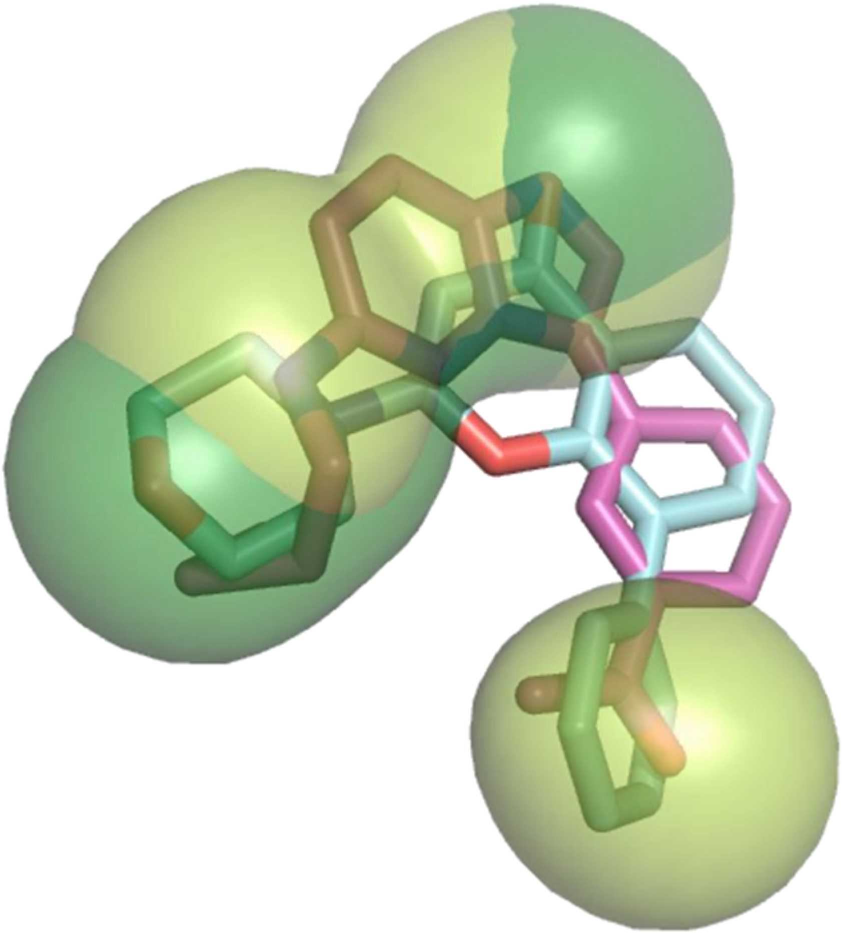
UNIVERSITAT DE BARCELONA

On the usage of lipophilic descriptors for molecular similarity evaluation

Javier Vázquez Lozano

On the usage of lipophilic descriptors for molecular similarity evaluation



Javier Vázquez Lozano

2019

UNIVERSITAT DE BARCELONA

Facultat de Farmàcia i Ciències de l’Alimnetació

Programa de doctorat en Biomedicina

On the usage of lipophilic descriptors for molecular similarity evaluation

JAVIER VÁZQUEZ LOZANO

2019

UNIVERSITAT DE BARCELONA

Facultat de Farmàcia i Ciències de l'Alimnetació

Programa de doctorat en Biomedicina

On the usage of lipophilic descriptors for molecular similarity evaluation

Memoria presentada por Javier Vázquez Lozano para optar al título de Doctor por la Universidad de Barcelona.

Dr. Javier Luque Garriga	Dr. Enric Herrero Abellanas
Director **Tutor**	**Director**

Javier Vázquez Lozano

Doctorando

A los que,
con más o menos atención,
escucharon.
Y en especial a ti, Carlos.

Acknowledgments

I would like to thank my thesis supervisors Prof. Javier Luque and Dr. Enric Herrero for their advice and support. I am also grateful to Dr. Enric Gibert, who always had time to talk, and Dr. Jordi Juarez, partly responsible for all of that.

My thanks to Albert Herrero, my first colleague to share the alignment nightmare, and Dr. Alessandro Deplano, with whom I published my first paper. To Ania Rudzinska, Alberto Salas, Elisa Gómez, Irene Montero, Lluís Campos. I learned something from everyone.

Dr. Toni Viayna, Dr. William Zamora, and Dr. Constantí Seira. Although we didn't share much time, the collaborations with them were fruitful. I hope we will continue discussing for years. I remember with an especial affection the classes teaching with Dr. Toni.

I am not forgetting of course, the helpful role of: Dr. Carol Estarelles, Dr. Salomé Llabrés, Dr. Marina Corbella, Dr. Tiziana Ginex, Dr. Sylvana Pinheiro, Dr. Axel Bidon-Chanal and Dr. Sergi Ruiz.

Thanks to Jess Tarrío, Pablo Garcilópez, and Enric Bartomeu for your advice and contributions.

Y en mi lengua materna agradecer con un cálido abrazo a mi familia y amigos de siempre. Adri, Serra, Marc, Cristian, Otero, Guerrero, Pablo, Enric y Jaume, que siempre habéis estado. A mi padre y mi madre, su apoyo. Y a mi hermano, a quien dedico este trabajo.

GENERAL INDEX

Index

1 INTRODUCTION 15

1.1 Molecular Similarity 16

1.1.1 Overview of Molecular Similarity 16

1.1.2 Methods for similarity Evaluation 17

1.1.3 3D Molecular Similarity Applications. 20

1.1.3.1 Quantitative structure-activity relationships 20

1.1.3.2 Ligand Based Virtual Screening Tools 20

1.1.3.3 3D similarity as a complement of structure-based methods. 23

1.1.4 In the interface of molecular comparison: basics of our proposal 24

1.1.4.1 Molecular alignment: the use of multipole moments. 24

1.1.4.2 3D Grid-based similarity function 25

1.2 Lipophilicity in drug design 26

1.2.1 The QM Continuum Solvation MST Model. 28

1.2.2 3D lipophilic profile from MST calculations. 29

2 OBJECTIVES 35

3 PUBLICATIONS 39

3.1 PAPER1: "Development and Validation of Molecular Overlays Derived From 3D Hydrophobic Similarity with PharmScreen" 41

3.2 PAPER 2: "Lipophilicity in drug design: an overview of lipophilicity descriptors in 3D-QSAR studies" 71

3.3 PAPER 3: "Similarity assessment of Lipophilic Distribution: A Boost for Structure-Based Methods" 95

3.4 PATENT: "Calculating molecular similarity" 132

4 RESULTS SUMMARY 167

4.1 Lipophilic descriptors in molecular alignment 167

4.1.1 Molecular descriptors and implementation 167

4.1.2 Weights calibration and QM methods 168

4.1.3 Overlay accuracy validation 170

4.2 Lipophilic descriptors in 3D-QSAR 173

4.2.1 On the QM method applied to derive the 3D-QSAR model 173

4.2.2 3D-QSAR validation 173

4.3 Three-dimensional similarity in combination with molecular docking 175

4.3.1 Combined score function exploration 176

4.3.2 Performance evaluation 178

5 DISCUSSION 183

6 CONCLUSIONS 187

Bibliography **190**

Other outputs related to the work of this thesis: 203

1 CHAPTER ONE Introduction

1 INTRODUCTION

From infancy to adulthood, similarities between people, objects, events, or situations are established. Similarity is inherently associated with daily activities, such as discerning between colors and shapes, it being an intrinsic feature of humans. This concept is also well established in scientific studies, as can be illustrated in the similarity between sequences to build up 3D structures of proteins by homology modeling, or phylogenetic analysis of organisms. The concept of similarity, thus, constitutes one of the basic processes of human thought: **the ability to learn through comparative analysis.**

Already in 1869, Dmitri Mendeleev based his reasoning on the similarity of chemical properties to create the first version of the periodic table. To our astonishment, Mendeleev even advanced the properties of certain elements that were still to be discovered.[1] Chemical informatics and medicinal chemistry also take profit of this concept to find relationships between molecules. Attempts to quantify how a given compound resembles to another have a long history in drug discovery. Furthermore, molecular similarity has been exploited to find guidelines in search of novel compounds with suitable pharmacological profiles. Therefore, the concept of similarity is linked to another characteristic: **the ability to establish predictions**.

Many methods that compute molecular similarities have been developed in the last 50[2] years, and new techniques continue to be proposed.[3] Molecular similarity is a complex issue that can only be assessed if the search space is limited to key molecular properties and under certain predefined conditions. Thus, each method defines a set of molecular descriptors and comparison algorithms that enable exploring the chemical space and quantify (di)similarities. In this framework, similarity measurements are associated with a variety of chemical features, such as bonding patterns, atomic positions, molecular conformation, shape and volume, and spatial disposition of molecular properties. Nevertheless, the inclusion of solvation properties in similarity measurements has been more elusive, though desolvation is recognized to be one of the major forces that modulate the binding of ligands to the target receptors.

In this context, the research group on Computational Biology and Drug Design (CBDD) at the University of Barcelona developed a methodology based on the use of continuous solvation models coupled with methods of quantum chemistry (QM)[4–8] to evaluate the solvation free energy of (bio)organic compounds, and to decompose this thermodynamic quantity into atomic contributions, yielding lipophilic profiles suited for similarity studies.

Under these premises, this thesis seeks to reconcile similarity and lipophilicity through the development of a 3D grid-based algorithm, called PharmScreen[9], for rigid-body molecular superposition and lipophilic similarity searching. This chapter is intended to briefly introduce the main blocks that constitute this work, considering (i) molecular similarity as a drug discovery tool, and (ii) lipophilicity as the primary descriptor.

1.1 Molecular Similarity

1.1.1 Overview of Molecular Similarity

The study of molecular similarity plays a main role in chemoinformatics[10,11] and is a key step in medicinal chemistry.[12,13] Its assessment has been closely merged with the history of chemical and physical science for more than one century,[14,15] pioneered by Kopp[16] in 1842, who reported the first relationship between structures and physicochemical properties. However, it was not until the middle of the 1980s that similarity methods came into wide use. This resulted from two major contributions: the implementation of the quantitative structure–activity relationships (QSAR) methodology developed by Hansch and Fujita[17], and the work carried out at Ledere[18] and Pfizer[19] Laboratories showing that similarity between two compounds can be evaluated in a computationally efficient way. Finally, the concept of molecular similarity was consolidated through the well-known *similarity-property principle* (SPP), which was introduced in the early 1990s and became a milestone in molecular similarity analysis.

The SPP states that "*structurally similar molecules are more likely to have similar properties*",[14] including biological activity. This principle summarizes the core issue of similarity searches, where molecular comparisons can be exploited to infer resemblance in the biological response of compounds. In this context, the existence of data relative to the biomolecular target is not a *sine qua non* condition for the development of drug discovery projects. Furthermore, by resorting to simplified information (e.g., a set of known ligands), it may be possible to identify novel hits (**compare**) and find structure-activity relationships (**predict**) at a low computational expense. These features made similarity evaluation to be a fundamental ingredient for the development of many computational methods.

Similarity helps in the design of mimetic and bioisosteric compounds, provides a measure of chemical diversity between molecules, defines a metrics for structure-activity correlations, and facilitates the determination of the bioactive molecular overlay in search of novel actives. Common **ligand-based drug design (LBDD)** techniques such as virtual screening (VS) and clustering methods,[20] have their origin in these applications and have become key resources in **computer-aided drug discovery (CADD)**.

In the last state-of-the-art review in tools that exploit molecular similarity[3], a total of 115 publications of LBDD techniques were analyzed, showing the significant impact and usefulness of these techniques in drug discovery. Computational, chemical, and life science journals that were operational and had an impact factor greater than 2.0 in 2009 were selected in this study. The analysis of these works showed that~60% of these studies reported hits with <1 μM potency and that~30% of the hits had a potency <10 μM. Thus, the majority of the ligand-based virtual screening (LBVS) hits were relatively potent. In addition, they determined that 78 of the analyzed LBDD studies had reported totally unique structures not published before. Taken together, the results of their analysis revealed that many practical LBDD applications resulted in the identification of new and interesting active compounds that recognize the value of using these techniques.

The practical use of similarity techniques in the last years is still a hot topic in rational drug design, either alone[21–27] or in combination with structure-based protocols.[28–37] To further reinforce its applicability, let us remark the fact that reference books provide a specific chapter or section for the application of molecular similarity in different areas of the field, such as "*Computational Methods for GPCR Drug Discovery"*[38], *"Computational Toxicology"*[39], *"Bioinformatics"*[11], and *"Successful Drug Discovery"*.[40]

Despite the existence of successful stories and the broad adoption of well-established techniques, molecular similarity still faces challenges, starting with its definition. Just like in any other comparison context, molecular similarity is linked to a certain degree of subjectivity, since it is difficult to use a unique similarity evaluation approach in all scenarios. Consider, for example, the neurotransmitters norephedrine (C_6H_5-CHOH-CH(CH_3)-NH_2), ephedrine (C_6H_5-CHOH-CH(CH_3)-$NHCH_3$), and pseudoephedrine (C_6H_5-CHOH-CHNH(CH_3)-CH_3). If we pay attention to the 2D structure of the whole molecule, these drugs appear visibly similar. The difference between ephedrine and pseudoephedrine is the stereochemistry of a single chiral center, while these two compounds differ from norephedrine only in the substitution of a hydrogen atom by a methyl group. However, the high 2D similarity does not correlate with their physiological behavior, which is driven by its 3D structure. Ephedrine is the most potent stimulant,[41,42] and it is used as a bronchodilator, vasoconstrictor, and cardiac stimulator. Pseudoephedrine is mainly employed in flu treatment as a decongestant, and norephedrine is used as an appetite suppressor and in cold and cough medications. The distinct physiological effects produced by these phenylpropanolamines exemplify how similarity is the consequence of the bounded comparison of a multifaceted nature: molecules that seem quite similar from a structural point of view (e.g., atom connections) could show significant differences from a medicinal chemistry perspective.[43]

Accordingly, similarity acquires a subjective meaning influenced by the molecular features relevant for similarity appraisal and the definition of the similarity measurement. Indeed, the **similarity paradox**[44] states that small modifications in molecules can cause them to modify their activity. To avoid this paradox it is convenient, if not necessary, to take into account two main factors: (i) to select series of comprehensive compounds for testing, and (ii) to consider that the biological activity usually stems from the interplay of a number of complex processes, which cannot be easily represented by a set of linear relationships.[15] To better describe these processes, non-linear estimations should be used, where the structural, topological, and molecular descriptors are independent of one another.

1.1.2 Methods for similarity Evaluation

In chemoinformatic approaches, the measure of molecular similarity involves mainly two major components[45]: the molecular descriptors and the similarity function. Descriptors capture the relevant features for molecular evaluation, and the similarity function transforms the comparison between pairs of properties in real numbers commonly within the interval [0,1]. For each component of the molecular representation, a certain degree of importance is assigned through the weighting of similarity coefficient. Some works even consider this weighting scheme as a third independent component.[15,46,47]

Molecular descriptors can be defined in one, two, or three dimensions:

I. **1D-molecular descriptors** include the simplest representation without atomic connections information: (i) physicochemical/biological properties, and (ii) chemical features[48] (e.g., number of atoms, bonds, functional groups, etc.). This chemical information can be expressed in a simple one-line[49–51] code, which allows a fast extraction and treatment of molecular information. For example, LINGO[52], a hologram of SMILES strings, provides a method to derive structure-related properties and to compute directly from one-dimensional representations.

II. **2D-molecular descriptors** are the most common similarity method derived from connection tables, 2D-electrotopological and topological descriptors: (sub)graphs and (sub)structure.[53–55] The approaches using electrotopological descriptors represent the electrical properties of atoms and molecules in a topological frame (e.g., the electron accessibility for each atom). On the other hand, the graph-based representation defines molecules as a function of their structure or substructure. For example, one molecule can be reduced to nodes where each one corresponds to ring systems, heteroatoms, acyclic components, or functional groups.[56]

III. **3D-molecular descriptors** were introduced and gained more attention recently because of their potential to explore the molecular projection in the surrounding space.[57] Since the binding affinity between molecules and target proteins is governed by atomic interactions in the 3D space, molecules with similar 3D shape and properties could have shared biological activities, even though their 1D and 2D representations are not similar. 1D and 2D methods tend to find close chemical analogs to known active compounds, but fail to predict activity differences between them[58]. What is lacking from 1D and 2D methods are obviously 3D structural information of compounds and target proteins.

    3D molecular descriptors provide molecular information in the context of molecular properties, chemical groups, or the spatial distribution of atoms. In general, it refers to molecular surfaces and volumes. Nevertheless, most CADD tools using 3D representations present alignment dependence.

3D molecular representations seem closer to the reality of the ligand-receptor binding process. However, some validation works[59–61] reports better performance from 2D methods in terms of the number of actives retrieved. Noise associated with incomplete conformational sampling, subtle deficiencies in the quality of descriptors, offering little additional information to 2D connectivity, and limitations in molecular alignments, are possible causes to explain this behavior. Nevertheless, the reason behind this response seems to be also conditioned by two main factors: (i) the benchmarking sets used for such comparison are biased towards 2D similarity, and (ii) the concept of correlating the number of actives retrieved with high accuracy instead of structural diversity. Methods making use of 3D features have less dependence on underlying atom connectivity, and thus, provide highly ranked structures based on new scaffolds.[62]

The *similarity function* provides a quantitative measure of the chemical resemblance from a global or local point of view. Local similarities are focused on a specific molecule section (heteroatoms, functional group, rings, etc.), which is known or can be inferred to be responsible for the activity, while global similarity considers the whole molecule. For example, a pharmacophore model where only specific features responsible for the activity are compared may apply a local similarity evaluation, whereas the confrontation of the whole volumes of different drug-like molecules implies global similarity measurements.

The **Tanimoto coefficient** is the most widely used global similarity coefficient for both binary fingerprints and continuous data representations. It is defined in Eq. 1, with $\alpha = \beta = 1$. It is also known as the **Dice coefficient** when $\alpha = \beta = {}^{1}\!/_{2}$.

$$Similarity\ coefficient_{A,B} = \frac{c}{\alpha(a-c)+\beta(b-c)+c} \qquad 1$$

where *a* represents the number of features present in molecule A, *b* is the number of features in molecule B, and *c* is the number of features in common between molecules A and B.

However, a partial similarity measurement, such as **Tversky coefficient**, can also be extracted from Eq 1. When values of $\alpha = 1$ and $\beta = 0$ are used, the number of features only present in molecule B disappears from the equation. This means that even if there are features only present in molecule B, it does not influence the similarity. Thus, the Tversky similarity between benzene and naphthalene is 1, since naphthalene completely contains benzene as a substructure. An extensive description of similarity coefficients has been provided elsewhere [63,64].

