic coordinates of the target protein, a three-dimensional (3D) cube was centered on the center of mass of the target and filled with a 1.5 Å-resolution grid defining voxels with a volume of 3.375 $Å^3$ each. To every voxel was associated a site point along with a property at its center. If the corresponding voxel encompassed a protein atom or if its center was less than 2.0 Å away from any protein heavy atom, the site point would be considered inaccessible ('IN' property). Any other point was then checked for buriedness by generating, from its coordinates, a set of 120 regularly spaced 8 Å-long rays. If the number of rays intersecting an 'IN' cell (Nri) was smaller than 55, the corresponding point would be deemed outside the enclosing cavity and was assigned the 'OUT' property. Remaining points were claimed to encompass the cavity and checked for direct neighborhood with other cavity points. If isolated (less than 3 neighbors in adjacent voxels), the points were deleted. Site points closer than 4.0 Å to a protein atom were assigned one of the eight possible pharmacophoric properties (hydrophobic, aromatic, H-bond acceptor, H-bond donor, H-bond acceptor and donor, negative ionizable, positive ionizable, metal-binding) complementary to that of the closest protein atom using previously-reported interaction rules**.**[32] Points with no neighboring protein atoms within a 4 Å distance were assigned the null property (Dummy). For each detected cavity, a set of site points (mol2 file format) and a druggability score (derived from a previously-described support vector machine model)[14] were given. Only cavities with positive druggability scores were further considered for the generation of cavity-based pharmacophores.

Step 2- High-resolution cavity description: For each cavity, the previously reported procedure (Step 1) was repeated with two modifications: (i) the center of the 3D lattice was defined as the center of mass of the corresponding coarse-grained cavity, (ii) the grid resolution was then set to 1 Å for a better

description of cavity points. Each cavity point was assigned a pharmacophoric feature as previously reported.

Step 3 - Pruning pharmacophoric features: To describe the properties of true pharmacophoric features, 'ideal pharmacophores' were deduced from 213 protein-ligand complexes of the sc-PDB Diverse Set. In an ideal pharmacophore, a feature is assigned to any ligand atom in interaction with the target protein with a property equal to that of the corresponding interaction, but using exactly the same IChem rules (atom types, distances, angles, planes) as those used to define pharmacophoric properties of cavity points. An analysis of these ideal pharmacophoric features enables setting threshold values for simple descriptors (buriedness, distance to the cavity center, interaction energy) to reduce the number of pharmacophoric features without losing crucial information. Three pruning rules were applied in the following order: (i) buriedness Nri <80, (ii) distance of the feature to the cavity center < 8 Å, (iii) piecewise linear potential (PLP)[46] interaction energy < feature-dependent threshold (hydrophobic, donor, acceptor, positive ionizable, negative ionizable: 0 kcal/mol; aromatic: -2.4 kcal/mol; metal-binding: -3.5 kcal/mol).

Step 4 - Refining and clustering pharmacophoric features: The remaining features were next refined with respect to three properties: hydrogen-bond acceptor, aromatic, and hydrophobic. Because hydrogen atoms were explicitly described in the target protein, a cavity point would keep a hydrogen-bond acceptor feature only at the condition that the nearest protein atom was a hydrogen-bond donor (previous definition in steps 1 and 2) and that the donor-hydrogen-feature angle was between 120 and 180 deg. Previously-defined acceptor features not fulfilling the new angular threshold were therefore re-assigned a novel property according to the second nearest protein atom and so on until a new property could be unambiguously assigned. If it was not possible (no clear assignment possible from any of the protein atoms closer than 4 Å from the feature), the feature was simply eliminated.

Remaining aromatic features were next reconsidered from their spatial location with respect to the aromatic plane to which the closest aromatic protein atom belonged. Apart from the previously applied distance criterion (< 4 Å between feature and protein atom), we herein applied a second distance threshold of 1.5 Å corresponding to the largest possible distance between the aromatic feature and two virtual points situated 4 Å away from the closest protein aromatic ring along a normal to the aromatic plane in both directions. Again, aromatic features not satisfying this additional filter were either reassigned a new property (starting from the second closest protein atom) or eliminated if no assignment was possible. Last, remaining hydrophobic features were also reconsidered and kept hydrophobic only if: (i) more than 50% of protein atoms located within 4.5 Å of the feature were hydrophobic, (ii) at least 50% of neighboring protein residues (less than 4.5 Å away) were considered hydrophobic (alanine, valine, leucine, isoleucine, proline, methionine, phenylalanine, tyrosine, and tryptophan). Please note that these refinements were applied at step 4 and not to the full set of pharmacophoric features (step 2) to speed up the overall protocol.

Remaining features were then clustered using a simple hierarchical clustering method by pharmacophoric property and inter-feature distance (< 3.1 Å). The final pharmacophoric features were saved in three possible file formats (TRIPOS mol2 format, CATALYST chm file format,[47] LigandScout pml format[8]). The pharmacophore describes for each feature the following items:

- property: hydrophobic, aromatic, acceptor, donor, negative ionizable, positive ionizable, metal-binding;

- atomic coordinates of the feature (head);

- a 3 Å-long projection vector to a tail (acceptor, donor, aromatic features) directed to the complementary protein atom;

- special attributes for aromatic features (centroid, normal, vector, plane);

- location spheres for directional features (acceptor, donor, aromatic) of 1.6 and 2.2 Å radius for head and tail spheres, respectively;

- exclusion volumes placed, for each cavity-lining residue (one exclusion volume per residue), on the geometric center of residue heavy atoms located at a distance range of 4.1-5.0 Å from any pharmacophoric feature. The radii of exclusion spheres are dependent on the number of close protein heavy atoms (1 close atom: 1.15 Å; 2 atoms: 1.25 Å; 3 atoms: 1.35 Å; 4 atoms: 1.45 Å; 5 atoms: 1.55 Å; 6 atoms: 1.60; 7 atoms: 1.65; ≥ 8 atoms: 1.70 Å).

Please note that features having the double property hydrogen-bond donor and hydrogen-bond acceptor were described by two separate properties (donor, acceptor) matched on the same point.

Ligand alignment to IChem pharmacophoric features (Shaper2)

The previously described Shaper algorithm,[14] designed to align cavities, was slightly modified to align ligand atoms (mol2 file format) onto the above-described set of cavity points. Shaper2 relies on OpenEye python libraries[43] to describe molecular shapes by a smooth Gaussian function and to align two molecular objects (ligand features, cavity features) by optimizing the intersection of their corresponding volumes.[48] During the alignment, cavity features are kept rigid while a maximum of 200 pre-defined conformers of the ligand to fit (fit object), constructed in Omega2 v.2.5.1.4[43, 49] undergo rigid body rotations and translations. Contrary to the original Shaper method, the updated version now allows the user to choose among different overlap methods (by default: Exact), different overlap minimization techniques (by default: Subrocs) and diverse similarity metrics (by default: TanimotoCombo). A detailed description of all options is available online.[50]

A specific force field (**Table S4**) has been set up to align ligand atoms to cavity features. It consists in SMARTS patterns for 9 pharmacophoric feature properties (hydrophobic, ring, donor, acceptor, donor and acceptor, cation, anion, Ca_Mg, Zn) and 56 pattern-matching rules to score the shape-based

alignment by pharmacophoric similarity (**Table S4**). All aligned poses were then subjected to a two-step structure optimization process using the MMFF94 force field[51] implemented in SZYBKI v. 1.8.0.1.[43] In the first step, each pose was minimized with the steepest descent algorithm with respect to the MMFF94 potential in full Cartesian coordinates using default settings. In the second step, a single point calculation was done with the Poisson-Boltzmann (PB) protein-ligand electrostatics,[52] calculating protein-ligand interaction energy including solvent effects.

All possible ligand-cavity matches were scored according to four metrics:

1) TanimotoCombo similarity score as follows:

$$TanimotoCombo = ShapeTanimoto + ColorTanimoto = \frac{OS_{C,L}}{IS_C + IS_L + OS_{C,L}} + \frac{OC_{C,L}}{IC_C + IC_L + OC_{C,L}}$$

where $OS_{C,L}$ is the overlap between shapes of cavity and ligand features, IS_C and IS_L are the non-overlapped shapes of each entity, $OC_{C,L}$ is the overlap between colors of cavity and ligand features, IC_C and IC_L are the non-overlapped colors of each entity. The metric is asymmetric and varies between 0 and 2.