These similarity coefficients can be used to select those compounds with a higher probability of being active and classify them according to the degree of similarity with selected templates. Compounds with higher similarity can be kept, while compounds with lower similarity can be discarded. While this approach is easy to use and well accepted, it has some limitations. On one side, a high similarity coefficient value does not always imply that two compounds will have the same activity. Some minor structural changes could greatly modulate the activity of a compound depending on how they affect the interactions with the protein, as has been noted previously. On the other, there is no universal cutoff of similarity value for determining that a compound will have similar activity to a reference molecule.

To avoid these inconveniences, VS can be complemented with SAR studies and an activity cliff description[65], thereby focusing similarity measurements on activity determinants using partial similarity functions. Moreover, the combination of similarity methods based on different criteria is recommended to discard compounds that may be incorrectly prioritized by a given methodology. Different types of coefficients perform differently in alternative situations, and the results obtained can thus complement each other.[66]

1.1.3 3D Molecular Similarity Applications.

It is common to see applications of 3D similarity in early stages of drug discovery,[3,67,68] although the foundations of QSAR were already established almost 40 years ago[69]. Accordingly, commercial tools have been developed using physicochemical abstractions like 3D shape or electrostatic potential to identify similar compounds and apply these molecular properties to the design of new compounds.

Methods using 3D similarity are divided into two main groups: QSAR and LBVS tools. 3D-QSAR is used to drive potency and identify active compounds in lead-optimization projects[70]. The Comparative molecular field analysis (CoMFA)[71] is the most paradigmatic method. On the other hand, LBVS is used to rank compounds based on the similarity relative to a given reference. Both categories are described in more detail in next sections. Finally, 3D similarity can also be used to complement structure-based virtual screening (SBVS).

1.1.3.1 Quantitative structure-activity relationships

3D QSAR approaches play a significant role in early phases of chemical exploration, where the aim is understanding the SAR and use this knowledge to build a predictive model. The wide applicability of these models can be noted by the huge number of works where CoMFA[71] and related techniques such as Comparative Molecular Similarity Indices Analysis (CoMSIA),[72] have been used.[73–76] CoMFA exploits the combination of electrostatic and steric fields. On the other hand, CoMSIA was formulated extending the numbers of descriptors, including hydrophobic and hydrogen-bond donor and acceptor properties, as well as including a distinct expression for the projection of these properties into the surrounding space.

3D-QSAR methodologies require a series of structural analogs that interact in the same way at the same binding site. Since the differences captured by the model must be only due to the accommodation of different functional groups to the binding site, molecular alignment becomes critical for the success of the process. Besides predicting activity, QSAR models can be used to understand the relevance of particular features for a specific effect.[77]

1.1.3.2 Ligand Based Virtual Screening Tools

LBVS can rank novel ligands by 3D similarity in order to find compounds that are related to known actives. Commercial software has been developed to explore databases of chemical structures that are similar to known actives or possess a pharmacophore or substructure in common with a known active.

The most representative approach within the tools that do not require molecular overlay (**non-superpositional methods**) is the Ultrafast Shape Recognition method (USR),[78] which involves the analysis of atomic distances to a set of reference positions. Later versions were adapted to include more molecular properties.[79,80]

Superpositional methods encompass protocols where the similarity between molecular pairs is computed after overlapping molecules (**Table 1**). Typical types of descriptors comprise Gaussian shape, molecular fields, and pharmacophoric features.

One of the most used approximations to represent molecular shape and volume in superpositional methods is the Gaussian function employed by ROCS (Rapid Overlay of Chemical Structures)[81,82], ShaEP (Shape and Electrostatic Potential)[83], SHAFTS (Shape-Feature Similarity)[84] and LIGSIFT[85]. Moreover, these approaches supplement the shape-based similarity by mapping chemical groups into pharmacophoric features[86–88]. An alternative representation is what Phase Shape[89,90] applies, where the whole molecular volume is treated as hard spheres. In contrast, Surflex-Sim[91] and SURFCOMP[92] explore the concept of local similarity exploiting specific properties of the molecular surface.

Molecular fields[93] represent a distinct approach where comparison relies on the spatial variation of interaction energies with a probe. The attributes that are assumed to lead the biological activity, such as the regions of positive and negative charge, are used to define fields. Blaze[94,95] applies this principle to incorporate off-atom charges to obtain a representation of the electronic environment. Moreover, steric and hydrophobic features are included. BRUTUS[96] or MIMIC[97] combines steric and electrostatic fields.

A 3D pharmacophore model extracts common chemical features that are essential for bioactivity from a series of active ligands. Pharmacophoric points generally include a core and features such as hydrogen bond acceptors/donors, heteroatoms, and charged groups[98]. Once the pharmacophoric points are determined, a triangle or tetrahedron is formed between them, and the distances are used for similarity measurements. These representations are further encoded into strings that have information on the selected features and the distances of the edges of the polygon built. As an example, FLAP[99] defines discrete points for fields of molecular interactions computed using GRID[100]. All possible combinations of 4 pharmacophoric points are generated (quadruplets). Similarly, Tuplets[101,102] encodes 2, 3, and 4 pharmacophoric points and determines five features: donor atom, acceptor atom, hydrophobic center, positive nitrogen, and negative center.

Table 1 | An overview of some superpositional methods

Method	**Description**	**Sub-Class**
ROCS[81,82]	Fast Gaussian overlay based shape comparison. Widely used shape based virtual screening tool.	Shape-guided Gaussian function.
ShaEP[83]	Generate consensus shape patterns based on structural features of known ligands.	Shape-guided Gaussian function
SHAFTS[84]	It combines shape similarity with pharmacophoric features. Employs a hybrid similarity metric combining shape and chemical similarity.	Shape-guided Gaussian function

LIGSIFT[85]	Uses Gaussian molecular shape overlay for fast small molecule alignment and a size-independent scoring function for efficient VS based on the statistical significance of the score.	Shape-guided Gaussian function
Phase Shape[89,90]	Phase Shape represents a structure as a set of hard atomic van der Waals spheres and uses atom triplets to align molecules.	Shape-guided not Gaussian function
SURFCOMP[92]	Molecular surface is divided into patches and corresponding patches are identified using geometrically invariant descriptors and physicochemical properties.	Shape-guided not Gaussian function
Blaze[94,95]	Exploits the local extrema of molecular interactions fields to align and score molecules.	Field-based
BRUTUS[96]	Aligns molecules using field information derived from charge distributions and van der Waals shapes of the compounds.	Field-based
MIMIC[103]	Molecular field based matching program (steric volume and electrostatic fields).	Field-based
FLAP[99]	Provides a common reference framework for comparing molecules, using GRID Molecular Interaction Fields (MIFs).	Pharmacophore
Tuplets[101,104]	Encodes the interfeature distances of a set of interesting pharmacophore features: pairs, triplets, quartets.	Pharmacophore

Overall, molecular overlap and similarity measurements are accomplished by using a variety of approaches, which exploit shape, electrostatic, and pharmacophoric features. Especially, 3D shape-based similarity analysis has become, in recent years, the method of choice in increasing number of virtual screening campaigns.[105] However, although the hydrophobic/hydrophilic balance is known to be critical in **pharmacokinetics** and **pharmakodinamics**[106–108] and ligand (de)solvation is a major contribution to the variation in maximal achievable binding free energy for a drug-like molecule,[109,110] the consideration of the differential solvation properties of molecules in similarity measurements and alignment procedures for VS has not been widely explored, partly due to the difficulty in defining an accurate 3D distribution of hydrophilic/hydrophobic properties.

1.1.3.3 3D similarity as a complement of structure-based methods.

VS tools are divided in two groups: ligand-based (LB) and structure-based (SB) methods. SB methods exploit receptor data[111,112] and have to take into account different structural conformations. In addition, they generally use oversimplified scoring functions and have a high computational cost. On the other side, LB methods are biased toward the existence of known ligands, and it is affected by the training set quality, and obviously the lack of protein structure information. Accordingly, combining molecular similarity with structure-based approaches, such as docking, has been proposed as a way to solve the shortcomings of both approximations using sequential, parallel, or hybrid approaches. [113–115]

The **sequential** approach divides the screening process in multiple steps with the goal of overcoming the expensive computational cost of SB methods. A prefiltering is used at the beginning where the less expensive LB approach is applied. The best compounds are further evaluated, usually using docking into the protein-binding site.[116–118]

In the **parallel** approach, both methods are run independently, and the top hits retrieved from each method are selected for biological testing. One of the first works that expose the prospective of this procedure was conducted in 2011[119]. Nevertheless, this approach has given rise to hybrid functions that represent a true combination of structural and ligand information. Protein-ligand pharmacophores concept arises from this idea. The observed protein-ligand interactions are directly translated into pharmacophore features which have demonstrated success in VS[120] and for profiling purposes.[121]

Hybrid approaches articulate SB and LB information in a unique core. Three alternatives are (i) the use of pharmacophore models to constrain poses generated by docking in a specific binding mode, (ii) the development of pseudoreceptors from an expansion of traditional QSAR methods, and (iii) the 3D similarity between docked compounds and a known crystallographic ligand is performed to re-score the docking ranking[122], Figure 1. (see ref.[123] for details about pharmacophoric constraints, refs.[114,124] for pseudoreceptors development, and ref[125] for the re-scoring of docked poses).

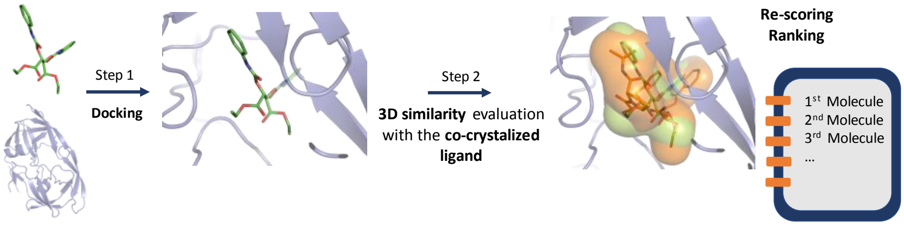


Figure 1 | Example of hybrid approach where the docking poses and co-crystallized ligand similarity are computed to re-score docking ranking.

1.1.4 In the interface of molecular comparison: basics of our proposal.

This section aims to present the antecedents of the two fundamental pillars on which our tool of molecular alignment and lipophilic similarity search, PharmScreen, is sustained: the multipolar expansion of electrostatic potential and the grid-based similarity assessment.

1.1.4.1 Molecular alignment: the use of multipole moments.

Several theories and methodologies exploit electron distribution to evaluate molecular similarity in alignment procedures,[83,126,127] a core step in superpositional ligand-based methods. These descriptors are intended to model the charge distribution in drug-like compounds, as electrostatics plays an essential role in the interaction with the target. Platt et al.[128] proposed to use the Cartesian multipole expansion moments as a solution to approximate the evaluation of the electrostatic interaction energy[127] and are the main procedure used to align molecules based on electron distribution.[128,129] In parallel, Grant et al.[130,131] proposed the so-called shape multipoles or moments as descriptors to align molecules based on a Gaussian model. Their work was a success and the starting point to develop the program ROCS.[132]

In this framework, the zero-order moments of the mass and charge distribution are the total sum of their contributions: the total mass and the net charge, respectively. If the mass distribution is our subject of study, the first-order moments located in the center of mass are zero, and the second-order moments are the moments of inertia, Eq. 2, can then be used for shape-based molecular alignment.

$$\boldsymbol{I} = \sum_{i=l}^{N} m_i\left(\vec{r_i}^{\,2}\,1 - \vec{r_i}\vec{r_i}\right) \qquad 2$$

where *i* is summed over the atomic centers, m*ᵢ* is the atomic weight of the *i*th atom, $\vec{\boldsymbol{r}}_{\boldsymbol{i}}$ is a vector from the center of rotation to the *i*th atom.

In the inertial tensor, the diagonal moments corresponding to the three spatial directions for which the angular velocities about these directions are parallel to their respective components of angular momenta are called the principal inertia moments. These moments depend, however, on the center of expansion. Since the origin is otherwise arbitrary, the moments of the mass distribution are calculated at the center-of-mass of each molecule.

In the case of the charge distribution, the first-order moment of the charge distribution is the dipole moment, which depends on the net charge. If the molecule is charged, the value of the dipole depends on the reference point used for its calculation, but its value is not affected by the origin of the axes for neutral molecules. This obeys to the fact that the lowest order nonvanishing moment of the electrostatic multipolar expansion does not depend upon the reference origin. The values of all the higher-order multipolar moments depend on the choice of the origin of the multipolar expansion, which may affect the calculation of second- and higher-order components of the multipolar expansion.

The solution proposed by Platt et al.[128] was the expansion of the electrostatic potential through multipolar decomposition at selected reference points. This method offers a reference frame were quadrupolar axes are calculated relative to the **"monopole center"** for charged molecules, and **"dipole center"** for neutral molecules. This choice is dictated by the convergence of the leading terms in the multipolar series expansion of the electrostatic potential. If the molecule is charged, the leading term is the monopole term, and the dipolar term is zero when determined relative to center-of-charge. Thus, the choice of center of monopole as the origin of expansion guarantees that the monopolar contribution to the electrostatic potential most closely approximates the total electrostatic potential over most of the space. Hence, the quadrupolar term emerges as the second most relevant contribution, and the principal quadrupolar axes can then be used to align molecules.

The first nonvanishing term for neutral polar molecules is the dipole moment. By using the center-of-dipole as reference point, the dipole moment lies along one of the principal axes of the quadrupole, whose value along this direction would be zero, and the quadrupolar tensor yields an orthogonal set of principal axes that can be utilized for molecular alignment, Eq. 3.

$$\boldsymbol{Q} = \sum_{i=l}^{N} \log P_{o/w,i} (3\vec{r}_i\vec{r}_i - |\vec{r}_i|^2 1) \qquad 3$$

1.1.4.2 3D Grid-based similarity function

Finding molecular diversity is one of the primary aims in early stages of drug discovery projects, especially in VS. However, changes in the functional groups of drug-like molecules give rise to changes in biological activity. Thus, structural agnostic methods based on molecular interactions fields (MIF) are an appropriate option to address the compromise offered by chemical diversity. This idea was initially introduced by Carbó et al.[133] to compare electron densities and subsequently applied in QSAR[72,134] and LVBS[94,96,103] tools to compute similarities between shape, electrostatic, and hydrophobic size surfaces.

The evaluation of overlapped molecules using spaced grid points surrounding them is a conventional approach to apply MIF in molecular similarity studies,[71,72,96] Figure 1.

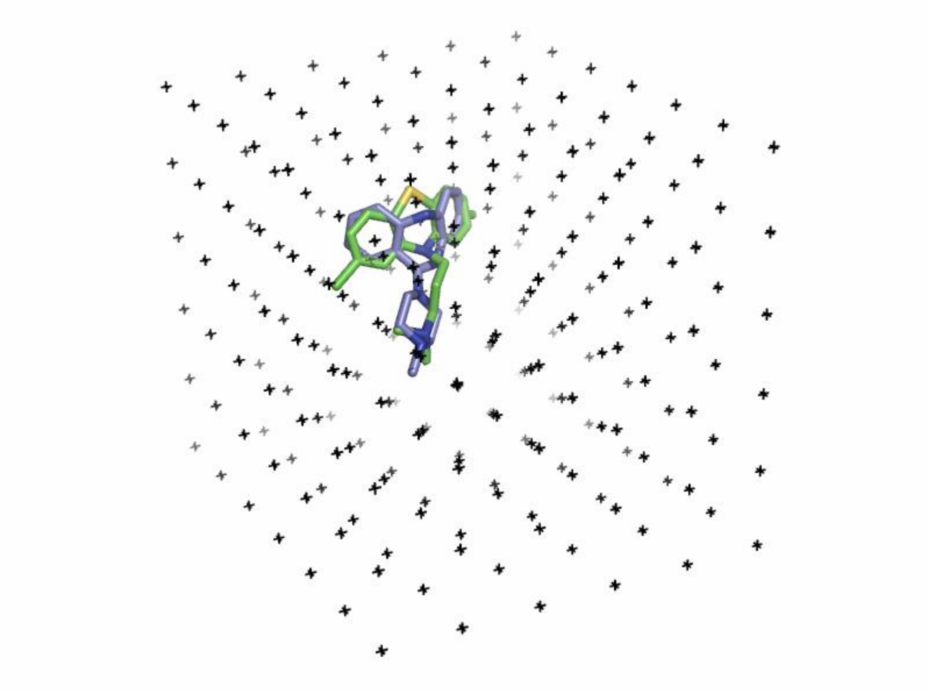


Figure 2 | Example of a superposition of two molecules in a set of field points used for similarity index calculations.

Under this scope, the energy between a probe atom and a molecule is computed at each grid point, typically using an exponential function,[72] p(q); eq 4.

$$\boldsymbol{p(q)} = \sum_{i=l}^{N} w_{probe,k}\, w_{ik} e^{-\alpha r_{iq^2}} \qquad 4$$

where i = summation index over all atoms of the molecule q under investigation, w_{ik} = value of the physicochemical property k of atom i; $w_{probe,k}$ = probe atom value of property k, a = attenuation factor, and r_{iq} = mutual distance between probe atom at grid point q and atom i of the test molecule.

Thus, a similarity index can be quantified theoretically by comparing the field values to compute similarities between pairs of aligned molecules.

1.2 Lipophilicity in drug design

The therapeutic effect of a drug is achieved when a molecule binds and modifies a druggable disease-protein target at a specific binding site. Studies of target druggability highlighted the relevance of shape and hydrophobicity in drug binding.[135–138] In particular, ligand desolvation was recognized to be mostly responsible for the variation in maximal achievable binding free energy for a drug-like molecule.[139] Otherwise, polar interactions in the binding sites play a primary role for both binding and selectivity.[140–142] Hence, the analysis of the 3D pattern of lipophilicity of ligands could be crucial to identify specific features of ligand recognition at druggable pockets.

In this context, Cheng et al.[109] modulated the **maximal achievable binding free energy** (ΔG_{MAP}) for a drug-like molecule from the observation of a clear trend toward higher fraction hydrophobic SASA and a lower radius of curvature of the pocket for druggable targets. Contemporary studies that analyzed the use of descriptors to discriminate small molecule-binding pockets reported similar tendency.[136,137] The druggability model assumes that favorable affinity is largely driven by the hydrophobic effect.[143] This agrees with the view that druggable binding sites appear to be closed and “greasy” cavities, whereas polar interactions are crucial for binding and selectivity[140–142]. Hence, it is reasonable to expect that the analysis of the 3D pattern of hydrophobicity/hydrophilicity of ligands could be a valuable feature to define molecular similarity between drug-like molecules.

Computational empirical methods have already been developed to estimate lipophilic interactions between ligand and receptors from octanol/water partition coefficient ($\log P_{o/w}$) determined by molecular fragments or atom types.[144–146] These empirical approaches to lipophilicity potential include the molecular lipophilicity potential (MLP),[147] or the Hydropathic INTeractions (HINT).[148] They are based on the concept that the spatial distribution of the empirically determined lipophilicity of molecules provides guidelines about the molecular determinants of ligand binding. The molecular lipophilicity potential (MLP) offers a quantitative 3D description of the lipophilicity potential to determine the hydrophobic pattern implicated in recognition of the biomolecular target. MLP combines fragment-based lipophilic contributions with distance-dependent function. This technique can be used in combination with 3D-QSAR or docking methods[149]. On the other hand, HINT provides an empirical and quantitative evaluation of the ligand–receptor complex as a sum of pairwise interactions between atomic hydrophobicities. Since these parameters are taken from experimental data of $\log P_{o/w}$, their use in diverse applications in biomolecular structure and drug discovery[150,151] has been easily considered.

Roger and Cammarata[152,153] proposed to rely on molecular properties derived from quantum mechanical descriptors instead of addressing determinations from an empirical perspective. They proposed to represent the partition coefficient by indices obtained from molecular orbital theory, particularly charge density and electrophilic superdelocalizability to represent the partitioning of aromatic molecules between nonpolar and polar phases. In a later approach, a nonlinear regressional model was presented for the estimation of octanol/water partition coefficients. The molecular surface, volume, weight, and charge densities on nitrogen and oxygen atoms of the molecule were the molecular descriptors employed to estimate of logP. All the descriptors were determined by using fully optimized structures based on AMI calculations[154]. Further studies have appeared that include more descriptors, and use alternative prediction systems such as regression models or neural networks.[155,156]

These efforts converged in the heuristic molecular lipophilic potential (HMLP).[157,158] HMLP is a structure-based technique requiring no empirical indices of atomic lipophilicity. In this model, the lipophilicity potential generated is derived from the electron density function and the electrostatic potential. The interactions of dipole and multipole moments, hydrogen bonds, and charged atoms in a molecule are included as hydrophilic interactions.

Alternatively, the computation of lipophilicity/hydrophilicity can have its origin on the description of the solvent as a continuum polarizable medium that reacts against the perturbing field created by the charge distribution of the solute.[159–161] These approaches are grouped under the heading of QM self-consistent reaction field (SCRF) methods, which provide a direct procedure to determine the solvation free energy, and hence the partition coefficient. This system offers the benefit of decomposing the total solvation free energy in atomic contributions, which allows us to carry out studies on the molecular determinants of bioactivity and be extended to studies of molecular similarity.[162–164]

In line with the criterion given in previous works of this extensive topic,[161,165] QM solvation models are classified into six categories, namely, (1) the apparent surface charge (ASC) methods, (2) the multipole expansion (MPE) methods, (3) the generalized Born approximation (GBA), (4) the image charge (IMC) methods, (5) the finite element methods (FEM), and (6) the finite difference methods (FDM). A complete description of these methods is beyond the purpose of this work, and we limit ourselves to address the reader to the reviews above cited[161,165] and to comparative studies of their performances[107,135]. Nevertheless, some aspects of the MST-SCRF model (included in the ASC methods), are discussed in the next section, since it is used o derived 3D distribution profile of lipophilicity for molecular similarity approaches.

1.2.1 The QM Continuum Solvation MST Model.

The Miertus-Scrocco-Tomasi (MST) model[165–167] is a reformulation of the formalism of dielectric polarizable continuous model (DPCM) optimized for organic and biological systems[168]. It has been parameterized at the HF/6-31G(d) level and with semiempirical AM1 and PM3 methods,[7,8,169,170] and to describe solvation in different solvents: water, dimethylsulfoxide, octanol, chloroform, and carbon tetrachloride.

This method provides fractional contributions to the solvation free energy for the reversible work necessary to transfer a molecule from the gas phase to a specific solvent at constant concentration, pressure, and temperature. As advantage, the influence of the whole molecule in the contribution of a given atom is considered.

The MST model computes the free energy of solvation as the sum of three contributions: cavitation (ΔG_{cav}), van der Waals (ΔG_{vW}), and electrostatic (ΔG_{ele}), which can be expressed as the sum of atomic contributions, eq 5. The first two components (ΔG_{cav} and ΔG_{vW}) are grouped in the "non-electrostatic contribution", in which the first term is the work required for creating a cavity shaped to accommodate the solute in the solvent and the second term accounts for dispersion-repulsion between solute and solvent. The third term (ΔG_{ele}) is responsible for the "electrostatic contribution", which measures the work needed to build up the solute charge distribution in the solvent (see ref.[5] for detailed review of MST model), Figure 3.

$$\Delta G_{sol} = \sum_{i=1}^{N} \Delta G_{sol,i} = \sum_{i=1}^{N} (\Delta G_{cav,i} + \Delta G_{vW,i} + \Delta G_{ele,i}) \qquad 5$$

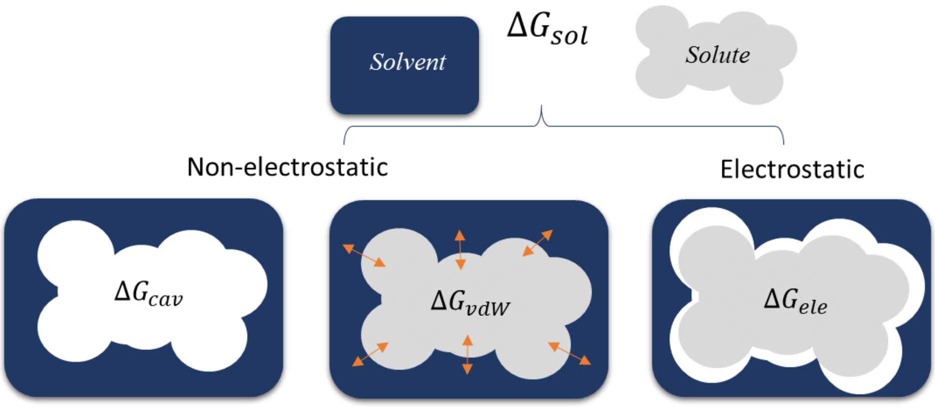


Figure 3 | Miertus-Scrocco-Tomasi (MST) Model: Framework for Continuum Solvation Calculations (ΔG_{sol}). The ΔG_{cav} is the work required for creating a cavity shaped to accommodate the solute in the solvent, ΔG_{vdW} term accounts for dispersion-repulsion (orange arrows), between solute and solvent, and ΔG_{ele} measures the work needed to build up the solute charge distribution in the solvent. See, eqs 6, 7, 8.

The cavitation free energy (ΔG_{cav}) is computed following Pierotti's scaled particle theory[171] adapted to molecular-shaped cavities according to the procedure proposed by Claverie.[172] In this model, the atomic cavitation and van der Waals free energy, non-electrostatic contributions, are computed according to:

$$\Delta G_{cav} = \sum_{i=1}^{N} \Delta G_{cav,i} = \sum_{i=1}^{N} \frac{S_i}{S_T} \Delta G_{P,i}^{o/w} \qquad 6$$

$$\Delta G_{vdW} = \sum_{i=i}^{N} \Delta G_{vdW,i} = \sum_{i=i}^{N} \Delta \xi_i^{o/w} S_i \qquad 7$$

where $\Delta G_{P,i}^{o/w} = \Delta G_{P,i}^{w} - \Delta G_{P,i}^{o}$ and its contribution is weighted by the ratio of the solvent-exposed surface (S_i) of atom i to the total surface (S_T), and $\Delta \xi_i^{o/w} = \Delta \xi_i^{w} - \Delta \xi_i^{o}$, where the atomic surface tension of atom i, ξ_i, is determined by fitting the experimental free energy of solvation.[7,173]

Otherwise, $\mathbf{\Delta G_{ele}}$, eq 7, encode electrostatic features of the molecule, eq 8.

$$\Delta G_{ele} = \sum_{i=1}^{N} \Delta G_{ele} = \sum_{i=1}^{N} \sum_{\substack{j=i \\ j \in i}}^{M} \Psi^{o} \frac{1}{2} \left\langle \left| \frac{q_j^{sol}}{|r_j - r_i|} \right| \Psi^{o} \right\rangle \qquad 8$$

where N is the total number of atoms, M is the total number of reaction field charges (q_j^{sol}, located at position r_j), and Ψ^{o} is the wave function of the solute in the gas phase.

1.2.2 3D lipophilic profile from MST calculations.

The hydrophobicity of a molecule is typically determined from the partitioning between octanol and water ($LogP_{o/w}$), which in turn is related to the free energy of transfer ($\Delta G_{o/w}$) of a given solute between these two solvents, eq 9. Therefore, hydrophobicity can be expressed in terms of the solvation free energy of the compound upon transfer from the gas phase to the water and organic phase, Figure 4.

$$\mathrm{logP_{o/w}} = -\frac{\Delta G_{tr}^{o/w}}{2.303\mathrm{RT}} = \frac{\Delta G_{tr}^{w} - \Delta G_{tr}^{o}}{2.303\mathrm{RT}} \qquad 9$$

where ΔG_w and ΔG_o denote the solvation free energy in water and octanol, respectively, and T is the temperature.

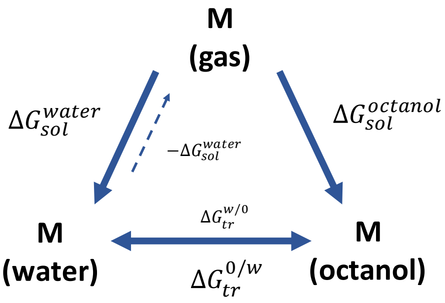


Figure 4 | Thermodynamic cycle for the determination of free energy of transfer of a molecule M between two immiscible solvents from the solvation free energies.

The decomposition scheme[174] formulated for the solvation free energy within the MST version of the PCM solvation model[5,8] offers a solution to define the hydrophobicity pattern of a molecule from the atomic contribution to $\mathrm{LogP_{o/w}}$ eqs 10 and 11. This arrangement allows us to evaluate the hydrophobic complementarity between a given molecule and its biological target via a fractional decomposition of $\mathrm{LogP_{o/w}}$ into atomic contributions.

$$\mathrm{logP} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{logP_{sol,i}} = \sum_{\mathrm{i=1}}^{\mathrm{N}} (\mathrm{logP_{cav,i}} + \mathrm{logP_{vdW,i}} + \mathrm{logP_{ele,i}}) \qquad 10$$

$$\mathrm{logP_X} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{logP_{X,i}} = \mathrm{logP_{X,i}} = \sum_{\mathrm{i=1}}^{\mathrm{N}} -\frac{\Delta \mathrm{G}_{\mathrm{X,i}}^{\frac{\mathrm{o}}{\mathrm{w}}}}{2.303\mathrm{RT}} \qquad 11$$

(x = ele, cav or vdW)

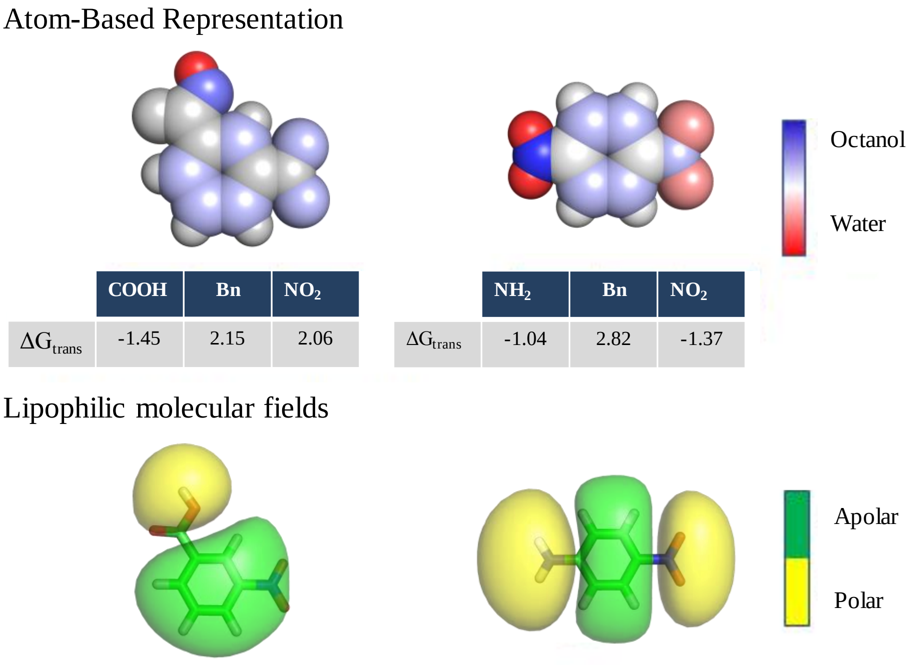
where N is the total number of atoms in the molecule, and $\Delta G_{w/o,i}$ is the atomic contribution of atom i to the transfer free energy from n-octanol to water ($\Delta G_{w/o,i} = \Delta G_{w,i} - \Delta G_{o,i}$).

MST-derived applications use the atomic contributions to the thermodynamic components of the differential solvation free energy in water and n-octanol. Accordingly, the computation of the 3D distribution pattern of molecular lipophilicity considers the effect of specific chemical features of the molecule, such as the existence of specific tautomers, conformational species, the formation of specific intramolecular interactions or the influence of other groups in one atom contribution, offering an advantage over experimental approaches, Figure 5.

These patterns have been previously exploited as logP descriptors to derive structure-activity relationchips[162,163,175], and play a decisive role in the development of PharmScreen atomic contributions.

The decomposition of the logP into three contributions: electrostatic ($logP_{ele}$), cavitation ($logP_{cav}$), and van der Waals ($logP_{vdW}$) allowed us the study of the relationships between the biological activity and the combination of the different descriptors. The fields derived from $logP_{cav}$ and $logP_{vW}$ are highly correlated[162,163], both contributions depend on the solute-exposed surface of atoms. They reflect the size and shape of the molecule, and therefore the information encoded for these descriptors is expected to relate to steric field.

Since the simultaneous inclusion of both non-electrostatic fields would be redundant and has been reported in previous works[162] that the best combinations of descriptors include the electrostatic contribution and one non-electrostatic representative. In particular, the best combination arises from the addition of $LogP_{ele}$ and $LogP_{cav}$ contribution,[162] which have been used by PharmScreen.



	COOH	Bn	NO_2
ΔG_{trans}	-1.45	2.15	2.06

	NH_2	Bn	NO_2
ΔG_{trans}	-1.04	2.82	-1.37

Figure 5 | The $\Delta G_{tranf\ o/w}$ of the nitro group change based on the other benzene substituent. Right, the nitro group is apolar (the other substituent acts an acceptor), left, the nitro group is polar (the other substituent acts as donor). $\Delta G_{tranf\ o/w}$ computed using MST model.

Previous studies had already addressed the use of these parameters in similarity applications, such as the self-hydrophobic similarities of molecular pairs were correlated with the inhibitory activity of a set of ACAT inhibitors and the binding affinities for a series of 5-HT3R agonist,[175] or the comparison of base pairs of nucleic acid bases with hydrophobic counterparts.[176]

2 CHAPTER TWO Objectives

2 OBJECTIVES

The use of computational methods for the search of active molecules is a core theme of chemo-informatics. One of its aims is to allow the high-throughput screening (HTS) of large libraries of compounds to identify potential hits against new pharmacological targets. Given the decreasing number of new drugs per billion US dollars spent on R&D approved by the Food and Drug Administration (FDA),[177] the efficiency halves every 9 years. The availability of efficient and comprehensive frameworks based on novel descriptors at the initial stages of drug discovery projects could alleviate this trend, complementing the output derived from traditional descriptors in LBVS tools.

In this context, this thesis proposes the development of a 3D VSLB tool (PharmScreen) to search for structurally diverse compounds with potential biological activity, contributing to reduce the high cost of experimental screening techniques. While molecular overlap and similarity measurements are traditionally accomplished by using approaches that primarily exploit shape, electrostatic, and pharmacophoric features, hydrophobicity, which plays a main role in pharmacodynamics and pharmacokinetics, has been relegated to the sidelines in VS methods.

Since atom- or fragment-based models have been used in VS tools[94,95] to obtain a qualitative picture of hydrophobic/hydrophilic areas, the main objective is to exploit the Miertus-Scrocco-Tomasi (MST) continuum solvation model, which relies on the integral equation formalism of the polarizable continuum model (IEFPCM), to account for the 3D lipophilic similarity between pairs of drug-like molecules.

With this general aim, the specific objectives of this work are indicated as follows:

1. Validation and development of a competent 3D alignment based on the partition of molecular lipophilicity into atomic contributions using the MST method.
2. To establish a balanced choice between accuracy and computational expensiveness to compute hydrophobic descriptors.
3. To calibrate the suitability of the MST-derived hydrophobic descriptors relative to traditional properties.
4. To examine the usefulness of the alignment descriptors to discern between active and inactive compounds.
5. To validate the lipophilic similarity framework developed as a competent tool for VS campaigns.

3 CHAPTER THREE Publications

3 PUBLICATIONS

Three papers have been compiled for the defense of this thesis. Each of them, preceded by an overview and a brief summary of the results and conclusions, are included in this chapter.

The first paper, entitled "*Development and Validation of Molecular Overlays Derived From 3D Hydrophobic Similarity with PharmScreen*", introduces PharmScreen in the scientific community as a new alignment tool. The overlap algorithm that exploits hydrophobic atomic contribution is presented and validated with the CCDC AstraZeneca Validation Overlays Data Test (more detail on the results summary in section 3.1)

In the second paper, entitled "*Lipophilicity in drug design: an overview of lipophilicity descriptors in 3D-QSAR studies*", encodes the use of quantum mechanical-based descriptors derived from continuum solvation models, as an open novel avenue for gaining insight into structure–activity relationships studies. In particular, the suitability of MST-based atomic lipophilicity contributions in combination with hydrogen bond pattern for 3D-QSAR studies is explored (in section 3.2).

The third paper, entitled "*Similarity assessment of lipophilic distribution: a boost for structure-based methods*", assesses the complementarity between 3D similarity using hydrophobic molecular profile derived from semi-empirical Quantum-Mechanical (QM) calculations and the scoring function of a wild accepted molecular docking package, Glide. Methodology development and validation details in section 3.3.

In addition, since the industrial framework on which this thesis has been developed, a patent, entitled "*Calculating molecular similarity*", was applied. In this document are described both the hydrophobic descriptors used to compute molecular similarity and the algorithm implemented to perform it (section 3.4).