2) The PLP interaction of each feature with the protein, as implemented in the original publication[46]

3) MMFF94 total energy: TotE= TotIE + IntE

TotIE (ligand MMFF intramolecular energy) = E_{VdW} + $E_{Coulomb}$ + E_{Bond} + E_{bend} + $E_{StretchBend}$ + $E_{Torsion}$ + $E_{Improper\_Torsion}$

IntE (protein-ligand interaction energy) = $E_{VdW\text{-}PL}$ + $E_{Coulomb\text{-}PL}$ + $E_{Protein\_desolv\_PB\text{-}PL}$ + $E_{Ligand\_desolv\_PB\text{-}PL}$ + $E_{Solvent\_screening\_PB\text{-}PL}$.

4) MMFF94 protein-ligand interaction energy IntE

For more details, the reader is directed to the SZYBKI implementation of the MMFF94 force field.[53]

Ligand alignment to IChem pharmacophores (Discovery Studio)

The input ligand 3D structure was converted from mol2 to sd file format using Corina v.3.4[42] and used as input to generate a set of 3D conformers using the 'Generate Conformations' protocol of Discovery Studio v.2017.[54] The conformer generation method was set to 'FAST', a maximum of 200 conformers were generated within an energy threshold of 20 kcal/mol with respect to the global minimum. Ligand conformers were next aligned to IChem pharmacophoric features (chm format) using the 'citest' online command of Discovery Studio. A maximum of 2,000 pharmacophores were generated with a minimum of 2 and a maximum of 6 features, to map ligand conformers in rigid mode. The best mapping conformer (highest fit value) was finally saved in sd file format.

Ligand alignment to IChem pharmacophoric features (LigandScout)

Ligands (sd file format) were converted to the internal LigandScout[55] v.4.1.10 ldb database format thanks to the idbgen script saving, for each ligand, up to 200 conformations using high quality settings of the iCon[56] conformer generator (icon-best option). The conformations were next aligned, with standard settings of the iscreen routine, to IChem-generated pharmacophores in LigandScout pml format. The best mapping conformer (highest fit value) was finally saved in sd file format

Docking (Surflex-Dock)

Surflex-Dock v.4.227[57] was used as prototypical docking engine. A protomol[57] was first generated from the list of residues, ions, cofactors and water molecules lining the ligand-binding site (any molecule with a heavy atom in a 4.5 Å-radius sphere centered on the ligand's center of mass) using default settings. The protomol was further used to dock a randomly generated conformation of the ligand using the –pgeom option of Surflex-Dock. Only the best-ranked pose (scored by pkd value) was saved.

ROCS shape overlap

A maximal number of 200 conformers (sd file format) were generated for every PubChem ligand using standard settings of Omega2 v.2.5.1.4.[43, 49] Every conformer was then compared to the query (protein-bound ligand X-ray pose, mol2 file format) with ROCS v. 3.2.0.4[43, 58] and the best matching conformers for every ligand were scored by decreasing TanimotoCombo value.

RESULTS AND DISCUSSION

The pharmacophore concept is more than one century old[59] and has been widely used in ligand-based[4] and more recently in protein-ligand-based[7, 8] virtual screening. When only structures of ligand-free proteins are available, defining simple and workable pharmacophore queries is more difficult for the simple reason that cavity structure-based pharmacophore perception is a complex and multi-step procedure. Cavities first need to be detected at the protein surface and evaluated for their potential druggability. The positions of pharmacophoric features mimicking a perfect ligand must then be inferred from the coordinates of cavity-lining protein residues. Very often, the number of ideal pharmacophoric features exceeds by far the upper complexity tolerated by standard 3D pharmacophore searches. Therefore, cavity features need to be pruned on a rational basis, usually from interaction energy maps, to downsize the number of features and enable the definition of a workable pharmacophore query (< 10 features). Moreover, many methods[20-22] rely on lengthy molecular dynamic simulations to locate the energetically preferred positions of probes, which prohibits their usage even at a low throughput. Although recent efforts have been reported to simplify the above described steps,[21, 28] there is still a need for a tool that is able to quickly and reliably automatize the entire process from early cavity detection to late final pharmacophore definition.

Cavity-based pharmacophore perception

The herein proposed cavity-based pharmacophore perception workflow is made of four consecutive steps (**Figure 1**). First, potentially druggable cavities were detected on-the-fly from the input protein structure using standard parameters of our in-house developed VolSite algorithm.[14] The method centers the protein in a 1.5 Å resolution lattice and assigns a pharmacophoric feature (hydrophobic, aromatic, hydrogen bond donor, hydrogen bond acceptor, negatively-charged, positively-charged, metal-binding) to every accessible voxel, depending on the pharmacophoric property of the nearest accessible protein atom. The structural druggability of every detected cavity was predicted thanks to a support vector machine model[14] showing a very good accuracy in comparison to state-of-the-art methods. For each cavity, the detection procedure was repeated using a tighter grid resolution (1.0 Å) centered on its center of mass. In a third step, the obtained cavity features were pruned in order to decrease their number.

The previously published VolSite algorithm[14] was modified to take into account the positions of explicit hydrogen atoms, added by the Protoss knowledge-based method.[38] The main advantage of using hydrogen coordinates of the target protein is that hydrogen acceptor features can be better assigned from the corresponding vectors (donor-hydrogen-voxel center) than using the previous protocol that just relied on distances. Along the same spirit, we have also refined the definition of cavity aromatic features by taking into account additional topological measurements for detecting face-to-face aromatic interactions (see Computational methods). Last, the assignment of hydrophobic features is stricter and now requires that the closest protein atom is also annotated as hydrophobic and that it lies in a global hydrophobic environment. The consequence of these changes is that the pharmacophoric assignment of cavity features may require several steps. For example, a hydrophobic protein atom (e.g. CB atom of an alanine) cannot be used to assign a hydrophobic property to a cavity voxel if the latter does not satisfy some above described proximity conditions, even if it is the closest protein atom of that particular voxel. In that case, a second assignment step is done by considering the second closest protein atom to the voxel, and so on until one protein atom perfectly suits all required

conditions. Therefore, contrary to the original VolSite implementation,[14] some cavity voxels may not be assigned a pharmacophoric property in the updated version.

A key issue in the current work is the implementation of knowledge-based rules to limit the number of pharmacophoric features to the lowest possible number. To reach this objective, we carefully analyzed the position of 'ideal' pharmacophoric features derived from a training set of 213 diverse protein-ligand structures. By ideal, we mean that pharmacophoric features are directly mapped onto protein-bound ligand atoms if the corresponding atom is in direct interaction, according to IChem rules, with the protein. To define a set of ideal features, 213 high-resolution protein-ligand X-ray structures were extracted from the sc-PDB archive of druggable protein-ligand complexes.[31] These structures present a maximal diversity of protein-ligand interaction patterns, as assessed by our previously described GRIM methodology[32] that directly computes pairwise similarity of protein-ligand interaction patterns. Out of the 213 most diverse complexes, we could identify 4,871 ideal features for which three properties were inspected: buriedness, distance to cavity center, PLP interaction energy (**Figure 2**).