3.1 PAPER1: *"Development and Validation of Molecular Overlays Derived From 3D Hydrophobic Similarity with PharmScreen"*

Javier Vazquez, †,‡ Alessandro Deplano, † Albert Herrero, † Tiziana Ginex, ‡ Enric Gibert, † Obdulia Rabal, § Julen Oyarzabal, § Enric Herrero, † and F. Javier Luque ‡

† Pharmacelera, Plaça Pau Vila, 1, Sector 1, Edificio Palau de Mar, Barcelona 08039, Spain

‡ Department of Nutrition, Food Science and Gastronomy, Faculty of Pharmacy and Food Sciences, Institute of Biomedicine (IBUB), and Institute of Theoretical and Computational Chemistry (IQTC-UB), University of Barcelona, Av. Prat de la Riba 171, Santa Coloma de Gramenet E-08921, Spain

§ Small Molecule Discovery Platform, Molecular Therapeutics Program, Center for Applied Medical Research (CIMA), University of Navarra, Avda. Pio XII 55, Pamplona E-31008, Spain

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: *J. Chem. Inf. Model.* 2018, 58, 1596−1609

pubs.acs.org/jcim

Development and Validation of Molecular Overlays Derived from Three-Dimensional Hydrophobic Similarity with PharmScreen

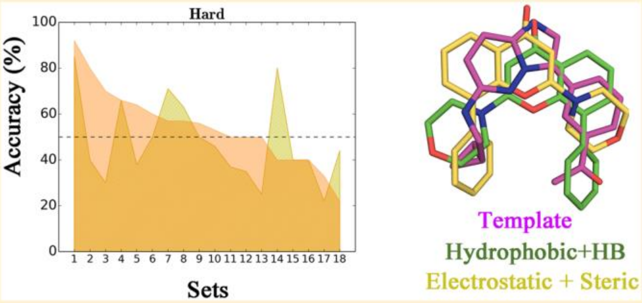
Javier Vázquez,[†,‡] Alessandro Deplano,[†] Albert Herrero,[†] Tiziana Ginex,[‡] Enric Gibert,[†] Obdulia Rabal,[§] Julen Oyarzabal,[§] Enric Herrero,[†] and F. Javier Luque*[,‡]

[†]Pharmacelera, Plaça Pau Vila, 1, Sector C 2a, Edifici Palau de Mar, Barcelona 08039, Spain

[‡]Department of Nutrition, Food Science and Gastronomy, Faculty of Pharmacy and Food Sciences, Institute of Biomedicine (IBUB), and Institute of Theoretical and Computational Chemistry (IQTC-UB), University of Barcelona, Av. Prat de la Riba 171, Santa Coloma de Gramenet E-08921, Spain

[§]Small Molecule Discovery Platform, Molecular Therapeutics Program, Center for Applied Medical Research (CIMA), University of Navarra, Avda. Pio XII 55, Pamplona E-31008, Spain

Supporting Information



ABSTRACT: Molecular alignment is a standard procedure for three-dimensional (3D) similarity measurements and pharmacophore elucidation. This process is influenced by several factors, such as the physicochemical descriptors utilized to account for the molecular determinants of biological activity and the reference templates. Relying on the hypothesis that the maximal achievable binding affinity for a drug-like molecule is largely due to desolvation, we explore a novel strategy for 3D molecular overlays that exploits the partitioning of molecular hydrophobicity into atomic contributions in conjunction with information about the distribution of hydrogen-bond (HB) donor/acceptor groups. A brief description of the method, as implemented in the software package PharmScreen, including the derivation of the fractional hydrophobic contributions within the quantum mechanical version of the Miertus−Scrocco−Tomasi (MST) continuum model, and the procedure utilized for the optimal superposition between molecules, is presented. The computational procedure is calibrated by using a data set of 402 molecules pertaining to 14 distinct targets taken from the literature and validated against the AstraZeneca test, which comprises 121 experimentally derived sets of molecular overlays. The results point out the suitability of the MST-based hydrophobic parameters for generating molecular overlays, as correct predictions were obtained for 94%, 79%, and 54% of the molecules classified into easy, moderate, and hard sets, respectively. Moreover, the results point out that this accuracy is attained at a much lower degree of identity between the templates used by hydrophobic/HB fields and electrostatic/steric ones. These findings support the usefulness of the hydrophobic/HB descriptors to generate complementary overlays that may be valuable to rationalize structure−activity relationships and for virtual screening campaigns.

■ INTRODUCTION

The assumption that structurally similar molecules have similar biological activities has been widely exploited in chemical informatics and drug discovery.[1−4] This premise underlies most practical applications in chemical and pharmaceutical research, such as the identification of new candidate compounds in screening studies through similarity searching against known actives. However, the concept of molecular similarity is subjective,[5] and its quantification depends on the representation of the chemical features present in the compounds by means of 1D, 2D, or 3D descriptors, the weighting of these descriptors, and the mathematical expression of the similarity function.

3D-based similarity methods rely on the molecular geometry, which can be used in different ways through non-super-positional and superpositional methods.[6−8] The former involves the analysis of atomic distances to a set of reference positions, as implemented in the Ultrafast Shape Recognition (USR) method,[9] which has been adapted to include other molecular properties.[10,11] Superpositional methods involve the overlay of compounds in a process intended to maximize overlap of molecular shape and/or pharmacophoric features. Shape-guided similarity can be achieved by exploiting specific properties of the molecular surface, as implemented in Surflex-Sim[12] and SURFCOMP,[13] or alternatively through the representation of the molecular volume with hard spheres, such as Phase Shape,[14] or Gaussian functions, which are used in ROCS (Rapid Overlay of Chemical Structures),[15,16] ShaEP (Shape and Electrostatic Potential),[17] SHAFTS (Shape-

Received: April 10, 2018
Published: July 16, 2018

DOI: 10.1021/acs.jcim.8b00216

Feature Similarity),[18] and LIGSIFT.[19] These latter methods supplement the shape-based similarity by mapping chemical groups into pharmacophoric features.[20−22] On the other hand, molecular fields[23] represent a distinct approach wherein comparison relies on the spatial variation of interaction energies with probes, as implemented in FieldScreen[24] and FLAP.[25]

Shape and electrostatics dominate the realm of chemical descriptors used in 3D-based similarity.[26] However, this hides the fundamental role played by other contributions to the binding affinity, such as the (de)solvation of both ligand and receptor.[27] At this point, it is worth noting that approximate models for estimating the maximal achievable affinity of target binding sites for drug-like compounds have shown the relevance of nonpolar desolvation.[28] This is consistent with studies that support the concept that favorable drug binding is largely driven by the hydrophobic effect,[29−32] whereas polar interactions provide "anchor points" contributing to ligand specificity and/or directionality in the binding pocket,[31] and modulate ligand binding kinetics.[33] In this context, one may question whether hydrophobicity alone may encode valuable information to guide similarity measurements between molecules.

While lipophilicity is of paramount importance for drug pharmacokinetics,[34] it has been rarely used as the primary descriptor in understanding ligand−receptor recognition. Elaborate implementations of this concept are the Molecular Lipophilicity Potential (MLP)[35] and the Hydropathic INTERaction (HINT) score.[36,37] The MLP offers a quantitative 3D description of the lipophilicity from all the molecular fragments on the surrounding space of a compound and has found applications in 3D-QSAR and docking.[38] On the other hand, HINT provides an empirical, but quantitative, evaluation of the ligand−receptor complex as a sum of pairwise interactions between atomic hydrophobicities. Since these parameters are taken from experimental data of the octanol/water partition coefficient ($\log P_{o/w}$), both enthalpy and entropy contributions are accounted for by the HINT score, which has been used in diverse applications in biomolecular structure and drug discovery.[39−41]

Here we present a novel strategy to evaluate molecular similarity from 3D distribution maps of hydrophobicity and to guide the overlay of compounds according to hydrophobic topology. Instead of using empirical data, as implemented in MLP and HINT, the method exploits hydrophobic maps estimated from quantum mechanical (QM) theoretical calculations of the differential solvation of a solute in water and *n*-octanol. These calculations are performed within the framework of self-consistent reaction fields methods and, particularly, the Miertus−Scrocco−Tomasi (MST) continuum solvation method.[42,43] It is noteworthy that the derivation of 3D distribution maps of the global $\log P_{o/w}$ is facilitated by the partition of the solvation free energies into atomic contributions following a perturbative treatment of the electrostatic coupling between solute and solvent and taking advantage of the dependence of nonelectrostatic contributions on the solvent-exposed surface of atoms. As an advantage, this method provides fractional contributions to the lipophilicity that incorporate the influence of the whole molecule in the contribution of a given atom. Moreover, they do not depend on the existence of suitable parameters for new chemical groups, which might not be present in the empirical databases. The algorithm has been implemented in a new tool called PharmScreen, which follows the successful application of these atomic contributions to the analysis of hydrophobic pharmacophores, which were found to have predictive potential comparable to other standard 3D-QSAR techniques.[44,45] The main application of PharmScreen is the 3D ligand-based virtual screening against single or multiple template targets, while exploiting a pregenerated ensemble of conformers for flexible compounds. The alignment to the template is completed by superimposition of molecular moments of the 3D hydrophobic distribution but may be subsequently refined by means of Monte Carlo sampling. The method has been calibrated by using a diverse set of ligands taken from known crystallographic complexes and further validated against the AstraZeneca benchmarking set, which contains 121 experimentally derived molecular overlays spanning across multiple protein families.[46] Finally, molecular overlays obtained from the hydrophobic contributions are discussed in light of the results obtained using standard electrostatic and steric fields, which are widely used in 3D molecular alignment studies.

METHODS

Derivation of 3D Atomic Hydrophobicity Maps. In the MST method, 3D hydrophobicity maps can be determined from the partition of the overall molecular hydrophobicity (estimated from the calculated $\log P_{o/w}$) into atomic contributions ($\log P_{o/w,i}$; eq 1), which in turn stem from the combination of atomic contributions to the solvation in water ($\Delta G_{w,i}$) and *n*-octanol ($\Delta G_{o,i}$).

$$\log P_{o/w} = \sum_{i=1}^{N} \log P_{o/w,i} = \sum_{i=1}^{N} \frac{\Delta G_{w/o,i}}{2.303RT} \tag{1}$$

where N is the total number of atoms in the molecule, and $\Delta G_{w/o,i}$ is the atomic contribution of atom i to the transfer free energy from *n*-octanol to water ($\Delta G_{w/o,i} = \Delta G_{w,i} - \Delta G_{o,i}$).

This decomposition scheme has been presented elsewhere,[47] and here we limit ourselves to remark the essential details needed for the overlay protocol described below. The solvation free energy is obtained by adding electrostatic (ΔG_{ele}) and nonelectrostatic (cavitation, ΔG_{cav}, and van der Waals, ΔG_{vW}) components. The work needed to build up the solute charge distribution in the solvent (ΔG_{ele}) is calculated from the interaction between the solvent reaction field (represented by a set of charges q_j spread on the surface of the solute/solvent boundary) and the polarized charge distribution of the solute. By using a perturbative treatment,[48] ΔG_{ele} can be expressed as the addition of atomic contributions ($\Delta G_{ele,i}$) generated from the interaction of the nonpolarized solute and the subset of point charges placed on the cavity surface of a given atom, and the electrostatic contribution to $\log P_{o/w}$ can be computed as noted in eq 2.

$$\log P_{o/w}^{ele} = \sum_{i=1}^{N} \log P_{o/w,i}^{ele} = \frac{1}{2}\left\langle \Psi^{o} \left| \sum_{\substack{k=1\\k\in i}}^{k} \frac{q_k^{w}}{|r_k^{w} - r|} - \sum_{\substack{l=1\\l\in i}}^{L} \frac{q_l^{o}}{|r_l^{o} - r|} \right| \Psi^{o} \right\rangle \tag{2}$$

where K and L stand for the total number of reaction field charges in water (q_k^w) and *n*-octanol (q_l^o), located at positions

$r_k^{\rm w}$ and $r_l^{\rm o}$, respectively (note that a distinct solvent-dependent boundary is used in the MST model for the two solvents),[47] and $\Psi^{\rm o}$ is the wave function of the solute in the gas phase. Therefore, $\log P_{\rm o/w,\it i}^{\rm ele}$ accounts for the atomic contribution to the differential electrostatic (free) energy due to the interaction of the whole solute with the reaction field charges spread onto the surface patch of atom *i*.

With regard to the nonelectrostatic terms, $\Delta G_{\rm cav}$ is determined following Pierotti's scaled particle theory[49] adapted to molecular-shaped cavities,[50] and $\Delta G_{\rm vW}$ is computed using a linear relationship to the solvent-exposed surface of each atom. Therefore, cavitation and van der Waals contributions to log $P_{\rm o/w}$ (eqs 3 and 4) permit a straightforward decomposition into atomic components depending on the contribution of a given atom to the molecular surface.

$$\log P_{\rm o/w}^{\rm cav} = \sum_{i=l}^{N} \log P_{{\rm o/w},i}^{\rm cav} = \sum_{i=l}^{N} \frac{S_i}{S_{\rm T}} \Delta G_{P,i}^{\rm o/w} \tag{3}$$

$$\log P_{\rm o/w}^{\rm vW} = \sum_{i=l}^{N} \log P_{{\rm o/w},i}^{\rm vW} = \sum_{i=l}^{N} S_i \Delta \xi^{\rm o/w} \tag{4}$$

where $\Delta G_{P,i}^{\rm o/w} = \Delta G_{P,i}^{\rm w} - \Delta G_{P,i}^{\rm o}$ and its contribution is weighted by the ratio of the solvent-exposed surface (S_i) of atom *i* to the total surface ($S_{\rm T}$), and $\Delta\xi^{\rm o/w} = \xi_i^{\rm w} - \xi_i^{\rm o}$, where the atomic surface tension of atom *i*, ξ_i, is determined by fitting the experimental free energy of solvation.[42,43]

Overall, the hydrophobicity of a molecule can be partitioned into atomic contributions, each decomposable into electrostatic ($\log P_{{\rm o/w},i}^{\rm ele}$), cavitation ($\log P_{{\rm o/w},i}^{\rm cav}$), and van der Waals ($\log P_{{\rm o/w},i}^{\rm vW}$) components (eq 5).

$$\begin{aligned} \log P_{\rm o/w} &= \sum_{i=l}^{N} \log P_{{\rm o/w},i} \\ &= \sum_{i=l}^{N} (\log P_{{\rm o/w},i}^{\rm ele} + \log P_{{\rm o/w},i}^{\rm cav} + \log P_{{\rm o/w},i}^{\rm vW}) \end{aligned} \tag{5}$$

Molecular Fields. In this context, molecular overlays may be guided by the similarities between the molecular fields generated from the projection of $\log P_{{\rm o/w},i}^{X}$ (*X*: ele, cav, vW) contributions in the 3D space around the molecules. However, the usage of these descriptors must be performed subject to two considerations:

(i) While there is little redundancy between electrostatic and nonelectrostatic components, there is a large correlation between $\log P_{{\rm o/w},i}^{\rm cav}$ and $\log P_{{\rm o/w},i}^{\rm vW}$, as expected from their dependence on the solvent exposure of atoms (eqs 3 and 4).[44] Since the simultaneous inclusion of these fields would be highly redundant, molecular overlays have been determined from the $\log P_{{\rm o/w},i}^{\rm cav}$ contributions alone, which would contain information about the size and shape of the molecule.

(ii) The atomic contribution to log $P_{\rm o/w}$ and $\log P_{\rm o/w}^{\rm ele}$ of polar atoms is negative, reflecting the tendency to be better solvated in water than in *n*-octanol. While the magnitude of $\log P_{{\rm o/w},i}$ and $\log P_{{\rm o/w},i}^{\rm ele}$ reflects the polarity of the corresponding atom, it does not contain information about its hydrogen-bond (HB) donor/acceptor character, which may be expected to be crucial for attaining a proper molecular alignment. As an example, Figure 1 shows the negative contributions of N, NH, and NH_2 determined for adenine, and for the NH_3^+ and COO^- groups in zwitterionic glycine. In the absence of information about the HB signature, self-alignment would lead to chemically counterintuitive overlays, such as the superposition of the carboxylate unit onto the protonated amine of glycine upon rotation around the bisector of the backbone CH_2 group, or the superposition of the NH and NH_2 donor groups onto the N acceptors upon rotation along the longitudinal axis of adenine. Therefore, the explicit addition of a HB field is necessary to preserve the proper HB recognition pattern of molecules.

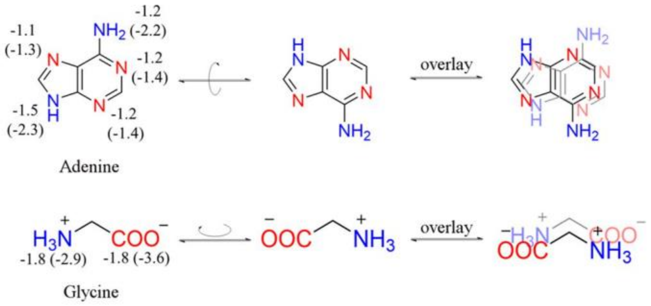


Figure 1. Contributions of selected polar atoms or functional groups (HB donor/acceptor groups shown in blue/red, respectively) to the molecular log $P_{\rm o/w}$ and $\log P_{\rm o/w}^{\rm ele}$ (in parentheses) for adenine and the zwitterionic form of glycine. Using these descriptors alone, self-alignment could lead to counterintuitive molecular overlays where HB donor/acceptor groups are superposed (right side of the plot).

On the basis of these considerations, molecular overlays have been determined using two or three molecular fields, which combine the atomic contributions to (i) the total log $P_{\rm o/w}$, and (ii) the electrostatic ($\log P_{\rm o/w}^{\rm ele}$) and cavitation ($\log P_{\rm o/w}^{\rm cav}$) contributions, supplemented in both cases with a HB field. The hydrophobic descriptors were obtained by using the MST solvation model parametrized for the semiempirical Hamiltonian RM1.[51,52] Choice of this level of theory was motivated by its low computational cost compared to *ab initio* methods. Nevertheless, to evaluate the influence of the QM method used to derive the hydrophobic descriptors, additional computations were performed using the MST version[42,43] parametrized at the B3LYP/6-31G(d) level. Calculations were performed using locally modified versions of MOPAC[53] and Gaussian 09.[54] With regard to the HB field, donor and acceptor sites were identified based on the classification of the various functional groups present in the molecule, with the subsequent assignment of an arbitrary parameter of +1 for all hydrogen atoms in HB donors, and −1 for N and O atoms that may act as acceptors. This description of the HB features of a molecule is simpler than more elaborate approaches that take into account experimental distributions of hydrogen-bonded atoms,[55] empirical scales,[56,57] or parameters derived from QM calculations.[58,59] Nevertheless, it is worth noting that the strength of the polar character is already contained in the magnitude of the atomic hydrophobic contribution ($\log P_{{\rm o/w},i}$, $\log P_{{\rm o/w},i}^{\rm ele}$) of donor/acceptor groups. Therefore, this should suffice to provide a parameter suitable to distinguish the HB signature of compounds.