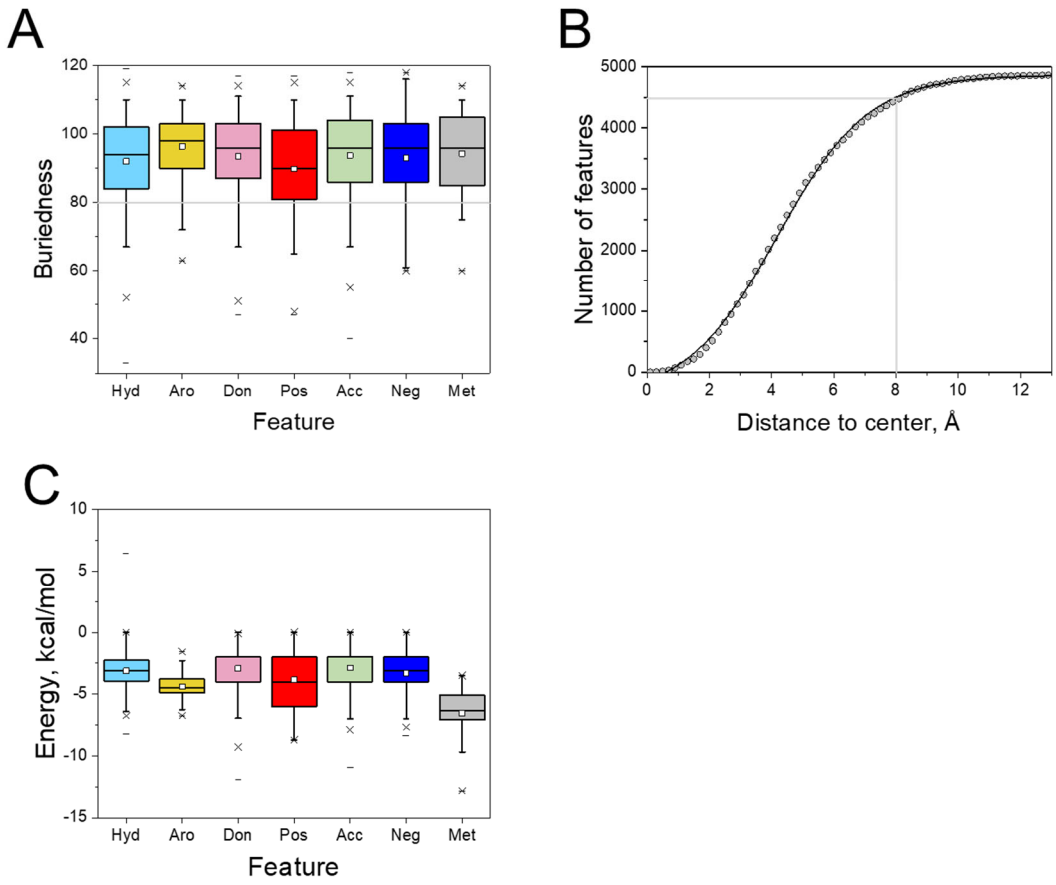

Figure 2. Properties of 4871 ideal pharmacophoric features generated from the sc-PDB Diverse Set (213 complexes); **A)** Box-and-whisker plot of the distribution of the buriedness of pharmacophoric features (Hyd, hydrophobic; Aro, aromatic; Don, hydrogen-bond donor; Pos, positive ionizable; Acc, hydrogen bond acceptor; Neg, negative ionizable; Met, metal-binding) expressed by the number of 8 Å-long rays (out of a total of 120) originating from the feature center and intersecting protein atoms. The boxes delimit the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box, respectively. Crosses delimit the 1% and 99th percentiles. Minimum and maximum values are indicated by a dash; **B)** Distance of the feature (in Å) to the cavity center, expressed by the cumulative number of features. The cumulative distribution follows a Boltzmann sigmoidal function (R^2 = 0.999); **C)** Box-and-whisker plot of the distribution of the PLP[46] interaction energy between features (Hyd, hydrophobic; Aro, aromatic; Don, hydrogen-bond donor; Pos, positive ionizable; Acc, hydrogen bond

acceptor; Neg, negative ionizable; Met, metal-binding) and their protein environment. The boxes delimit the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box, respectively. Crosses delimit the 1% and 99th percentiles. Minimum and maximum values are indicated by a dash.

Whatever the feature type, more than 75% of ideal features had a buriedness value higher than 80 (**Figure 2A**). Likewise, over 90% of all these features were closer than 8 Å from the corresponding cavity center (**Figure 2B**). Last, recording the PLP interaction energy of each feature with its protein environment clearly indicated, as to be expected, that such interaction energies are negative and feature type-dependent (**Figure 2C**). Applying feature-dependent cut-off values (0 kcal/mol for hydrophobic, donor, acceptor, donor and acceptor, positive ionizable, negative ionizable features; -2.4 kcal/mol for aromatic features, -3.5 kcal/mol for metal-binding features) ensured that at least 95% of these ideal features would be selected.

The application of the above-described pruning rules all along the flowchart (**Figure 3A**) indeed limited the number of output features from 326 ± 90 at the beginning of the process (fine-grained cavity description) to 259 ± 95 after buriedness evaluation, 253 ± 88 after cavity center-feature distance calculation, 37 ± 7 after clustering, and finally 27 ± 7 after PLP interaction energy calculation (**Figure 3B**). The chronological order in applying these three filters does not affect the obtained results. To avoid repeating the PLP interaction energy evaluation before and after clustering, we decided to place this step at the end of the protocol. Here again, we verified that this choice did not bias the obtained results.

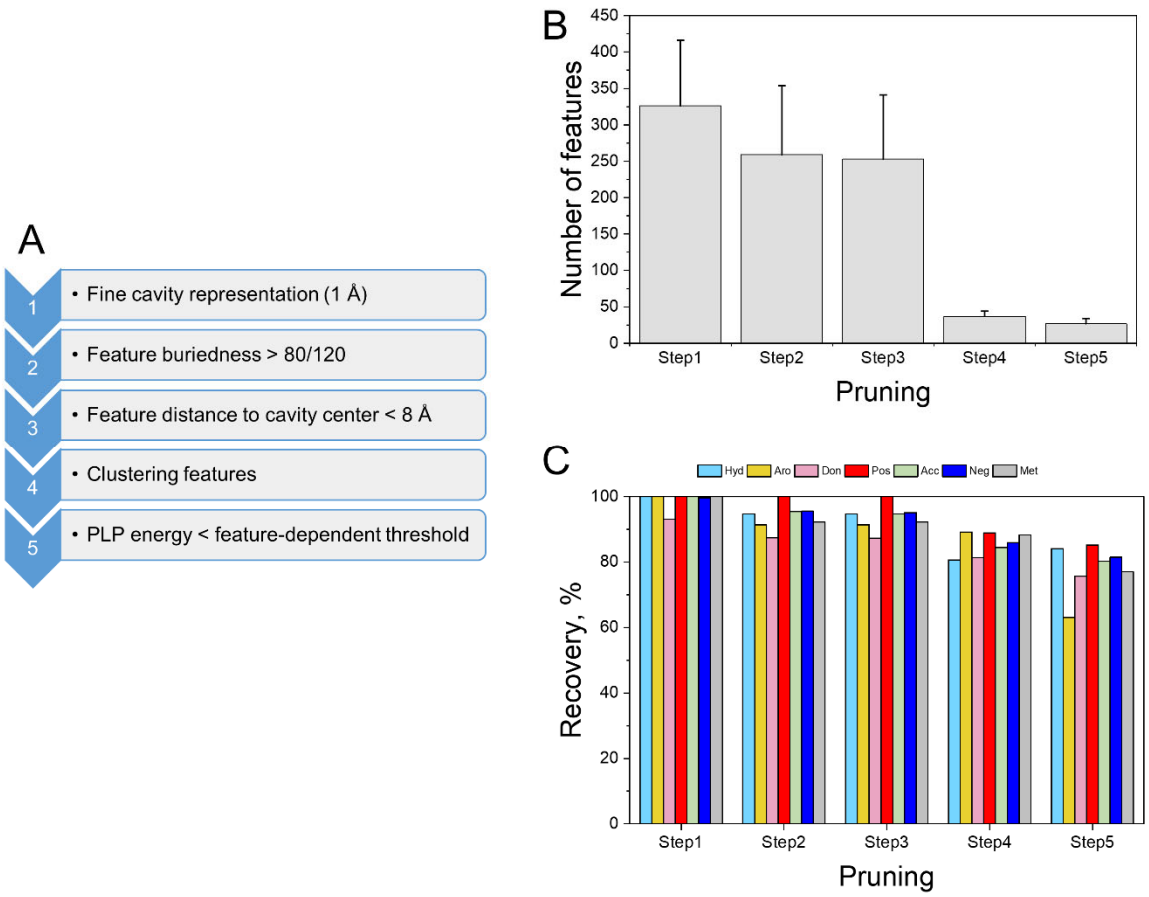

Figure 3. 5-step protocol to prune cavity-based pharmacophoric features in IChem. Features were defined from the IChem detected ligand binding-site of the 213 entries of sc-PDB Diverse Set. **A)** Flowchart; **B)** Decreasing number of pharmacophoric features all along the protocol; **C)** Percentage of recovery of ideal features all along the protocol. A predicted feature was defined as recovered if closer than 2.0 Å from an ideal feature of the same type, generated for the same test set and identical topological rules by matching pharmacophoric properties to protein-interacting ligand atoms.

Of course, we verified that the observed drastic reduction of the number of features did not lead to a global loss of information. For that purpose, we estimated the percentage of recovery of ideal features. For every IChem feature, we computed the closest distance between an ideal feature and an IChem predicted feature of compatible pharmacophoric type. If the distance is less than 2.0 Å, we considered that the predicted feature is close enough to the ideal one and that the latter is recovered. Estimating

the percentage of recovery of ideal features at every step of the pruning stage (**Figure 3C**) indicated that the filtering process did not discard a significant proportion of key features. After the last filtering step, about 80% of all feature types (except for aromatic features for which the recovery rate was about 70%) were indeed within 2 Å of a predicted feature of the same type. We therefore assume that our feature selection process is accurate enough to simplify the final cavity-based pharmacophore without any major loss of information.

Ligand posing accuracy

Ligands were aligned onto the above-described cavity-based pharmacophoric features using a modified version (Shaper2) of our Shaper algorithm[14] that uses a smooth Gaussian function to maximize the shape overlap of ligand atoms and cavity features, and score the alignment by both shape and color (feature type) similarity. With respect to the previous Shaper version that was designed for pairwise cavity comparisons, we modified the force field (**Table S4**) to enable aligning ligands to cavity features. A test set of 85 high-quality protein-ligand complexes (Astex Diverse Set),[34] specifically designed for assessing docking performance, was used for that purpose. To estimate the posing quality, we compared the results obtained with Shaper2 alignment to IChem features (this work) to those of a state-of-the-art docking tool (Surflex-Dock).[57] Moreover, we also compared the alignment accuracy of Shaper2 to that of two standard pharmacophore search methods (Discovery Studio, LigandScout), using the same set of IChem-derived features. Four scoring functions were evaluated to analyze Shaper2 matching poses to IChem pharmacophores. The first one (Tc) just computes the TanimotoCombo similarity (shape + color) between the aligned poses and the protein-bound ligand X-ray coordinates. The second one (PLP) computes the PLP interaction energy of the feature with its protein environment. The third and fourth ones (TotE, IntE) register the MFF94 total interaction energy and MMFF94 protein-ligand interaction energy using a Poisson-Boltzmann treatment of desolvation effects.

Plotting, for each Astex Diverse Set entry, the root-mean square deviation (rmsd) of the best Surflex-Dock pose (heavy atoms only) to the true X-ray pose, defines the base line for applying a structure-based docking tool to this dataset (**Figure 4**).

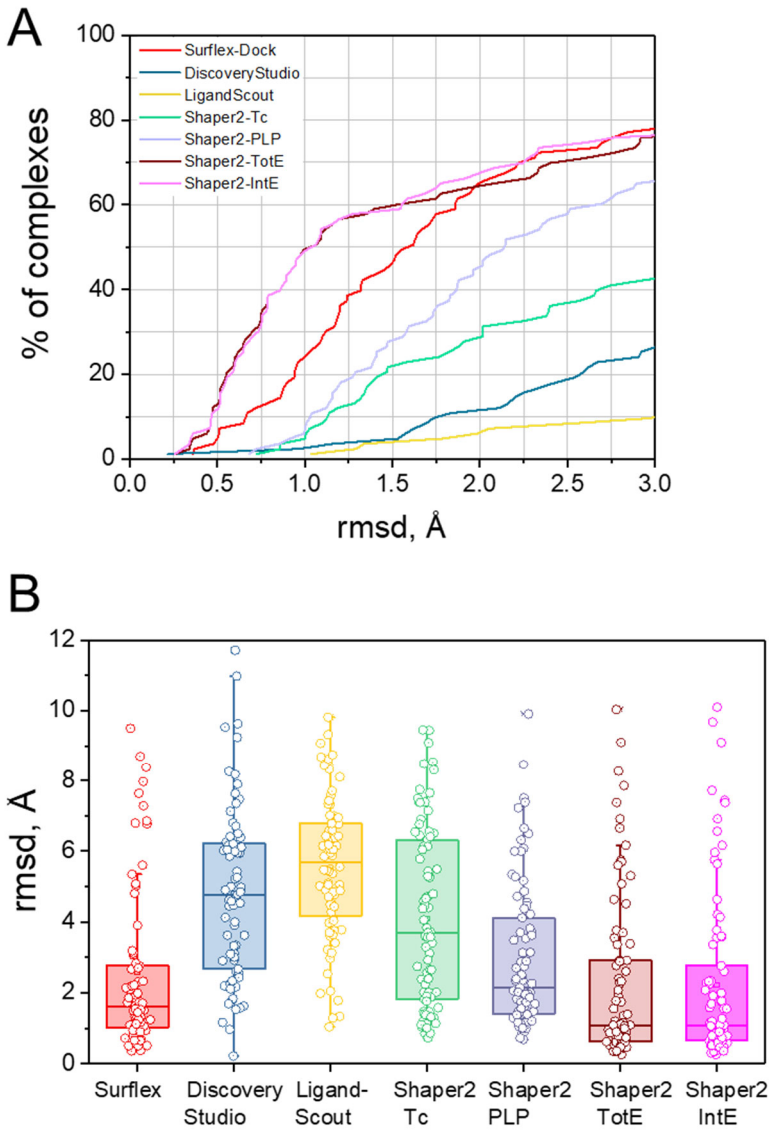

Figure 4. Performance of different methods in predicting the bound pose of 85 ligands from the Astex Diverse Set. Posing was done using docking (Surflex-Dock), ligand-based pharmacophore (DiscoveryStudio, LigandScout), and cavity-based pharmacophore (IChem) searches. IChem alignments were scored by four different scoring functions: Tc, TanimotoCombo similarity; PLP, PLP interaction energy in kcal/mol; TotE, total MMFF94 energy in kcal/mol; IntE, MMFF94 protein-ligand interaction

energy. **A)** Cumulative percentage of entries form the Astex Diverse Set, for which the top-ranked pose of the cognate ligand is within a certain rmsd to the X-ray pose. **B)** Distribution of rmsd values to the X-ray pose. The boxes delimit the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box, respectively. Crosses delimit the 1% and 99th percentiles. Minimum and maximum values are indicated by a dash.

Surflex-Dock indeed posed quite accurately the Astex ligands with a median rmsd of 1.62 Å. 65% of all ligands were docked with rmsd values to the X-ray pose below 2 Å (**Table 1**). This docking performance is quite similar to previous results obtained on this peculiar dataset[60] and on other sets by us[61] and other groups.[5, 62] We can therefore assess that no particular bias is present in both the dataset and the manner we set the input files. In our hands, the two ligand-based pharmacophore tools (Discovery Studio, LigandScout) failed in predicting a correct pose (rmsd < 2.0 Å) in ca. 90% of the cases (**Figure 4, Table 1**). In other words, the complexity of IChem cavity-based features (28 features on average for the Astex Diverse Set) is still too important for hard sphere-based alignment tools. The quality of IChem cavity-based pharmacophores is not responsible for this observation since Shaper2 alignments to the same pharmacophores produced much better results, albeit with significant differences with respect to the chosen scoring function (**Figure 4, Table 1**). Just relying on the similarity of shapes and colors (Tc metric) was not sufficient to yield high-quality poses (average rmsd = 4.10 Å) although obtained results were already better than that observed with Discovery Studio and LigandScout. Rescoring Shaper2 poses according to the PLP energy potential improved significantly the alignment (average rmsd = 2.95 Å, **Table 1**). This scoring method remains inferior to Surflex-Dock in producing high-quality poses (**Figure 4**).