Molecular Overlays. An initial set of alignments is generated from the set of molecular moments that describe the 3D hydrophobic distribution in a compound. The definition of these moments was inspired on the multipolar expansion of the electrostatic potential (see below).[60,61] The pool of initial alignments is subsequently used to identify the most suitable overlay according to a score function that takes into account the similarity of molecular fields. Finally, refinement of the overlaid pose is accomplished through a Metropolis Monte Carlo algorithm.

Multipole Expansion of log $P_{o/w}$. Since the atomic log $P_{o/w,i}$ contributions comprise positive/negative values that denote the apolar/polar character of atoms, the 3D hydrophobic distribution is described with respect to the center of the "hydrophobic monopole" ($\vec{R}_m$; eq 6).

$$\vec{R}_m = \frac{\sum_{i=l}^{N} \log P_{o/w,i}\vec{r}_i}{\log P_{o/w}} \tag{6}$$

where $\vec{r}_i$ denotes the position of atom i.

This definition minimizes the contribution of the "hydrophobic dipole", which is zero, and the first nonvanishing term is the "hydrophobic quadrupole" ($\boldsymbol{Q}$; eq 7). It defines two independent principal values (note that the quadrupole tensor is traceless) and three principal axes, which represent the canonical axes that define the molecular orientations of the compound and are invariant to the translation of the molecule.

$$\boldsymbol{Q} = \sum_{i=l}^{N} \log P_{o/w,i}(3\vec{r}_i\vec{r}_i - |\vec{r}_i|^2 1) \tag{7}$$

For compounds with log $P_{o/w}$ equal to zero, the leading term is the hydrophobic dipole, and hence the center of expansion is defined as the center of dipole, which minimizes the contribution of the quadrupolar term. The dipole direction coincides with one of the quadrupolar principal axes, which has a null principal value. Therefore, the quadrupolar tensor yields an orthogonal set of principal axes that can be used for molecular alignment.

Multipole Expansion of log $P_{o/w}^{ele}$. The atomic log $P_{o/w,i}^{ele}$ contributions may adopt zero or negative values due to the more favorable electrostatic interaction arising upon hydration compared to solvation in n-octanol. Then, the molecular alignment is performed using the hydrophobic quadrupole, following the same formalism described above for log $P_{o/w}$.

Multipole Expansion of log $P_{o/w}^{cav}$. In this case, the alignment is accomplished through calculation of the moments of inertia ($\boldsymbol{I}$; eq 8) obtained from the atomic log $P_{o/w,i}^{cav}$ contributions, which upon diagonalization provides the principal axes of rotation.

$$\boldsymbol{I} = \sum_{i=l}^{N} \log P_{o/w,i}^{cav}(|\vec{r}_i|^2 1 - \vec{r}_i\vec{r}_i) \tag{8}$$

Score Function. For each molecular alignment, field values are computed by projecting the atomic (hydrophobic + HB) contributions into a 3D grid using an exponential function ($p(q)$; eq 9) as implemented in COMSIA.[62]

$$p(q) = \sum_{i=l}^{N} w_i e^{-\alpha r_{iq}^{\ 2}} \tag{9}$$

where w_i is the actual value of the atomic property of atom i, α is the attenuation factor, which was set to 0.3,[44] and r_{iq} is the distance between the grid point q and atom i.

The similarity between each projected field (k) is evaluated using the Tanimoto coefficient (T_k), and the global score (S) is determined by combining the Tanimoto index obtained for the distinct molecular fields using normalized weighting factors (λ_k; eq 10).

$$S = \sum_{k} \lambda_k T_k \tag{10}$$

Molecular Systems. Calibration of the weighting factors was achieved using a training set that consists of 14 molecular systems (Table 1). A subset of nine systems (subset I) containing considered X-ray data on 168 ligand−protein complexes was used as a benchmark data set in previous studies.[63,64] To obtain the reference structure for molecular overlays, X-ray structures retrieved from the Protein Data Bank[65] were cleaned, leaving only the protein and the ligand of interest, and in the case of structures containing multiple chains, only chain A was retained. Then, a multiple protein alignment was performed for each target using PyMOL,[66] and the aligned ligands were extracted and used as reference structures. With the exception of cyclin dependent kinase 2 (CDK2) and human immunodeficiency virus (HIV), these systems contain a limited number of compounds (between 7 and 13). Therefore, we also included a second subset of five systems (subset II) containing between 32 and 74 structurally related molecules, which were aligned by using X-ray crystallographic data and pharmacophoric constraints in previous studies (see ref 44 and references therein).

Table 1. Molecular Systems Considered As a Training Set

subset	system	number of molecules
I[a]	cyclin dependent kinase 2 (CDK2)	57
	elastase	7
	estrogen receptor (ER)	13
	human immunodeficiency virus (HIV; 1)	28
	mitogen activated protein kinase 14 (MAPK14)	13
	rhinovirus	8
	thermolysin (1)	12
	trypsin	7
	human immunodeficiency virus (HIV; 2)[b]	10
II[c]	chromanone	34
	cruzain	32
	dopamine D2/D4	41
	GSK-3β	74
	thermolysin (2)	74

[a]From ref 63. [b]From ref 64. [c]From ref 44 and references therein.

Validation of PharmScreen was subsequently performed using the AstraZeneca Overlays Validation Test Set (AZ test).[46] It contains 121 experimentally derived molecular overlays from 119 targets. The targets were classified into four categories (easy, moderate, hard, and unfeasible, comprising 22, 73, 18, and 8 systems, respectively) according to the expected difficulty to reproduce the experimental overlay.[67] The categorization of the AZ test set was made according to three parameters: (i) the average shape match, calculated on all the possible pairwise combinations of ligands within a set, (ii) the average Color score (i.e., a measure of overlap for predefined pharmacophore points obtained by aligning groups with the related properties), which accounts for feature similarity, and (iii) the average Tanimoto coefficient used to measure 2D fingerprint similarity. Then, using the Borda tallies determined for the three parameters, a consensus ranking scale was defined for classification of the targets into easy, moderate, hard, and unfeasible sets (the lower the Borda tally, the easier the prediction for the compounds pertaining to a given target).

The list of targets is provided in Supporting Information Table S1. Let us note that some targets, like trypsin, elastase, cruzain, and p38 MAP kinase, are present both in the training set and AZ test set. However, the overlap between the compounds pertaining to these targets is low: no overlap for trypsin and elastine, a single case (X-ray entry 3I06) for cruzain, and three cases (X-ray entries 1M7Q, 1OZL, and 1YQJ) for p38 MAP kinase.

All the molecular overlays were obtained for the molecules in aqueous solution at neutral pH. To this end, they were prepared with OpenBabel[68] and the final state was checked in all cases. Finally, to avoid any influence of the initial alignment on the molecular overlay, the molecules were randomly positioned in space without changing the internal geometry.

Performance Evaluation. Previous studies have indicated that the success in predicting molecular overlays is influenced by the choice of the template, the degree of similarity between compounds and templates, and the accuracy of the template's conformation.[64] Accordingly, the overall performance can be affected by the conformational aspects of ligands and the intrinsic accuracy of the overlay methods. For this reason, this study is focused on the X-ray conformations of the compounds, which permits to concentrate the comparative analysis in the intrinsic features of the molecular descriptors. Two metrics were used to check the performance of PharmScreen. Following Chen et al.,[64] every single molecule was used as template for the alignment of the remaining molecules in the corresponding set. The molecular overlay was considered to be correct when the root-mean-square deviation (RMSD) of the heavy atoms was less than or equal to 2.0 Å from the X-ray arrangement. The performance was then calculated as the *average* value (in percentage) of correct overlays found for all molecule-template pairs in each set. Alternatively, the accuracy in predicting the molecular overlay for a given compound was determined by taking into account only the pose with the *best* similarity score, irrespective of the template utilized in the molecular alignment.

Comparison against Electrostatic/Steric Fields. Finally, to further assess the suitability of the hydrophobic descriptors, molecular overlays were also obtained by combining electrostatic and steric fields. The aim was 2-fold: (i) to check the influence of the molecular descriptors on the number of successful overlays and (ii) to verify whether the two sets of descriptors lead to identical molecular alignments. For the sake of comparison, the electrostatic field was determined from the atomic partial charges obtained by fitting the semiempirical RM1 electrostatic potential calculated around the molecule using the NDDO-based strategy.[69,70] On the other hand, the steric field was calculated by using the cube of the atomic radii[71] (taken from the Tripos MMFF94 force field).[72] Furthermore, the initial alignment of the molecules was performed following the same formalism described above for the hydrophobic descriptors (i.e., multipolar expansion for atomic charges, and moments of inertia for the atomic radii) and the molecular overlays were refined using a Metropolis Monte Carlo algorithm.

RESULTS AND DISCUSSION

Calibration of Weighting Factors in the Similarity Function. Calibration of the similarity score was performed with the aim of deriving optimal weighting factors of the molecular fields using the training set. Table 2 reports the results obtained for the combination of total log $P_{\mathrm{o/w}}$ and HB descriptors, which were weighted by factors of 70% and 30%, respectively. Inclusion of the HB distribution retains or improves the average accuracy of the molecular overlays compared to the performance obtained exclusively from the atomic log $P_{\mathrm{o/w}}$ contributions. The improvement is largely dependent on the specific set of compounds, as it is primarily observed for the molecules pertaining to HIV and thermolysin in subset I, and thermolysin in subset II. Nevertheless, this effect is less apparent when the alignments with the best similarity are considered, which reflects the relevant influence exerted by the template on the molecular alignment. On the other hand, the overall performance is only slightly affected by the specific weight of the HB field (see Table S2 in the Supporting Information).

Table 2. Weighting Factors Chosen for Molecular Overlays upon Combination of Total Hydrophobicity and HB Fields[a]

		weight (log $P_{\mathrm{o/w}}$/HB)			
		100/0		70/30	
subset	system	*average*	*best*	*average*	*best*
I	CDK2	9	64	11	63
	elastase	14	0	14	0
	ER	30	76	33	76
	HIV (1)	17	53	22	57
	MAP14	21	61	21	61
	rhinovirus	72	100	72	100
	thermolysin (1)	21	50	33	50
	trypsin	55	57	55	57
	HIV (2)	26	50	30	50
	total (155)	20.8	59.4	24.0	60.0
II	chromanone	92	100	88	100
	cruzain	68	96	69	96
	dopamine	22	90	21	90
	GSK-3β	53	100	53	100
	thermolysin (2)	23	87	27	87
	total (255)	46.0	94.1	52.9	94.1
I + II	total (410)	36.7	81.0	41.9	81.1

[a] *Average* value (%) of successful overlaps for all ligand-template pairs, and *best* value (%) obtained from the pose with the highest similarity score for each molecule.

The optimal weights obtained for the combination of electrostatic ($\log P^{\mathrm{ele}}_{\mathrm{o/w}}$) and cavitation ($\log P^{\mathrm{cav}}_{\mathrm{o/w}}$) components of the hydrophobicity with the HB field are reported in Table 3. Preliminary analysis led to an optimal weighting of $\log P^{\mathrm{ele}}_{\mathrm{o/w}}$ and $\log P^{\mathrm{cav}}_{\mathrm{o/w}}$ fields close to 30/70 (see Table S3 in Supporting Information), which was subsequently refined upon inclusion of the HB field. Upon inclusion of the HB field, the largest accuracy was obtained for weighting factors of 15 ($\log P^{\mathrm{ele}}_{\mathrm{o/w}}$), 55 ($\log P^{\mathrm{cav}}_{\mathrm{o/w}}$), and 30 (HB), although the overall performance was only slightly affected by the specific weights of these fields (see Table S4–S5 in the Supporting Information).

Comparison of Tables 2 and 3 shows that replacement of the total log $P_{\mathrm{o/w}}$ contribution by its electrostatic and cavitation components ameliorates the overall accuracy for subsets I and II, increasing the average accuracy in the predicted molecular overlays from 24.0% to 38.0% for subset I, and from 52.9% to 65.6% for subset II. This trait is also observed using the best score, as the accuracy is enlarged from 60.0% to 78.7% for the

Table 3. Weighting Factors Chosen for Molecular Overlays upon Combination of Electrostatic and Cavitation Components of the Hydrophobicity and HB Fields[a]

		weight (log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB)			
		30/70/0		15/55/30	
subset	system	*average*	*best*	*average*	*best*
I	CDK2	18	85	22	87
	elastase	39	42	29	42
	ER	44	100	43	92
	HIV (1)	45	64	47	67
	MAP14	20	61	24	84
	rhinovirus	61	87	81	100
	thermolysin (1)	30	58	35	66
	trypsin	59	71	73	85
	HIV (2)	64	50	66	60
	total (155)	34.1	73.6	38.0	78.7
II	chromanone	94	100	93	100
	cruzain	97	100	95	100
	dopamine	42	100	42	100
	GSK-3β	75	100	75	100
	thermolysin (2)	37	95	44	94
	total (255)	64.0	98.5	65.6	98.2
I + II	total (410)	52.7	89.1	54.0	90.80

[a] *Average* value (%) of successful overlaps for all ligand-template pairs, and *best* value (%) obtained from the pose with the highest similarity score for each molecule.

compounds in subset I, although this improvement is lower in the series of compounds included in the subset II (from 94.1% to 98.2%), which can be ascribed to the larger congeneric character of the compounds included in these molecular sets. On the other hand, the alignment results are very sensitive to the templates, as noted in the range of the average accuracy, which varies from 22% for CDK2 to 81% for rhinovirus in subset I, and from 42 for dopamine D2/D4 antagonists to 95% for cruzain inhibitors (Table 3).

Figure 2 shows the comparison of the average value (%) of molecules correctly superposed for the distinct proteins of the training set. A relevant trait is the consistent improvement in the number of correctly predicted overlays observed in all cases upon decomposition of the total log $P_{o/w}$ into log $P^{ele}_{o/w}$ and log $P^{cav}_{o/w}$, which is larger than 15% in 9 sets. Indeed, the combination of log $P^{ele}_{o/w}$ and log $P^{cav}_{o/w}$ performs significantly better ($p < 0.05$) in 10 out of 14 sets, and the statistical significance is even higher ($p < 0.001$) in 7 targets.

Overlays from RM1 and B3LYP Hydrophobic Contributions. In order to evaluate the influence of the QM method used to derive the hydrophobic contributions, the accuracy of the molecular overlays obtained from fractional contributions obtained from MST/RM1 and MST/B3LYP computations was examined. The overall performance of the molecular overlays obtained from these computations is shown in Figure 3, which indicates that similar results are obtained from RM1 and B3LYP fragmental contributions for all data sets. Indeed, the slight differences observed in few cases are not found to be statistically significant.

This finding is remarkable when one takes into account the important saving in computational time due to the use of semiempirical calculations, as noted in a reduction by a factor of ~250 in the time required for the computation of atomic contributions for the ligand in PDB entry 1AQ1 (CDK2 set; RM1 calculations performed in an Intel Xeon CPU ES-2666v3 at 2.9 GHz, and B3LYP/6-31G(d) computations carried out in an Intel Xeon CPU E5645 at 2.4 GHz; the factor of ~250 was adjusted taking into account the differences in the clock computer cycle). Therefore, these findings support the suitability of the semiempirical RM1 Hamiltonian, which offers a much better balance between overlay accuracy and computational expensiveness.

Validation with the AstraZeneca Data Set. The AZ test comprises 121 molecular systems encoded in four categories (easy, moderate, hard, and unfeasible) based on how easy or difficult it would be for a program to reproduce the experimental overlay.[46] The accuracy of the molecular overlays predicted from log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB fields (weights of 15/55/30, respectively) is shown in Figure 4, which shows the overlay accuracy determined using the poses with the best similarity score. For the sake of comparison, Figure 4 also displays the results obtained by using electrostatic/steric fields (see Methods). Let us note that the weights of these two fields were previously adjusted using the compounds included in the training set, leading to contributions of 50% for each field (see Table S6 in the Supporting Information).

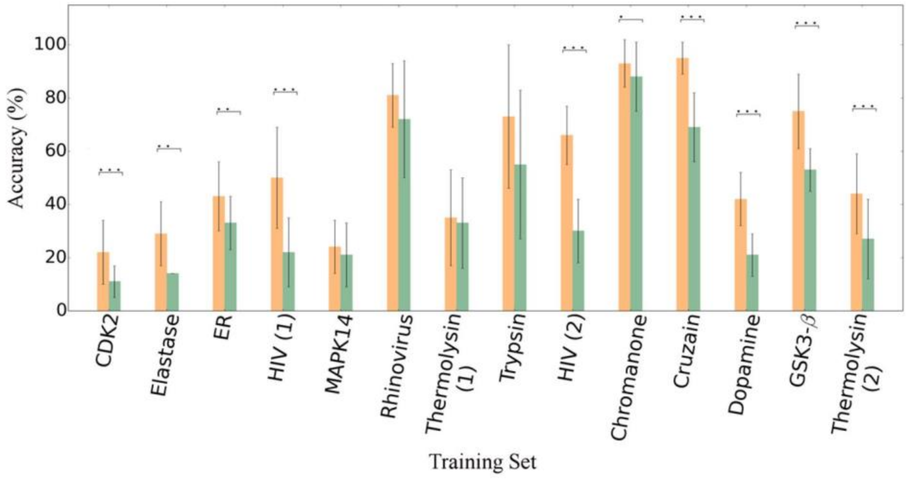


Figure 2. Number (%) of correctly predicted overlays obtained from log $P_{o/w}$/HB (weights 70/30; green) and log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB (weights 15/55/30; orange) fields. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

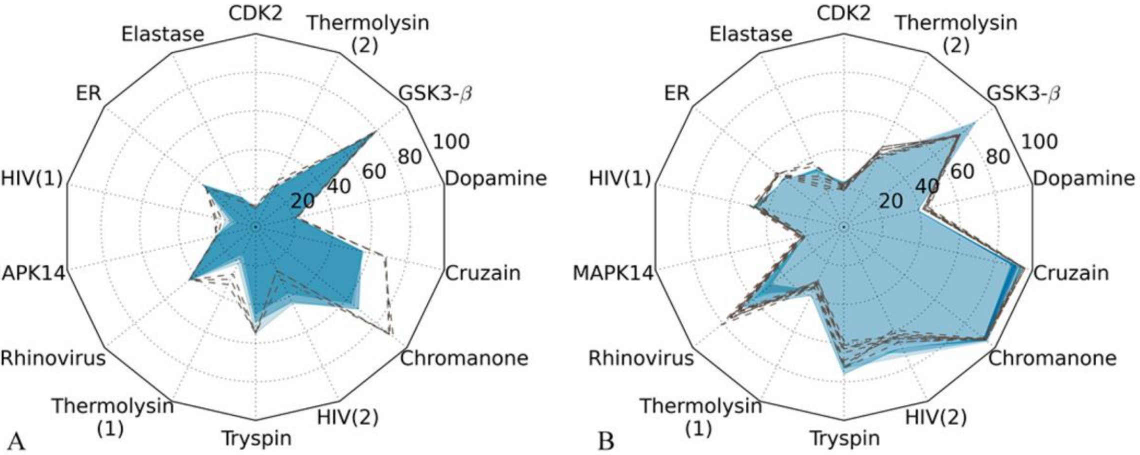


Figure 3. Overlay accuracy (%) determined for (A) log $P_{o/w}$/HB (weights 70/30) and log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB (weights 15/55/30) fields. The average values of correctly predicted alignments from semiempirical RM1 and B3LYP hydrophobic contributions are shown as dashed lines and solid area, respectively.

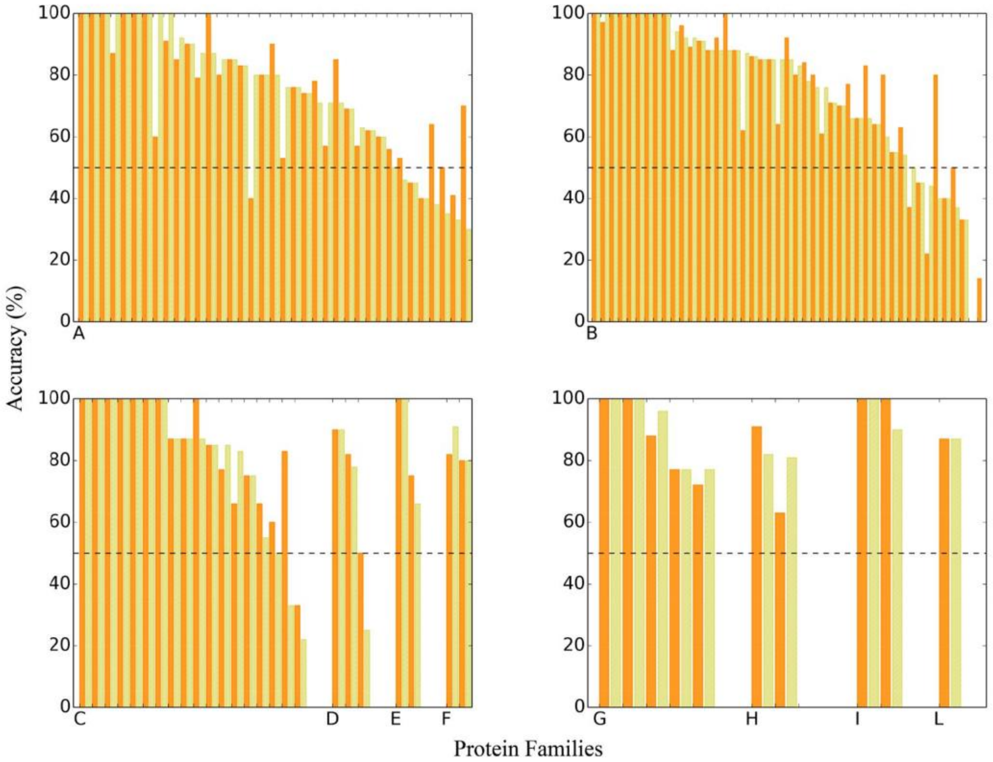


Figure 4. Comparison of experimental and predicted molecular overlays by target family. Accuracy (%) of molecular overlays predicted from log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB (orange) and electrostatic/steric (yellow) fields for the different protein families. A transferase, B hydrolase, C oxidoreductase, D lyases, E ligases, F isomerases, G DNA-binding proteins, H chaperones, I BET, and L iLBP. The results are sorted from highest to lowest performance as obtained from electrostatic/steric results.

The combination of log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB fields leads to an overall success of 88% for the AZ data set. This performance is slightly better than the results obtained from the molecular overlays generated by combining electrostatic and steric fields, which yielded a global accuracy of 82%. This tendency is also found when the average metrics is considered (38% versus 35.6% for of log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB and electrostatic/steric fields, respectively; see Table S7 in the Supporting Information). The larger differences between the protein families are obtained for transferases, where 87% and 79% of the molecules were correctly aligned from log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB and electrostatic/steric descriptors, respectively, followed by oxidoreductases (94% and 88%, respectively), while the differences were generally less notable for the rest of protein categories. Nevertheless, it must be noted that the performance of log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB and electrostatic/steric descriptors may exhibit significant differences for individual targets, as can be observed in Figure S1 in the Supporting Information. On the other hand, there is no apparent correlation between the accuracy and the global hydrophobicity of the compounds, which was estimated using the AlogP method, as noted in Figure S2 (Supporting Information). This reveals the fact that the relevant property is not the total logP but the 3D distribution pattern of the atomic hydrophobic contributions.

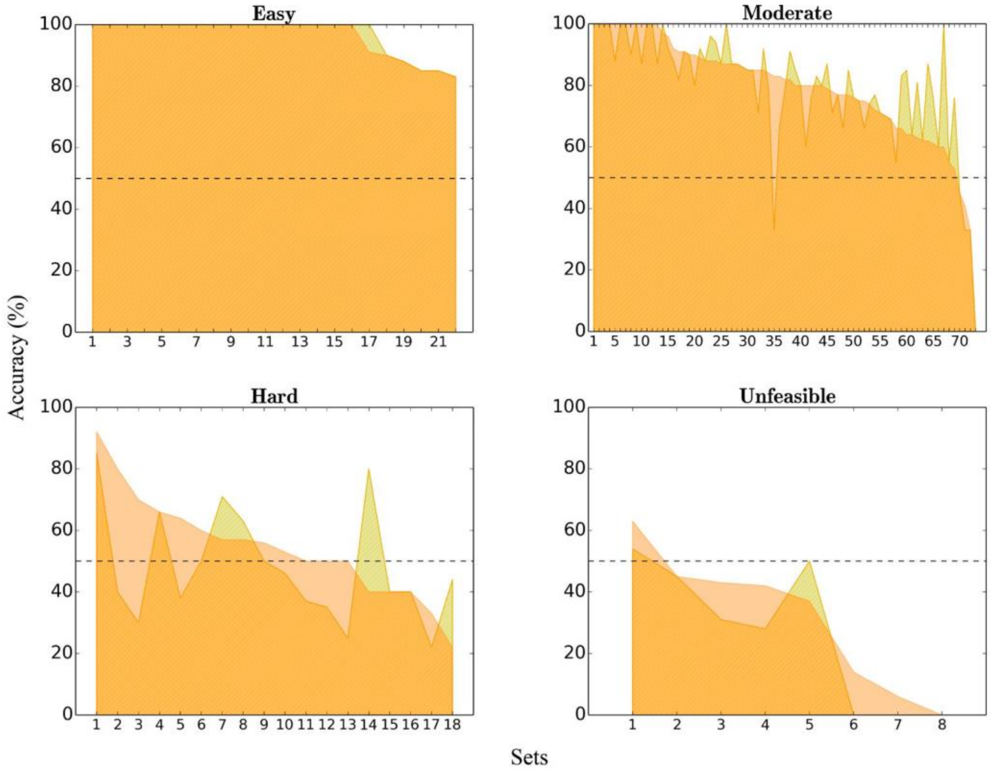


Figure 5. Comparison of experimental and predicted molecular overlays by target category in the AZ data set. Accuracy (%) of molecular overlays predicted from log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB (orange) and electrostatic/steric (yellow) fields for the different categories in the AZ data set. The results are sorted from highest to lowest performance as obtained from log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB descriptors. The correspondence between the numbering of the sets in each category and its target name is reported in Table S1 in the Supporting Information.

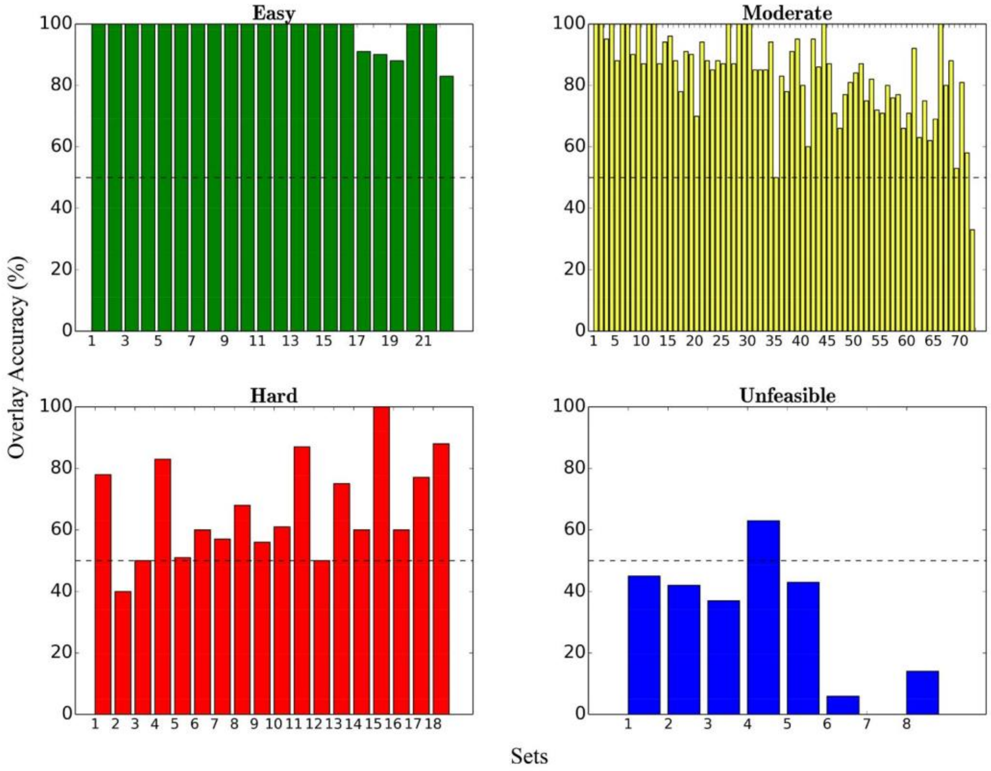


Figure 6. Degree of identity (%) between the molecular overlays obtained predicted from log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/ log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB and electrostatic/steric fields for the four categories of AZ sets. The ordering of the sets for each category follows the ordering shown in Figure 5.

Almost all the ligands included in the 22 sets within the easy category were correctly aligned (96.5% accuracy; Figure 5), a trait also found for most of the ligands pertaining to the 73 sets in the moderate category (accuracy of 79.4%). The overall

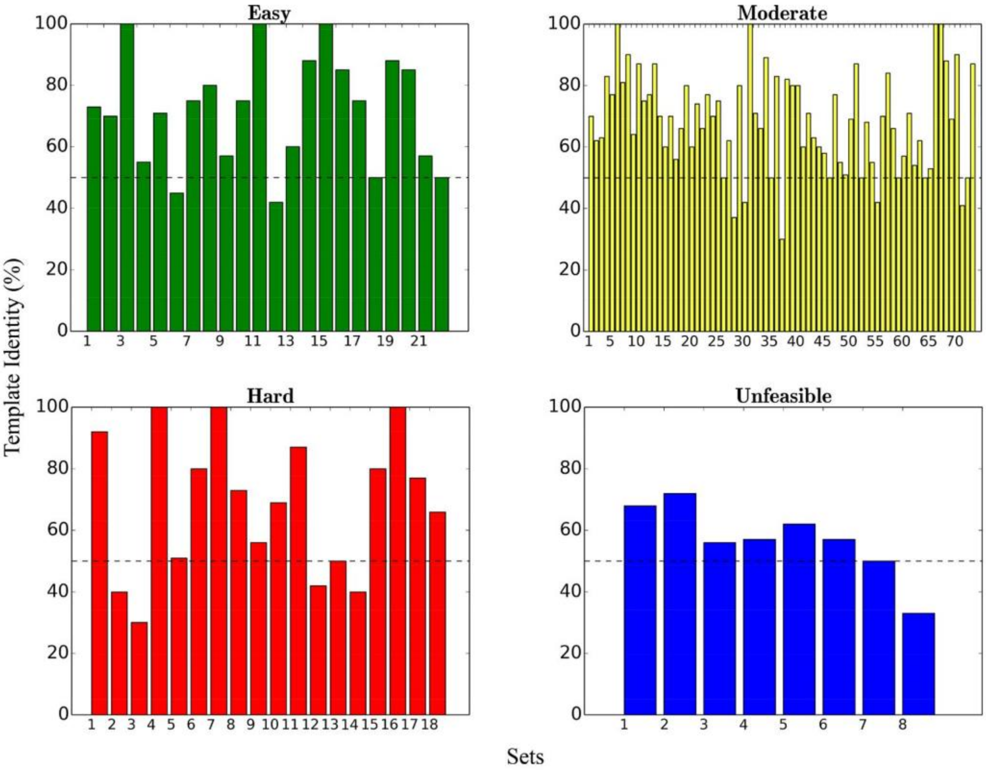


Figure 7. Degree of identity (%) between the templates associated with the best pose in the molecular overlays obtained from log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB and electrostatic/steric fields for the four categories of AZ sets. The ordering of the sets for each category follows the ordering shown in Figure 5.

performance obtained for the molecular overlays predicted from electrostatic/steric fields is identical (96.6% and 79.4% for easy and moderate sets, respectively), although there are differences in the accuracy attained for individual targets in the moderate category. As expected, the percentage of correctly overlaid ligands was reduced to 54% for the 18 sets under the hard category. In this case, the accuracy obtained with log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB descriptors exceeded the behavior found for electrostatic/steric fields, with an overall success of 48%. Finally, the behavior for the unfeasible category was reduced to 31%, reflecting the difficulty in predicting the molecular overlays in cases where the molecules explore different subpockets within the binding cavity of the targets in this category (see Figure S3 in the Supporting Information).

These results can be compared with the performance reported by Giangreco et al. in the assessment of a Cambridge Structural Database-driven overlay program using the AZ data set.[67] Thus, using as reference the accuracy obtained using the best *AlignScore*, the number (%) of correctly predicted overlays was 95, 73, and 39 for easy, moderate, and hard sets, whereas no ligand was successfully predicted in the unfeasible category. On the other hand, Chan has recently reported a novel algorithm, named MolAlign, which was also checked using the AZ data set.[73] In this case, the percentage of correct overlays was 95, 68, 44, and 13 (results derived by using conformers generated with Ballon and Confect, considering a geometrically successful arrangement in any of the top five solutions). Although caution is required for a quantitative comparison due to the differences in the computational protocol and performance metrics adopted in these studies, present results suggest that PharmScreen leads to a slight improvement for the targets pertaining to the hard set.

It is worth noting that the apparently similar overall performance of log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB and electrostatic/steric fields does not necessarily imply that these descriptors lead to identical overlays for a given compound. This can be verified in Figure 6, which shows the number (%) of identical superpositions (i.e., those with a RMSD ≤ 2 Å) between the best poses predicted by the two combinations of molecular fields for the four categories of compounds present in the AZ data set. In general, the larger the degree of difficulty expected for the four categories, the lower the identity between the predicted overlays for the best poses. Thus, for the 22 sets included in the easy category, both log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB and electrostatic/steric fields lead to the same molecular alignments in 18 cases, and the agreement is larger than 80% in the remaining 4 cases. This level of identity is attained in 50 out of the 73 sets included in the moderate category, and it is found only in 4 cases out of the 18 sets pertaining to the hard category. Finally, only a single set reaches a number of identical overlays larger than 50% for the unfeasible targets. Therefore, although the overall accuracy achieved from log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB and electrostatic/steric fields is similar for every separate category of targets (Figure 5), Figure 6 shows that these descriptors may lead to different molecular overlays for the compounds pertaining to individual targets, especially for those included in the hard category.

It may also be questioned whether the best pose obtained by using log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB and electrostatic/steric fields comes from the same template or from distinct reference molecules. To answer this question, Figure 7 represents the number (%) of common templates shared in the best alignment of the compounds by the two sets of descriptors. Even for the subset of easy targets, comparison of the plots in

Figures 6 and 7 shows that there is not a correspondence between the degree of common templates and the identity in molecular overlays obtained from the two classes of descriptors. Thus, there is a perfect agreement between the chosen templates only in 3 out of the 22 cases pertaining to the easy category, while the template identity lies close to or below 60% in 5 cases (sets 4, 6, 9, 12, and 13) that, nevertheless, yielded 100% identity in the molecular overlays of the compounds. A perfect identity between templates is limited to 4 out of the 73 cases in the moderate family, and 3 out of the 18 sets in the hard category.

Finally, we compared the overlay accuracy (%) predicted for the best pose obtained from log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB fields versus the identity (%) between the templates that led to the correct alignment from the experimental pose (RMSD ≤ 2 Å) when these descriptors and the electrostatic/steric ones were used. This comparison is shown in Figure 8 using different symbols for the targets included in easy, moderate, hard, and unfeasible categories. The results point out that a successful prediction of the molecular overlay can be achieved with a low degree of identity between the templates chosen for the two sets of molecular fields. For instance, there are cases in the moderate category where molecular overlays are predicted with 80% accuracy, but the correspondence between the templates used with log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$ /HB and electrostatic/steric fields can be as low as 30%. Overall, this indicates that the two sets of descriptors provide complementary information that may lead to plausible molecular overlays useful for predicting the relative arrangement between molecules.

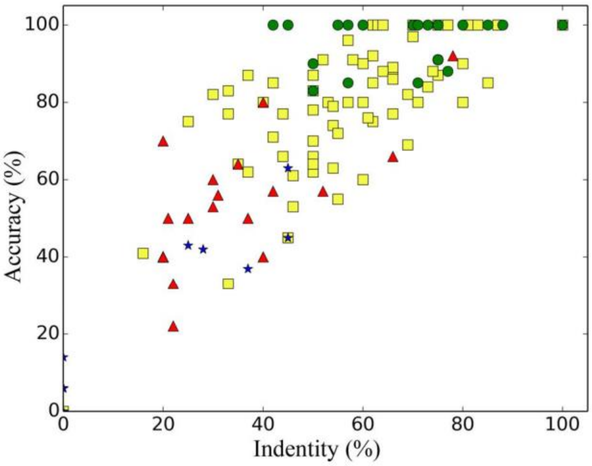


Figure 8. Representation of the overlay accuracy (%) obtained from log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB fields versus the identity (%) of templates leading simultaneously to correct molecular alignments using log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB and electrostatic/steric fields. Results obtained for the sets in easy, moderate, hard, and unfeasible categories are shown as green circles, yellow squares, red triangles, and blue stars, respectively.

Discussion of Selected Cases. In this section we examine the molecular overlays for selected representative cases pertaining to the easy, moderate and hard systems, as the lack of common recognition features in the unfeasible set makes the prediction of molecular overlays to be difficult in the absence of additional structural information, such as key pharmacophoric features.