Table 1. Posing accuracy of docking (Surflex-Dock), ligand-based pharmacophore (Discovery Studio, LigandScout), and receptor-based pharmacophore (IChem) searches, applied to 85 protein-ligand complexes from the Astex Diverse Set.

Program	Average rmsd, Å[a]	Median rmsd, Å[b]	% entries rmsd <1 Å	%entries rmsd <2 Å
Surflex-Dock[c]	2.57	1.62	24	65
Discovery Studio[d]	4.80	4.77	3	12
LigandScout[e]	5.53	5.70	0	6
Shaper2-Tc[f]	4.10	3.70	4	28
Shaper2-PLP[g]	2.95	2.14	6	45
Shaper2-TotE[h]	2.23	1.06	49	64
Shaper2-IntE[i]	2.22	1.06	48	67

[a] average rms deviation (heavy atoms) to the ligand X-ray pose

[b] median rms deviation (heavy atoms) to the ligand X-ray pose

[c] Surflex-Dock pose with the lowest internal score (pkd)

[d] Discovery Studio pose with the highest fit score

[e] LigandScout pose with the highest fit score

[f] Shaper2 pose with the highest TanimotoCombo score

[g] Shaper2 pose with the lowest PLP interaction energy

[h] Shaper2 pose with the lowest MMFF94 total energy

[l] Shaper2 pose with the lowest MMFF94 ligand-protein interaction energy

We therefore minimized the pose (ligand in its protein environment) with the MMFF94 force field including an explicit Poisson-Boltzmann treatment of desolvation effects.[52] Using either the total MMFF94 energy (TotE: ligand strain energy + protein-ligand interaction energy) or just the protein-ligand interaction energy term (IntE) yielded very accurate poses (identical median rmsd to the X-ray pose of 1.06 Å). Interestingly, although the fraction of high-quality poses (rmsd < 2.0 Å) was almost identical to that obtained with Surflex-Dock (ca. 65%), the latter two scoring functions were much more efficient in producing very high-quality poses (rmsd to the X-ray pose < 1.0 Å; **Figure 1**).

Altogether, Shaper2 alignment to IChem cavity-based pharmacophores is therefore competitive with a standard docking tool with respect to posing accuracy. The competitive advantage of a Gaussian function (Shaper2) in comparison to either the Kabsch algorithm[63] (Discovery Studio) or the Hungarian matcher[64] (LigandScout) appears quite significant for the complexity of pharmacophore queries (27 features on average) produced by our method. However, the scoring function used to rank Shaper2 poses is very important. Energy-based scoring functions are preferred to faster shape/color overlap estimations. Moreover, an explicit treatment of desolvation effects yields a very accurate pose ranking, albeit at the cost of an extra computational demand (ca. 5 sec per pose)

Virtual screening accuracy (DUD-E set)

In the next challenge, we probed the accuracy of Shaper2 alignment to IChem cavity-based pharmacophores to discriminate between true actives and chemically similar decoys for a set of ten DUD-E targets[32, 39] (**Table S3**). Although results obtained on such benchmarks are not fully predictive of real-life prospective virtual screening studies,[65] we still wanted to compare our approach to Surflex-Dock in this exercise. Ten targets were selected to span major target families (G protein-coupled

receptors, kinases, nuclear hormone receptors, proteases, globular enzymes) and caution was given to discard easy test cases (targets leading to areas under the ROC curve above 0.85) as suggested by the seminal paper.[39] The chosen subset is believed to be rather difficult for docking (DUD-E authors used the Dock3.6 docking program as screening engine) with an average AUC value of 0.66, well below the mean AUC value (0.76) observed for the entire DUD-E dataset.[39] Results obtained with Surflex-Dock generally confirmed the previous report with a mean AUC value of 0.73 (**Table 2**). For two targets (GCR, FGFR1), the observed ROC AUC values were statistically better than random selection but still below 0.70, therefore indicating just a fair performance. Shaper2 alignment to IChem pharmacophores scored by the PLP potential led to a poor performance in this challenge (mean AUC value of 0.57; **Table 2**). Conversely to the above-described challenge, scoring matching poses by either MMFF94 protein-ligand interaction energy or MMFF94 total energy marginally enhanced the virtual screening accuracy of the method (mean AUC values of 0.62 and 0.65, respectively; **Table 2**) despite significant ameliorations (AUC ≥ 0.70) for five out of the ten targets (ADRB2, GCR, ACE, FGFR1, AKT1), using the MMFF total energy as a scoring function. Given that the MMFF94 total energy led to the best performance, we tried to decouple the scoring function used to select the best poses from that utilized to sort compounds. The best combination was obtained by selecting the poses by MMFF94 total energy and sorting compounds (actives and decoys) by PLP energy (**Table 2**). Using this approach, a mean AUC value of 0.68 comparable to that observed with the docking program Dock3.6, was obtained. The performance was excellent for two targets (ADRB2, RENI; AUC > 0.80), good for two other entries (FGFR1, AKT1; 0.70 < ROC AUC < 0.80), fair for four targets (AA2AR, GCR, ADA, ACE; AUC ≥ 0.57) and remained poor but still better than random picking for two entries (ANDR, PGH2). Despite the small sample size, the distribution of ROC values observed from the three Shaper2 protocols with MMFF94 refinement (IntE, TotE, TotE+PLP) is statistically different from that seen when only PLP energy was taken into account in a two-sample t-test assuming either equal or unequal variance at a confidence interval of 95% ($p < 0.05$). The differences observed with respect to each pair of the refinement protocols are however statistically not significant in the same test. Compared to Surflex-Dock, the

mixed approach gave a better performance for three targets (ADRB2, ANDR, FGFR1), a rather similar accuracy for three entries (GCR, RENI, AKT1), and gave a poorer performance in four entries (AA2AR, ADA, PGH2, ACE; **Table 2**, **Figure 5**).

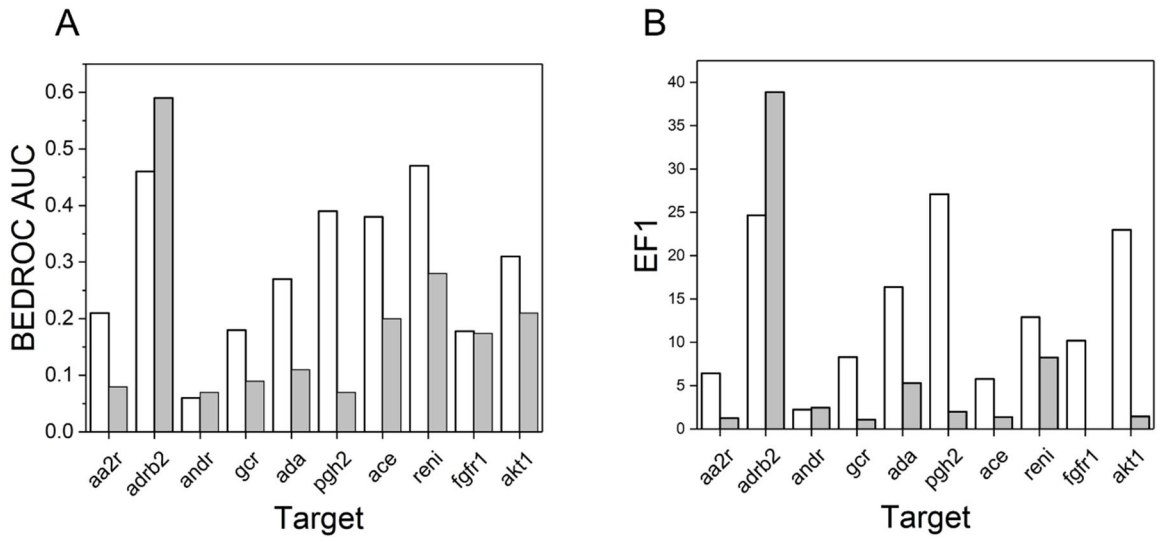

Figure 5. Virtual screening performance of Surflex-Dock (white bars) and Shaper2 (gray bars) on 10 entries of the DUD-E set.[32] Shaper2 alignments to IChem cavity-based pharmacophores were scored by MFF94 total energy, whereas DUD-E compounds were ranked by increasing PLP interaction energy. **A)** Area under the BEDROC curve (α = 20), **B)** Enrichment in true actives at a constant 1% false positive rate.