Easy Set. For the targets in this category, the overlay accuracy is larger than 80% (Figure 5), although the template identity can be notably lower (Figure 7). This is illustrated by the molecular alignments obtained for the 7 ligands in target P30291 (Tyrosine kinase Wee1; set 12 in this category), which are predicted correctly in all cases, but the template identity is only 42%. As shown in Figure 9, the ligand in X-ray structure 1X8B obtained the highest similarity against the template taken from PDB entry 3CR0 when log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB descriptors were used. However, when electrostatic/steric fields were adopted, the highest similarity was found against the template taken from PDB entry 2Z2W. This illustrates how the correct molecular overlay can be obtained for a given compound from distinct templates depending on the specific set of descriptors used in similarity measurements.

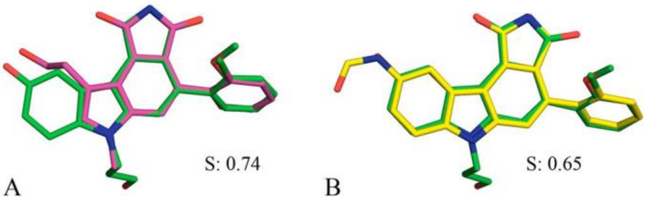


Figure 9. Molecular overlays obtained for the compound bound to tyrosine kinase Wee1 in PDB entry 1 × 8B (C atoms in green). Molecular overlays were obtained for template molecules taken from (A) PDB entry 3CR0 (C atoms in magenta) and (B) 2Z2W (C atoms in yellow) when log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB and electrostatic/steric fields, respectively, were used. S denotes the similarity score.

Moderate Set. The overall overlay accuracy obtained when log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB and electrostatic/steric descriptors are used is almost identical (79%; Figure 5). In most cases the two sets of descriptors give rise to similar poses, but the overlaid ligands may often differ by most than 2 Å, which is a widely accepted threshold in checking the accuracy of predicted poses. This is exemplified in Figure 10 for the ligand bound to cytochrome P450 2A6 (taken from PDB entry 1ZL0; set 35 in moderate category), as the best alignment obtained against the template taken from PDB entry 3EBS leads to arrangements that differ by more than 2 Å from the experimental pose when electrostatic/steric descriptors were considered.

Hard Set. This is the set that gives rise to most notable differences in the molecular overlays obtained between the two sets of descriptors. For instance, the log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB descriptors led to a correct prediction for 7 (1QPE, 2OF2, 2ZMI, 3ACJ, 3AD4, 3AD5, and 3AD6) out of the 10 ligands bound to target P06239 (tyrosine-protein kinase LcK; set 3 in this category), whereas only 3 cases (2OF2, 3AD4, and 3AD6) were correctly predicted from electrostatic/steric fields. As an example, Figure 11 shows the alignment of the ligand present in PDB entry 1QPE and the compound in the X-ray structure 2ZM1, which was used as template. Figure 11 shows the reversal of the ligand orientation obtained between the two sets of descriptors.

For P11309 (serine/threonine protein kinase pim-1; set 5 in hard category) the log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB fields generated a correct prediction for 20 out of the 31 ligands in the set, whereas electrostatic/steric descriptors succeeded in the molecular alignment in 12 cases. Two representative examples are shown in Figure 12, which shows the correct alignment of ligands taken from X-ray structures 1YI3 and 3MAE using as templates the compounds extracted from 2C3I and 3R00 when hydrophobic descriptors are used. Note the reversed

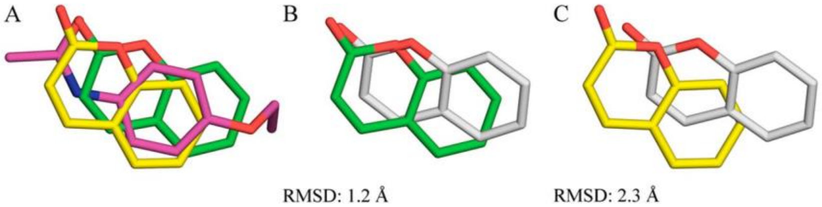


Figure 10. Molecular overlays obtained for the compound bound to cytochrome P450 2A6 in PDB entry 1ZL0. (A) Representation of the overlaid ligand (C atoms in green: log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB; C atoms in yellow: electrostatic/steric) against the template taken from PDB entry 3EBS (C atoms in magenta). (B) Comparison of the experimental pose (C atoms in white) and the ligand (C atoms in green) overlaid using log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB descriptors. (C) Comparison of the experimental pose (C atoms in white) and the ligand (C atoms in yellow) overlaid using electrostatic/steric descriptors.

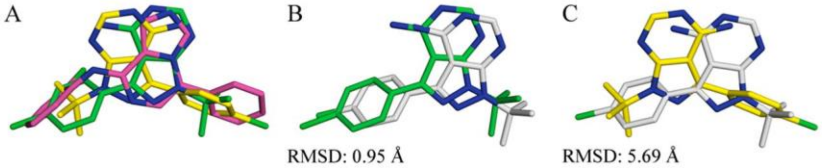


Figure 11. Molecular overlays obtained for the compound bound to tyrosine-protein kinase LcK in PDB entry 1QPE. (A) Representation of the overlaid ligand (C atoms in green log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB; C atoms in yellow electrostatic/steric) against the template taken from PDB entry 2ZM1 (C atoms in magenta). (B) Comparison of the experimental pose (C atoms in white) and the ligand (C atoms in green) overlaid using log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB descriptors. (C) Comparison of the experimental pose (C atoms in white) and the ligand (C atoms in yellow) overlaid using electrostatic/steric descriptors.

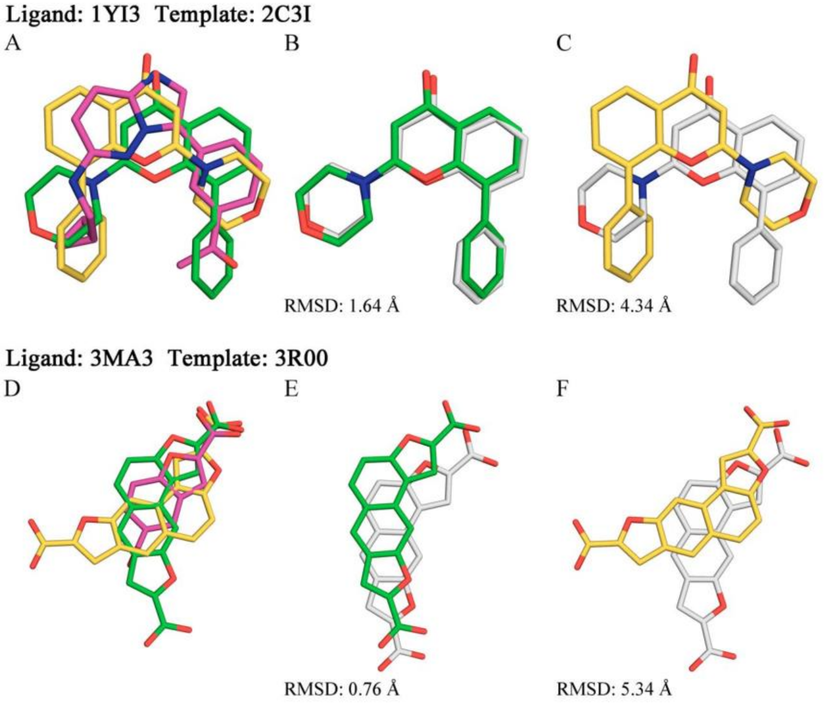


Figure 12. Molecular overlays obtained for compounds bound to serine/threonine protein kinase pim-1 in PDB entries (top) 1YI3 and (bottom) 3MA3. (A, D) Representation of the overlaid ligand (C atoms in green log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB; C atoms in yellow electrostatic/steric) against the template taken from PDB entry (top) 2C3I and (bottom) 3R00 (C atoms in magenta). (B, E) Comparison of the experimental pose (C atoms in white) and the ligand (C atoms in green) overlaid using log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB descriptors. (C, F) Comparison of the experimental pose (C atoms in white) and the ligand (C atoms in yellow) overlaid using electrostatic/steric descriptors.

orientations of the two ligands obtained for these templates by the two sets of descriptors.

Finally, an example of the correct overlay predicted only from electrostatic/steric fields is the allosteric site in O15530 (serine/threonine protein kinase pdpk1; set 14 in this category). For this system the hydrophobic descriptors generated a correct prediction for two (3OTU and 4A07) out of the five cases, while four compounds were correctly aligned when electrostatic/steric fields were used (3ORZ, 3OTU, 4A06, and 4A07). As noted in Figure 13, even though the alignment obtained from the log $P^{\mathrm{ele}}_{\mathrm{o/w}}$/log $P^{\mathrm{cav}}_{\mathrm{o/w}}$/HB descriptors resembles the X-ray pose (panel B), the chlorobenzene unit is deviated, making the RMSD slightly larger than the threshold, presumably due to the tendency to match the chlorobenzene unit of the ligand onto the benzofuran unit of the template.

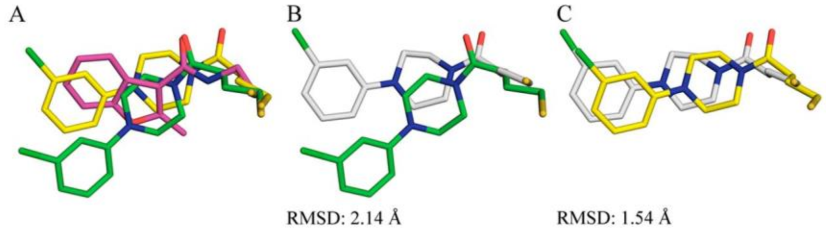


Figure 13. Molecular overlays obtained for compounds bound to serine/threonine protein kinase pdpk1 in PDB entry 3ORZ. (A) Representation of the overlaid ligand (C atoms in green log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB; C atoms in yellow electrostatic/steric) against the template taken from PDB entry 3ORX (C atoms in magenta). (B) Comparison of the experimental pose (C atoms in white) and the ligand (C atoms in green) overlaid using log $P^{ele}_{o/w}$/log $P^{cav}_{o/w}$/HB descriptors. (C) Comparison of the experimental pose (C atoms in white) and the ligand (C atoms in yellow) overlaid using electrostatic/steric descriptors.

■ CONCLUSIONS

Since the maximal binding affinity that can be attained by a target is primarily due to the curvature and apolar surface of the binding pocket,[28] an accurate representation of the 3D pattern of hydrophobic/hydrophilic regions may be valuable as a source of descriptors in rational drug design. In particular, we have presented PharmScreen, a tool that generates ligand overlays based on MST fractional contributions to the octanol/water partition coefficients.

The results obtained for the systems included in the AZ calibration data set give support to the assumption that the hydrophobic/hydrophilic balance in a molecule, supplemented with the HB features, may provide a useful signature to enrich molecular alignment studies performed traditionally based only on electrostatic and steric properties. The results point out the suitability of the MST based-hydrophobic parameters, as correct overlays were predicted for 94%, 79%, and 54% of the molecules classified into easy, moderate, and hard sets, respectively. Moreover, the results point out that this accuracy is attained at a much lower degree of identity between the templates used by the combination of electrostatic/steric fields, which reinforces the complementarity between these descriptors in order to take into account the increasing complexity of the targets under investigation. These findings support the usefulness of PharmScreen as a valuable alternative for molecule superposition and virtual screening of chemical libraries, and future studies will be conducted to calibrate its performance in ligand-based virtual screening studies.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00216.

> List of targets in the AstraZeneca data set, effect of weighting factors on the alignment accuracy, average values of successful overlays obtained for the AstraZeneca test set using the electrostatic/steric fields, comparison of the accuracy of molecular overlays predicted for every individual target in the AZ data set, and representation of selected molecular overlays (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: fjluque@ub.edu. Tel.: +34 934033788.

ORCID

Obdulia Rabal: 0000-0002-3224-0987
Julen Oyarzabal: 0000-0003-1941-7255
F. Javier Luque: 0000-0002-8049-3567

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. E. Jamois for valuable suggestions. We thank the Ministerio de Economía y Competitividad (MINECO: SAF2017-88107-R, DI-14-06634) and the Generalitat de Catalunya (2017SGR1746, 2015-DI-052) for financial support. The Foundation for Applied Medical Research (FIMA) and the University of Navarra (Pamplona, Spain) are acknowledged for financial support, and the Consorci de Serveis Universitaris de Catalunya (CSUC), for computational resources (Molecular Recognition project). F.J.L. acknowledges the support from ICREA Academia. J.V. has a fellowship from the Generalitat de Catalunya.

■ REFERENCES

(1) Lemmen, C.; Lengauer, T. Computational Methods for the Structural Alignment of Molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.

(2) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.

(3) Bultinck, P.; Gironés, X.; Carbó-Dorcaz, R. Molecular Quantum Similarity: Theory and Applications. *Rev. Comput. Chem.* **2005**, *21*, 127−207.

(4) Medina-Franco, J. L.; Maggiora, G. M. Molecular Similarity Analysis. In *Chemoinformatics for Drug Discovery*; Bajorath, J., Ed.; John Wiley & Sons, Inc., 2013; pp 343−399.

(5) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186−3204.

(6) Tresadern, G.; Bemporad, D. Modeling Approaches for Ligand-Based 3D Similarity. *Future Med. Chem.* **2010**, *2*, 1547−1561.

(7) Finn, P. W.; Morris, G. M. Shape-Based Similarity Searching in Chemical Databases. *WIRES Comput. Mol. Sci.* **2013**, *3*, 226−241.

(8) Shin, W.-H.; Zhu, X.; Bures, M. G.; Kihara, D. Three-Dimensional Compound Comparison Methods and Their Application in Drug Discovery. *Molecules* **2015**, *20*, 12841−12862.

(9) Ballester, P. J.; Richards, W. G. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* **2007**, *28*, 1711−1723.

(10) Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Moretti, L.; Cooper, R. I.; Richards, W. G. ElectroShape: Fast Molecular Similarity Calculations Incorporating Shape, Chirality and Electrostatics. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 789−801.

(11) Armstrong, M. S.; Finn, P. W.; Morris, G. M.; Richards, W. G. Improving the Accuracy of Ultrafast Ligand-Based Screening: Incorporating Lipophilicity into ElectroShape as an Extra Dimension. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 785−790.

(12) Jain, A. N. Ligand-Based Structural Hypotheses for Virtual Screening. *J. Med. Chem.* **2004**, *47*, 947−961.

(13) Hofbauer, C.; Lohninger, H.; Aszódi, A. SURFCOMP: A Novel Graph-Based Approach to Molecular Surface Comparison. J. *Chem. Inf. Comput. Sci.* **2004**, *44*, 837−847.

(14) Sastry, G. M.; Dixon, S. L.; Sherman, W. Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlap Scoring. *J. Chem. Inf. Model.* **2011**, *51*, 2455−2466.

(15) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein−Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.

(16) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(17) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential. *J. Chem. Inf. Model.* **2009**, *49*, 492−502.

(18) Liu, X.; Jiang, H.; Li, H. SHAFTS: A Hybrid Approach for 3D Molecular Similarity Calculation. 1. Method and Assessment of Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 2372−2385.

(19) Roy, A.; Skolnick, J. LIGSIFT: An Open-Source Tool for Ligand Structural Alignment and Virtual Screening. *Bioinformatics* **2015**, *31*, 539−544.

(20) Horvath, D. Pharmacophore-Based Virtual Screening. *Methods Mol. Biol.* **2010**, *672*, 261−298.

(21) Tosco, P.; Balle, T.; Shiri, F. Open3DALIGN: An Open-Source Software Aimed at Unsupervised Ligand Alignment. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 777−783.

(22) Moser, D.; Wittmann, S. K.; Kramer, J.; Blöcher, R.; Achenbach, J.; Pogoryelov, D.; Proschak, E. PENG: A Neural Gas-Based Approach for Pharmacophore Elucidation. Method Design, Validation, and Virtual Screening for Novel Ligands of LTA4H. *J. Chem. Inf. Model.* **2015**, *55*, 284−293.

(23) Artese, A.; Cross, S.; Costa, G.; Distinto, S.; Parrotta, L.; Alcaro, S.; Ortuso, F.; Cruciani, G. Molecular Interaction Fields in Drug Discovery: Recent Advances and Future Perspectives. *WIRES Comput. Mol. Sci.* **2013**, *3*, 594−613.

(24) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. *J. Chem. Inf. Model.* **2008**, *48*, 2108−2117.

(25) Cross, S.; Baroni, M.; Carosati, E.; Benedetti, P.; Clementi, S. FLAP: GRID Molecular Interaction Fields in Virtual Screening, Validation Using the DUD Data Set. *J. Chem. Inf. Model.* **2010**, *50*, 1442−1450.

(26) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010**, *53*, 3862−3886.

(27) Spyrakis, F.; Ahmed, M. H.; bayden, A. S.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **2017**, *60*, 6781−6827.

(28) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-Based Maximal Affinity Model Predicts Small-Molecule Druggability. *Nat. Biotechnol.* **2007**, *25*, 71−75.

(29) Davis, A. M.; Teague, S. J. Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angew. Chem., Int. Ed.* **1999**, *38*, 736−749.

(30) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *J. Med. Chem.* **2005**, *48*, 2518−2525.

(31) Egner, U.; Hillig, R. C. A Structural Biology View of Target Drugability. Expert Opin. *Expert Opin. Drug Discovery* **2008**, *3*, 391−401.

(32) Schmidtke, P.; Barril, X. Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *J. Med. Chem.* **2010**, *53*, 5858−5867.

(33) Schmidtke, P.; Luque, F. J.; Murray, J. B.; Barril, X. Shielded Hydrogen Bonds as Structural determinants of Binding Kinetcs: Application in Drug Design. *J. Am. Chem. Soc.* **2011**, *133*, 18903−18910.

(34) Tsopelas, F.; Giaginis, C.; Tsantili-Kakoulidou, A. Lipophilicity and Biomimetic Properties to Support Drug Discovery. Expert Opin. *Expert Opin. Drug Discovery* **2017**, *12*, 885−896.

(35) Gaillard, P.; Carrupt, P. A.; Testa, B.; Boudon, A. Molecular Lipophilicity Potential, a Tool in 3D QSAR: Method and Applications. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 83−96.

(36) Eugene Kellogg, G. E.; Abraham, D. J. Hydrophobicity: Is LogP(o/w) More than the Sum of its Parts? *Eur. J. Med. Chem.* **2000**, *35*, 651−661.

(37) Sarkar, A.; Kellogg, G. E. Hydrophobicity-Shake Flasks, Protein Folding and Drug Discovery. *Curr. Top. Med. Chem.* **2010**, *10*, 67−83.