We must acknowledge that we have no clear explanation on the positive role of PLP rescoring on poses selected by MMFF94 total energy. We could not either explain successes and failures of the approach with respect to target and/or ligand properties. To account for early enrichment in true actives, the area under the Boltzmann-enhanced discrimination of the ROC (BEDROC) curve, as well as the enrichment in true actives at 1% decoys retrieval, were also computed for each of the entries (**Figure 5**). Disappointingly, BEDROC curves clearly show that our method was inferior to Surflex-Dock in early enrichment in true actives for seven out of the ten cases.

Table 2. Area under the ROC plot of a binary classification (actives, decoys) of DUD-E ligand poses to the X-ray structure of 10 representative targets.[32]

Posing	Dock3.6[a]	SF-Dock	Shaper2	Shaper2	Shaper2	Shaper2
Conformer selection	Dock3.6	SF-Dock	PLP	IntE	TotE	TotE
Scoring	Dock3.6	Sf-Dock	PLP	IntE	TotE	PLP
G Protein-Coupled receptors						
Adenosine A2A receptor (AA2AR)	0.83	0.74	0.57	0.61	0.56	0.58
Beta2 adrenergic receptor (ADRB2)	0.76	0.85	0.51	0.61	0.71	0.96
Nuclear hormone receptors						
Androgen receptor (ANDR)	0.51	0.47	0.56	0.52	0.59	0.54
Glucocorticoid receptor (GCR)	0.44	0.56	0.56	0.64	0.73	0.57
Other enzymes						
Adenosine deaminase (ADA)	0.76	0.83	0.60	0.56	0.53	0.63
Prostaglandin G/H synthase 2 (PGH2)	0.62	0.76	0.57	0.62	0.54	0.55
Proteases						

Angiotensin-converting enzyme (ACE)	0.72	0.84	0.58	0.60	0.75	0.64
Renin (RENI)	0.66	0.88	0.56	0.68	0.66	0.82
Protein kinases						
Fibroblast growth factor receptor 1 (FGFR1)	0.73	0.67	0.60	0.78	0.70	0.76
RAC-alpha protein kinase (AKT1)	0.72	0.76	0.57	0.61	0.72	0.74
Mean ROC area under the curve	0.67	0.73	0.57	0.62	0.65	0.68

[a] Report from the original paper describing the DUD-E dataset[39]

Virtual screening accuracy (PubChem bioassays)

The real value of DUD-E ligands for providing virtual screening guidelines is currently debated because of severe ligand and target-based drawbacks in selecting decoys.[65] Most docking tools were notably reported to overestimate the real discrimination between actives and decoys, for the simple reason that DUD-E actives tend to be chemically similar to the co-crystallized ligand in the target 3D structure selected for docking.[65]

We therefore challenged our method with respect to true experimental screening data from the PubChem bioassay archive[66] in which both true active and true inactive compounds have been explicitly defined from a single *in vitro* assay. Three targets of pharmaceutical importance (one kinase, one nuclear hormone receptor, one G protein-coupled receptor) for which both high-quality screening data (primary assay, confirmatory dose-response assay) and 3D structural information (ligand-bound high resolution X-ray structure) are available were selected as test cases (**Table 3**).

Virtual screening was done using one ligand-based method (ROCS),[58] and two structure-based methods (Surflex-Dock, Shaper2). The virtual screening accuracy was simply estimated from the number of true actives ranked among the top 1% and top 5% scorers, respectively. The experimentally determined hit rate is low (ca. 0.1%) for two screens (ROCK2, OPR1) and much higher (3.71%) for the ESR1 challenge. Activity data ranged from low nanomolar to two-digit micromolar values, the ESR1 dataset being the most enriched in low nanomolar ligands (**Table 3**). The latter screening data should therefore be easier to predict, an assumption that is confirmed by analyzing the results of the ROCS screening for which spectacular enrichments over random picking were already observed when considering the top 1% ESR1 ligands (**Table 3**). This means that the true actives in this dataset are similar in shape and pharmacophoric properties to the reference ligand (4-hydroxytamoxifen) co-crystallized in the protein structure used for the structure-based approaches.

Table 3. Virtual screening of PubChem bioassay data.

Target	Rho kinase 2		Estrogen receptor α		Kappa opioid receptor	
Encoding gene	ROCK2		ESR1		OPRK1	
PubChem bioassay AID	604,644		743080		1777	
Number of actives	67		59		35	
Number of inactives	59,714		1,530		34,048	
Activity range, μM	0.03-9.78		0.03-9.69		0.06-18.1	
Hite rate, %	0.11		3.71		0.10	
Virtual screening[a]	Top1%	Top5%	Top1%	Top5%	Top1%	Top5%
ROCS[b]	2(3)	3 (0.9)	11 (18.5)	11 (3.7)	1 (2.9)	1 (0.6)
Surflex-Dock[c]	1 (1.5)	2 (0.6)	1 (1.7)	6 (2.0)	3 (8.8)	4 (2.3)
Shaper2[d]	1 (1.5)	2 (0.6)	2 (3.4)	18 (6.1)	1 (2.9)	6 (3.5)

[a]number of true actives among the top 1% and top 5% scoring molecules. Numbers in brackets indicate the observed enrichment over random picking

[b]ranked by TanimotoCombo similarity to ROCK2-bound inhibitor (ligand ID 3SG, PDB ID 4WOT), ESR1-bound antagonist (ligand ID OHT, PDB ID 3ERT), and OPRK1-bound agonist (ligand ID CVV, PDB ID 6B73).

[c] ranked by pkd (Surflex score)

[d] ranked by PLP energy after MMFF94 energy minimization

For the two other targets (ROCK2, OPRK1), ROCS screening performed three times better than random selection when the top 1% scorers were considered, the enrichment logically decreased when selecting more compounds from the screen with a performance equal or even inferior to random picking when the top 5% scoring compounds were considered (**Table 3**). In other words, two screening sets (ROCK2, OPRK1) were considered difficult for structure-based approaches whereas the third one (ESR1) should be much easier.

Surflex-Dock and Shaper2 gave identical results when considering the top 1% scorers of the ROCK2 screen, however inferior to that observed with ROCS (**Table 3**). Considering a higher percentage of top scoring compounds (5%) allowed retrieving one additional active, but at the cost of a lower hit rate. For the easier ESR1 test case, Shaper2 gave much better results than Surflex-Dock whatever the fraction considered to qualify virtual hits. Enrichment factors over random picking of 3.4 and 6.1 were observed for the top 1% and top 5% scoring molecules, respectively (**Table 3**). Noteworthy, Shaper2 continued to retrieve novel actives upon increasing the number of selected virtual hits, and even outperformed ROCS when considering the top 5% scoring hits. For the last dataset (OPRK1), both Surflex-Dock and Shaper2 gave statistically relevant enrichment over random picking (8.8 and 2.9 at 1% top scorers, 2.3 and 3.5 at top 5% scorers). Docking performed better than cavity-based pharmacophore search in the initial enrichment. However, Shaper2 retrieved more actives than Surflex-Dock among the top 5% scorers (**Table 3**).

In agreement with many previous studies,[67-69] we observed that the three virtual screening methods used in this study tend to retrieve different true actives and most importantly different chemotypes (**Figure 6**). In all screens, Shaper2 was able to identify true actives (one ROCK2 inhibitor, seven ESR1 antagonists, four OPRK1 agonists, **Figure 6**) not found by any other method. If one restricts the analysis to the retrieval of unique scaffolds, Shaper2 was the method producing the highest number of uniquely retrieved chemotypes (**Figure 6**), thereby demonstrating its utility and orthogonality to other virtual screening methods.

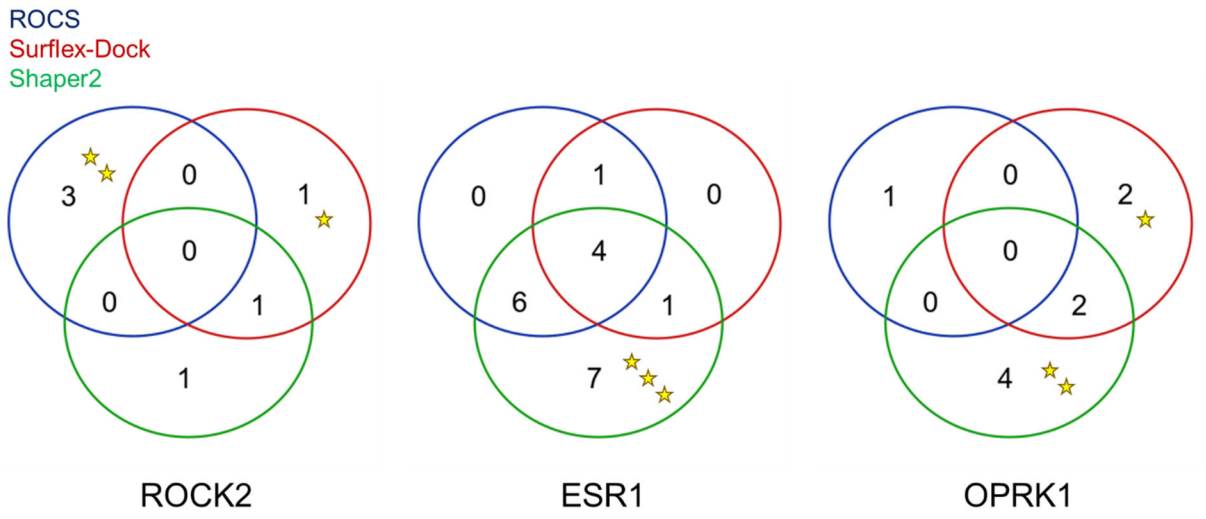

Figure 6. Orthogonality of three virtual screening methods (ROCS, Surflex-Dock, Shaper2) in retrieving true actives among the top 5% ranking hits, from three PubChem bioassay high-throughput screens (ROCK2 inhibitors, PubChem bioassay AID 644; ESR1 antagonists, PubChem bioassay AID 743080; OPRK1 agonists, PubChem bioassay AID 1777). The numbers of true actives recovered by each method are displayed by Venn diagrams[70] highlighting molecules uniquely found by a single method or common to two or three hit lists. Each chemotype retrieved by a single method is highlighted by a star.

The motivations for retrieving the top 5% scorers were two-fold: (i) Since we were really mining HTS data with very few high affinity ligands, the number of hits retrieved among the top 1% scorers was low (even for the ligand-based ROCS shape matching method). We therefore increased the threshold to select the top 5% scoring molecules in order to begin to see statistically meaningful differences between the screening methods; (ii) retrieving a higher proportion of virtual hits enabled us to cluster them by scaffolds (maximum common substructures) and pick a more representative set of hits for experimental validation (in terms of scaffold coverage) than a strategy based on a harder cut-off (say, pick the top 100 scoring compounds). Of course, no definitive conclusion can be drawn from the present benchmarking exercise focusing on three independent HTS data. However, it appears that

Shaper2 alignment to IChem cavity-based pharmacophores is at least as efficient as other virtual screening methods (shape alignment, docking) when applied to three test cases for which the entire screening results were known. The good performance of Shaper2 in true virtual screening benchmarks is in contradiction with the previously reported poorer performance observed with artificially reconstituted DUD-E training sets for which severe target and ligand-bias have been noticed.[65] We therefore recommend benchmarking virtual screening methods with true experimentally determined high-throughput screening data. Fortunately, the PubChem bioassay repository[66] proposes an increasing number of high-quality screening sets with both primary and confirmatory dose-response data to guide computational method development and validation.

Comparison to other cavity-based pharmacophore perception methods

With respect to current structure-based pharmacophore perception methods,[11-29] the herein described approach presents five noticeable assets.

First, the pharmacophore perception method is fully automated, does not rely on any third party tool, and is freely available for non-profit research. The last criterion is particularly important to enable fair benchmarking. Second, in contrast with many alternative approaches,[11,15,16] IChem does not require user intervention in defining grid lattice coordinates. It scans the entire surface and can therefore generate as many pharmacophores as non-overlapping binding sites. Third, IChem offers the unique opportunity to restrict pharmacophore perception to binding cavities predicted as structurally druggable. Druggability (or ligandability) is predicted on the fly thanks to a robust support vector machine model, immediately after cavity detection. Fourth, IChem rules to select the most valuable pharmacophoric features have been derived from an exhaustive training set of 213 high resolution protein-ligand x-ray structures featuring non-redundant interaction patterns and 4,871 pharmacophoric features. Fifth, the method has been extensively validated on different test sets (Astex diverse set, DUD-E, PubChem bioassays) for its accuracy in ligand posing and virtual screening.

To the best of our knowledge, we provide here for the first time several high throughput screening data mimicking real life scenarios in which all true positives and true negatives are known. Such benchmarking data are, to our opinion, much more valuable than commonly used benchmarks in which actives (usually high affinity ligands) are mixed with chemically similar decoys of unknown affinity for the intended target.

CONCLUSIONS

We herewith propose an alternative computational method (IChem-Shaper2) to molecular docking to identify ligands from the single knowledge of a protein 3D structure. The concept of structure-based pharmacophores has already been exploited, but rarely led to pharmacophore queries truly adapted to virtual screening purposes. The proposed approach is fully automatized and consists in three consecutive steps that each can be customized if necessary: (i) detection of druggable cavities at the surface of the target of interest, (ii) generation of cavity-based pharmacophore queries, (iii) alignment of library compounds to the structure-based pharmacophore. The method appears to be quite robust in producing high-quality poses, distinguishing true actives from decoys, and retrieving confirmed hits from high throughput experimental screens. It should be considered as a novel weapon to the arsenal of current virtual screening methods such as protein-ligand docking or ligand-centric similarity searches. Since virtual screening benchmarks suggest its strong orthogonality to existing methods, we recommend its usage in parallel to docking and/or ligand-based methods in order to retrieve different chemotypes and optimize the real value of virtual screening hits for medicinal chemistry optimization.

ACKNOWLEDGMENTS

The Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) is acknowledged for allocation of computing time and excellent support. We sincerely thank Prof. M. Rarey (University of Hamburg,

Germany) for providing an executable version of Protoss. All datasets are freely accessible upon request to the authors.

Supporting Information. Sc-PDB Diverse Set of 213 protein-ligand complexes, Astex Diverse Set of 85 protein-ligand complexes, 10 DUD-E entries, Shaper2 force field for aligning ligand atoms to cavity features. This material is available free of charge via the Internet at http://pubs.acs.org.

References

1. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr. Computational methods in drug discovery. *Pharmacol. Rev.* **2014**, *66*, 334-395.
2. Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-art in ligand-based virtual screening. *Drug Discov. Today* **2011**, *16*, 372-376.
3. Spyrakis, F.; Cavasotto, C. N. Open challenges in structure-based virtual screening: Receptor modeling, target flexibility consideration and active site water molecules description. *Arch. Biochem. Biophys.* **2015**, *583*, 105-119.
4. Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **2010**, *53*, 539-558.
5. Plewczynski, D.; Lazniewski, M.; Augustyniak, R.; Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **2011**, *32*, 742-755.
6. Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today* **2008**, 13, 23-29.
7. Meslamani, J.; Li, J.; Sutter, J.; Stevens, A.; Bertrand, H. O.; Rognan, D. Protein-ligand-based pharmacophores: generation and utility assessment in computational ligand profiling. *J. Chem. Inf. Model.* **2012**, *52*, 943-955.
8. Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160-169.
9. Salam, N. K.; Nuti, R.; Sherman, W. Novel method for generating structure-based pharmacophores using energetic analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2356-2368.
10. Koes, D. R.; Camacho, C. J. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Res.* **2012**, *40*, W409-W414.

11. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849-857.
12. Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **1999**, *289*, 1093-1108.
13. Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **2009**, *25*, 621-627.
14. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287-2299.
15. Mortier, J.; Dhakal, P.; Volkamer, A. Truly Target-Focused Pharmacophore Modeling: A Novel Tool for Mapping Intermolecular Surfaces. *Molecules* **2018**, *23*, 1959.
16. Schuetz, D. A.; Seidel, T.; Garon, A.; Martini, R.; Korbel, M.; Ecker, G. F.; Langer, T. GRAIL: GRids of phArmacophore Interaction fieLds. *J. Chem. Theory Comput.* **2018**, *14*, 4958-4970.
17. Ahlstrom, M. M.; Ridderstrom, M.; Luthman, K.; Zamora, I. Virtual screening and scaffold hopping based on GRID molecular interaction fields. *J. Chem. Inf. Model.* **2005**, *45*, 1313-1323.
18. Ortuso, F.; Langer, T.; Alcaro, S. GBPM: GRID-based pharmacophore model: concept and application studies to protein-protein recognition. *Bioinformatics* **2006**, *22*, 1449-1455.
19. Radoux, C. J.; Olsson, T. S.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Identifying Interactions that Determine Fragment Binding at Protein Hotspots. *J. Med. Chem.* **2016**, *59*, 4314-4325.
20. Miranker, A.; Karplus, M. Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins* **1991**, *11*, 29-34.
21. Yu, W.; Lakkaraju, S. K.; Raman, E. P.; Fang, L.; MacKerell, A. D., Jr. Pharmacophore modeling using site-identification by ligand competitive saturation (SILCS) with multiple probe molecules. *J. Chem. Inf. Model.* **2015**, *55*, 407-420.

22. Hu, B.; Lill, M. A. Protein pharmacophore selection using hydration-site analysis. *J. Chem. Inf. Model.* **2012**, *52*, 1046-1060.

23. Bohm, H. J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput.- Aided. Mol. Des.* **1992**, *6*, 61-78.

24. Schuller, A.; Fechner, U.; Renner, S.; Franke, L.; Weber, L.; Schneider, G. A pseudo-ligand approach to virtual screening. *Comb. Chem. High Throughput Screen.* **2006**, *9*, 359-364.

25. Kirchhoff, P. D.; Brown, R.; Kahn, S.; Waldman, M.; Venkatachalam, C. M. Application of structure-based focusing to the estrogen receptor. *J. Comput. Chem.* **2001**, *22*, 993-1003.

26. Barillari, C.; Marcou, G.; Rognan, D. Hot-spots-guided receptor-based pharmacophores (HS-Pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores. *J. Chem. Inf. Model.* **2008**, *48*, 1396-1410.

27. Roland, W. S.; Sanders, M. P.; van Buren, L.; Gouka, R. J.; Gruppen, H.; Vincken, J. P.; Ritschel, T. Snooker structure-based pharmacophore model explains differences in agonist and blocker binding to bitter receptor hTAS2R39. *PLoS One* **2015**, *10*, e0118200.

28. Johnson, D. K.; Karanicolas, J. Ultra-High-Throughput Structure-Based Virtual Screening for Small-Molecule Inhibitors of Protein-Protein Interactions. *J. Chem. Inf. Model.* **2016**, *56*, 399-411.

29. Hu, B.; Lill, M. A. Exploring the potential of protein-based pharmacophore models in ligand pose prediction and ranking. *J. Chem. Inf. Model.* **2013**, *53*, 1179-1190.

30. Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem* **2018**, *13*, 507-510.

31. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. sc-PDB: a 3D-database of ligandable binding sites--10 years on. *Nucleic Acids Res.* **2015**, 43, D399-D404.

32. Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding protein-ligand interaction patterns in fingerprints and graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623-637.

33. http://bioinfo-pharma.u-strasbg.fr/scPDB (accessed Sep. 29, 2018).

34. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726-741.
35. https://www.ccdc.cam.ac.uk/support-and-resources/Downloads/ (accessed Sep.29, 2018).
36. Certera USA, Inc, Princeton, NJ 08540 , U.S.A.
37. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, 28, 235-242.
38. Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J. Cheminform.* **2014**, *6*, 12.
39. 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. *J. Med. Chem.* **2012**, *55*, 6582-6594.
40. National Center for Biotechnology Information. PubChem BioAssay Database; AID=604, https://pubchem.ncbi.nlm.nih.gov/bioassay/604 (accessed Sep. 26, 2018).
41. National Center for Biotechnology Information. PubChem BioAssay Database; AID=644, https://pubchem.ncbi.nlm.nih.gov/bioassay/644 (accessed Sept. 26, 2018).
42. Molecular Networks GmbH, Erlangen, Germany.
43. OpenEye Scientific Software, Santa Fe, NM 87508, U.S.A.
44. National Center for Biotechnology Information. PubChem BioAssay Database; AID=743080, https://pubchem.ncbi.nlm.nih.gov/bioassay/743080 (accessed Sept. 26, 2018).
45. National Center for Biotechnology Information. PubChem BioAssay Database; AID=1777, https://pubchem.ncbi.nlm.nih.gov/bioassay/1777 (accessed Sept. 26, 2018).
46. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317-324.
47. Kurogi, Y.; Guner, O. F. Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr. Med. Chem.* **2001**, *8*, 1035-1055.

48. Grant, J. A.; Gallardo, M.; Pickup, B. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653-1666.
49. Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572-584.
50. Shape Toolkit 2.0.1, https://docs.eyesopen.com/toolkits/python/shapetk/index.html (accessed Aug. 18, 2018).
51. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comp. Chem.* **1996**, *17*, 490-519.
52. Nicholls, A.; Wlodek, S.; Grant, J. A. SAMPL2 and continuum modeling. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 293-306.
53. https://docs.eyesopen.com/szybki/szybkitheory.html (accessed Aug. 17, 2018).
54. Dassault Systèmes, BIOVIA Corp., San Diego, CA 92121, U.S.A.
55. Inte:Ligand GmbH, Austria, Vienna.
56. Poli, G.; Seidel, T.; Langer, T. Conformational Sampling of Small Molecules With iCon: Performance Assessment in Comparison With OMEGA. *Front. Chem.* **2018**, *6*, 229.
57. Jain, A. N. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281-306.
58. Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74-82.
59. Guner, O. F.; Bowen, J. P. Setting the record straight: the origin of the pharmacophore concept. *J. Chem. Inf. Model.* **2014**, *54*, 1269-1283.
60. Spitzer, R.; Jain, A. N. Surflex-Dock: Docking benchmarks and real-world application. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 687-699.
61. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225-242.

62. Cleves, A. E.; Jain, A. N. Knowledge-guided docking: accurate prospective prediction of bound configurations of novel ligands using Surflex-Dock. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 485-509.
63. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Cryst. Sect. A* **1976**, *32*, 922-923.
64. Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics* **1955**, 2.
65. Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. *J. Cheminform.* **2016**, *8*, 56.
66. Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 update. *Nucleic Acids Res.* **2017**, *45*, D955-D963.
67. Kruger, D. M.; Evers, A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem* **2010**, *5*, 148-158.
68. Tanrikulu, Y.; Kruger, B.; Proschak, E. The holistic integration of virtual screening in drug discovery. *Drug Discovery Today* **2013**, *18*, 358-364.
69. Tian, S.; Sun, H.; Li, Y.; Pan, P.; Li, D.; Hou, T. Development and evaluation of an integrated virtual screening strategy by combining molecular docking and pharmacophore searching based on multiple protein structures. *J. Chem. Inf. Model.* **2013**, *53*, 2743-2756.
70. http://bioinformatics.psb.ugent.be/webtools/Venn/ (accessed Sept. 26, 2018).

For Table of Contents use only

All in One: Cavity Detection, Druggability Estimate, Cavity-Based Pharmacophore Perception and Virtual Screening.

Viet-Khoa Tran-Nguyen, Franck Da Silva, Guillaume Bret and Didier Rognan