(38) Carrupt, P. A.; Gaillard, P.; Billois, F.; Weber, P.; Testa, B.; Meyer, C.; Pérez, S. The Molecular Lipophilicity Potential (MLP): A New Tool for log P Calculations and Docking, and in Comparative Molecular Field Analysis (CoMFA). In *Lipophilicity in Drug Action and Toxicology*; Pliska, V., Testa, B., Van De Waterbeemd, H., Eds.; Weinheim VCH: Germany, 1996; pp 49−71.

(39) Fornabaio, M.; Spyrakis, F.; Mozzarelli, A.; Cozzini, P.; Abraham, D. J.; Kellogg, G. E. Simple, Intuitive Calculations of Free Energy of Binding for Protein-Ligand Complexes. 3. The Free Energy Contribution of Structural Water Molecules in HIV-1 Protease Complexes. *J. Med. Chem.* **2004**, *47*, 4507−4516.

(40) Marabotti, A.; Spyrakis, F.; Facchiano, A.; Cozzini, P.; Alberti, S.; Kellogg, G. E.; Mozzarelli, A. Energy-Based Prediction of Amino Acid-Nucleotide Base Recognition. *J. Comput. Chem.* **2008**, *29*, 1955−1969.

(41) Ahmed, M. H.; Spyrakis, F.; Cozzini, P.; Tripathi, P. K.; Mozzarelli, A.; Scarsdale, J. N.; Safo, M. A.; Kellogg, G. E. Bound Water at Protein-Protein Interfaces: Partners, Roles and Hydrophobic Bubbles as a Conserved Motif. *PLoS One* **2011**, *6*, e24712.

(42) Curutchet, C.; Orozco, M.; Luque, F. J. Solvation in Octanol: Parametrization of the Continuum MST Model. *J. Comput. Chem.* **2001**, *22*, 1180−1193.

(43) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Luque, F. J. Extension of the MST Model to the IEF Formalism: HF and B3LYP Parametrizations. *J. Mol. Struct.: THEOCHEM* **2005**, *727*, 29−40.

(44) Ginex, T.; Muñoz-Muriedas, J.; Herrero, E.; Gibert, E.; Cozzini, P.; Luque, F. J. Development and Validation of Hydrophobic Molecular Fields Derived from the Quantum Mechanical IEF/PCM-MST Solvation Models in 3D-QSAR. *J. Comput. Chem.* **2016**, *37*, 1147−1162.

(45) Ginex, T.; Muñoz-Muriedas, J.; Herrero, E.; Gibert, E.; Cozzini, P.; Luque, F. J. Application of the Quantum Mechanical IEF/PCM-MST Hydrophobic Descriptors to Selectivity in Ligand Binding. *J. Mol. Model.* **2016**, *22*, 136.

(46) Giangreco, I.; Cosgrove, D. A.; Packer, M. J. An Extensive and Diverse Set of Molecular Overlays for the Validation of Pharmacophore Programs. *J. Chem. Inf. Model.* **2013**, *53*, 852−866.

(47) Javier Luque, F. J.; Curutchet, C.; Muñoz-Muriedas, J.; Bidon-Chanal, A.; Soteras, I.; Morreale, A.; Gelpí, J. L.; Orozco, M. Continuum Solvation Models: Dissecting the Free Energy of Solvation. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3827−3836.

(48) Luque, F. J.; Bofill, J. M.; Orozco, M. New Strategies to Incorporate the Solvent Polarization in Self-Consistent Reaction Field and Free-Energy Perturbation Simulations. *J. Chem. Phys.* **1995**, *103*, 10183−10191.

(49) Pierotti, R. A. A. Scaled Particle Theory of Aqueous and Nonaqueous Solutions. *Chem. Rev.* **1976**, *76*, 717−726.

(50) Claverie, P. Elaboration of Approximate Formulas for the Interactions Between Large Molecules: Applications in Organic Chemistry. In *Intermolecular Interactions: From Diatomics to*

Biopolymers; Pullman, B., Ed.; Wiley: New York, 1978; Vol. *1*, pp 69−305.

(51) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. RM1: A Reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.* **2006**, *27*, 1101−1111.

(52) Forti, F.; Barril, X.; Luque, F. J.; Orozco, M. Extension of the MST Continuum Solvation Model to the RM1 Semiempirical Hamiltonian. *J. Comput. Chem.* **2008**, *29*, 578−587.

(53) *MOPAC 6.0.*; version locally modified by Luque, F. J.; Orozco, M.; University of Barcelona, 2008.

(54) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stramann, R. R.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, S.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Oritiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian09*, Revision D.01; Gaussian Inc., Wallingford CT, 2009.

(55) Klebe, G.; Abraham, U. Comparative Molecular Similarity Index Analysis (CoMSIA) to Study Hydrogen-Bonding Properties and to Score Combinatorial Libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 1−10.

(56) Laurence, C.; Brameld, K. A.; Graton, J.; Le Questel, J.-Y.; Renault, E. The P K BHX Database: Toward a Better Understanding of Hydrogen-Bond Basicity for Medicinal Chemists. *J. Med. Chem.* **2009**, *52*, 4073−4086.

(57) Hunter, C. A. Quantifying Intermolecular Interactions: Guidelines for the Molecular Recognition Toolbox. *Angew. Chem., Int. Ed.* **2004**, *43*, 5310−5324.

(58) Salichs, A.; López, M.; Segarra, V.; Orozco, M.; Luque, F. J. Fast Estimation of Hydrogen-Bonding Donor and Acceptor Propensities: A GMIPp Study. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 569−583.

(59) Kenny, P. W.; Montanari, C. A.; Prokopczyk, I. M.; Ribeiro, J. F. R.; Sartori, G. R. Hydrogen Bond Basicity Prediction for Medicinal Chemistry Design. *J. Med. Chem.* **2016**, *59*, 4278−4288.

(60) Platt, D. E.; Silverman, B. D. Registration, Orientation, and Similarity of Molecular Electrostatic Potentials through Multipole Matching. *J. Comput. Chem.* **1996**, *17*, 358−366.

(61) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129−2140.

(62) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.

(63) Lemmen, C.; Lengauer, T.; Klebe, G. FlexS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.

(64) Chen, Q.; Higgs, R. E.; Vieth, M. Geometric Accuracy of Three-Dimensional Molecular Overlays. *J. Chem. Inf. Model.* **2006**, *46*, 1996−2002.

(65) 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.

(66) *Pymol Molecular Graphics System*, version 2.0; Schrödinger, LLC, 2015.

(67) Giangreco, I.; Olsson, T. S. G.; Cole, J. C.; Packer, M. J. Assessment of a Cambridge Structural Database-Driven Overlay Program. *J. Chem. Inf. Model.* **2014**, *54*, 3091−3098.

(68) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.

(69) Ferenczy, G. G.; Reynolds, C. A.; Richards, W. G. Semi-empirical AM1 Electrostatic Potentials and AM1 Electrostatic Potential Derived Charges: A Comparison With ab Initio Values. *J. Comput. Chem.* **1990**, *11*, 159−169.

(70) Alhambra, C.; Luque, F. J.; Orozco, M. Comparison of NDDO and Quasi-Ab Initio Approaches to Compute Semiempirical Molecular Electrostatic Potentials. *J. Comput. Chem.* **1994**, *15*, 12−22.

(71) Boström, J.; Böhm, M.; Gundertofte, K.; Klebe, G. A 3D QSAR Study on a Set of Dopamine D4 Receptor Antagonists. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1020−1027.

(72) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(73) Chan, S. L. MolAlign: An Algorithm for Aligning Multiple Small Molecules. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 523−546.

Supporting Information

Development and Validation of Molecular Overlays Derived From 3D Hydrophobic Similarity with PharmScreen

Javier Vazquez,[†,‡] Alessandro Deplano,[†] Albert Herrero,[†] Tiziana Ginex, [‡] Enric Gibert,[†] Obdulia Rabal,[§] Julen Oyarzabal,[§] Enric Herrero,[†] and F. Javier Luque[‡,*]

[†] Pharmacelera, Plaça Pau Vila, 1, Sector 1, Edificio Palau de Mar, Barcelona 08039, Spain

[‡] Department of Nutrition, Food Science and Gastronomy, Faculty of Pharmacy and Food Sciences, Institute of Biomedicine (IBUB), and Institute of Theoretical and Computational Chemistry (IQTCUB), University of Barcelona, Av. Prat de la Riba 171, Santa Coloma de Gramenet E-08921, Spain

[§] Small Molecule Discovery Platform, Molecular Therapeutics Program, Center for Applied Medical Research (CIMA), University of Navarra, Avda. Pio XII 55, Pamplona E-31008, Spain

* E-mail: fjluque@ub.edu

Table S1. List of targets included in the AstraZeneca dataset.

Family	ID[a]	Name	Category	N_{PDB}[b]	no.[c]	Source
transferase	A9JQL9	dehydrosqualene synthase	Unfeasible	6	4	*Staphylococcus Aureus*
transferase	O14757	serine/threonine-protein kinase Chk1	Moderate	42	53	*Homo Sapiens*
transferase	O14965	serine/threonine-protein kinase 6	Moderate	11	71	*Homo Sapiens*
transferase	O15530	3-phosphoinositide dependent protein kinase-1	Moderate	18	46	*Homo Sapiens*
transferase	O15530*	3-phosphoinositide dependent protein kinase-1	Hard	5	14	*Homo Sapiens*
transferase	O60674	tyrosine-protein kinase JAK2	Moderate	13	70	*Homo Sapiens*
bromodomain and extra-terminal protein	O60885	glutathione-requiring prostaglandin D synthase	Moderate	8	1	*Homo Sapiens*
hydrolase	O76074	cGMP-specific 3'.5'-cyclic phosphodiesterase	Moderate	9	22	*Homo Sapiens*
oxidoreductase	O76290	pteridine reductase	Moderate	9	25	*Trypanosoma Brucei Brucei*
oxidoreductase	P00374	dihydrofolate reductase	Easy	15	1	*Homo Sapiens*
transferase	P00469	thymidylate synthase	Moderate	5	67	*Lactobacillus Casei*
transferase	P00489	protein (glycogen phosphorylase)	Easy	42	2	*Oryctolagus Cuniculus*
transferase	P00509	aspartate aminotransferase	Easy	6	22	*Escherichia Coli*
transferase	P00517	cAMP-dependent protein kinase. alpha-catalytic subunit	Moderate	21	39	*Bos Taurus*
transferase	P00520	proto-oncogene tyrosine-protein kinase ABL	Moderate	8	40	*Mus Musculus*
transferase	P00523	proto-oncogene tyrosine-protein kinase Src	Moderate	13	72	*Gallus Gallus*
hydrolase	P00730	lysozyme	Moderate	11	2	*Enterobacteria Phage T4*
hydrolase	P00734	alpha thrombin	Moderate	78	15	*Homo Sapiens*
hydrolase	P00742	coagulation factor XA	Moderate	37	14	*Homo Sapiens*
hydrolase	P00749	protein (urokinase-type plasminogen activator)	Moderate	27	16	*Homo Sapiens*
hydrolase	P00760	trypsin	Moderate	71	3	*Bos Taurus*
hydrolase	P00772	elastase	Hard	5	2	*Sus Scrofa*
hydrolase	P00797	renin	Easy	6	3	*Homo Sapiens*
hydrolase	P00808	beta-lactamase	Hard	8	11	*Bacillus Licheniformis*
hydrolase	P00811	beta-lactamase	Unfeasible	24	1	*Escherichia Coli*
lyase	P00918	carbonic anhydrase II	Moderate	135	37	*Homo Sapiens*
lyase	P00929	tryptophan synthase	Moderate	10	19	*Salmonella Typhimurium*
chaperone	P02829	HSP82	Moderate	11	63	*Saccharomyces Cerevisiae*
DNA-binding proteins	P03372	oestrogen receptor	Moderate	27	23	*Homo Sapiens*
oxidoreductase	P04035	protein (HMG-COA reductase)	Moderate	7	30	*Homo Sapiens*
hydrolase	P04058	acetylcholinesterase	Unfeasible	8	5	*Torpedo Californica*
oxidoreductase	P04642	l-lactate dehydrogenase A	Moderate	8	51	*Rattus Norvegicus*

		chain				
oxidoreductase	P05326	isopenicillin N synthase	Easy	9	4	*Emericella Nidulans*
transferase	P06239	LCK kinase	Hard	10	3	*Homo Sapiens*
DNA-binding proteins	P06401	progesterone receptor	Moderate	9	47	*Homo Sapiens*
hydrolase	P07688	cathepsin B	Easy	7	20	*Bos Taurus*
chaperone	P07900	eat shock protein HSP 90-alpha	Moderate	74	17	*Homo Sapiens*
transferase	P08069	insulin-like growth factor 1 receptor precursor	Moderate	8	64	*Homo Sapiens*
DNA-binding proteins	P08235	mineralocorticoid receptor	Easy	7	5	*Homo Sapiens*
hydrolase	P08254	stromelysin-1	Moderate	9	48	*Homo Sapiens*
transferase	P08581	hepatocyte growth factor receptor	Moderate	14	50	*Homo Sapiens*
hydrolase	P08709	coagulation factor VII	Moderate	5	41	*Homo Sapiens*
hydrolase	P09467	fructose-1.6-bisphosphatase 1	Moderate	6	4	*Homo Sapiens*
hydrolase	P09955	procarboxypeptidase B	Moderate	9	5	*Sus Scrofa*
hydrolase	P09960	leukotriene A-4 hydrolase	Moderate	23	34	*Homo Sapiens*
oxidoreductase	P0A017	dihydrofolate reductase	Easy	11	6	*Staphylococcus Aureus*
hydrolase	P0A5J2	methionine aminopeptidase	Moderate	6	6	*Mycobacterium Tuberculosis*
transferase	P0ABP9	purine nucleoside phosphorylase	Easy	8	7	*Escherichia Coli*
hydrolase	P0AD64	beta-lactamase SHV-1	Moderate	6	73	*Klebsiella Pneumoniae*
hydrolase	P0AE18	methionine aminopeptidase	Moderate	21	42	*Escherichia Coli*
hydrolase	P0C5C1	beta-lactamase	Moderate	8	65	*Mycobacterium Tuberculosis*
DNA-binding protein	P10275	androgen receptor	Moderate	16	7	*Homo Sapiens*
transferase	P11309	proto-oncogene serine/threonine-protein kinase Pim-1	Hard	31	5	*Homo Sapiens*
oxidoreductase	P11509	Cytochrome P450. family 2. subfamily A. polypeptide 6	Moderate	6	35	*Homo Sapiens*
hydrolase	P11838	endothiapepsin	Unfeasible	12	2	*Cryphonectria Parasitica*
transferase	P12758	uridine phosphorylase	Easy	5	8	*Escherichia Coli*
isomerase	P14174	macrophage migration inhibitory factor	Unfeasible	16	3	*Homo Sapiens*
transferase	P14324*	farnesyl pyrophosphate synthetase	Moderate	7	26	*Homo Sapiens*
transferase	P14325	farnesyl pyrophosphate synthetase	Hard	15	7	*Homo Sapiens*
Ileal lipid binding protein	P15090	fatty acid-binding protein. adipocyte	Moderate	8	27	*Homo Sapiens*
oxidoreductase	P15121	aldose reductase	Moderate	27	49	*Homo Sapiens*
oxidoreductase	P16184	dihydrofolate reductase	Easy	7	9	*Pneumocystis Carinii*
transferase	P17612	cAMP-dependent protein kinase. alpha-catalytic subunit	Moderate	13	58	*Homo Sapiens*
hydrolase	P18031	protein (protein-tyrosine phosphatase 1B)	Moderate	30	43	*Homo Sapiens*
oxidoreductase	P22906	dihydrofolate reductase	Easy	8	10	*Candida Albicans*
hydrolase	P23470	receptor-type tyrosine-	Easy	5	11	*Homo Sapiens*

		protein phosphatase gamma				
ligase	P24182	biotin carboxylase	Moderate	12	52	*Escherichia Coli*
hydrolase	P24627	lactotransferrin	Unfeasible	12	8	*Bos Taurus*
transferase	P24941	cyclin-dependent kinase 2	Moderate	105	45	*Homo Sapiens*
bromodomain and extra-terminal protein	P25440	bromodomain-containing protein 2	Moderate	11	8	*Homo Sapiens*
hydrolase	P25774	cathepsin S	Moderate	14	9	*Homo Sapiens*
hydrolase	P25779	cruzain	Moderate	7	56	*Trypanosoma Cruzi*
hydrolase	P27487	dipeptidyl peptidase IV soluble form	Moderate	39	21	*Homo Sapiens*
transferase	P28482	mitogen-activated protein kinase 1	Moderate	7	31	*Homo Sapiens*
transferase	P28523	casein kinase II. alpha chain	Hard	19	8	*Zea Mays*
oxidoreductase	P28845	corticosteroid 11-beta-dehydrogenase isozyme 1	Moderate	9	59	*Homo Sapiens*
transferase	P30291	wee1-like protein kinase	Easy	7	12	*Homo Sapiens*
isomerase	P30405	peptidyl-prolyl cis-trans isomerase F. mitochondrial	Moderate	7	44	*Homo Sapiens*
transferase	P35557	glucokinase isoform 2	Moderate	7	32	*Homo Sapiens*
transferase	P35968	vascular endothelial growth factor receptor 2	Moderate	8	10	*Homo Sapiens*
transferase	P36897	TGF-beta receptor type I	Easy	5	13	*Homo Sapiens*
hydrolase	P39900	gacrophage metalloelastase	Moderate	17	24	*Homo Sapiens*
chaperone	P41148	endoplasmin	Moderate	8	54	*Canis Lupus Familiaris*
oxidoreductase	P42330	aldo-keto reductase family 1 member C3	Hard	10	6	*Homo Sapiens*
hydrolase	P42574	caspase-3	Unfeasible	8	6	*Homo Sapiens*
hydrolase	P43235	cathepsin K	Moderate	14	66	*Homo Sapiens*
hydrolase	P45452	collagenase 3	Moderate	12	18	*Homo Sapiens*
transferase	P47811	mitogen-activated protein kinase 14	Moderate	10	20	*Mus Musculus*
transferase	P48736	phosphatidylinositol-4.5-bisphosphate 3-kinase catalytic subunit gamma isoform	Hard	5	15	*Homo Sapiens*
transferase	P49841	glycogen synthase kinase-3 beta	Hard	13	10	*Homo Sapiens*
hydrolase	P50579	protein (methionine aminopeptidase)	Hard	9	18	*Homo Sapiens*
oxidoreductase	P51857	3-oxo-5-beta-steroid 4-dehydrogenase	Moderate	6	60	*Homo Sapiens*
transferase	P51955	serine/threonine-protein kinase NEK2	Easy	10	18	*Homo Sapiens*
hydrolase	P52700	metallo-beta-lactamase L1	Hard	7	4	*Stenotrophomonas Maltophilia*
transferase	P53779	mitogen-activated protein kinase 10	Hard	16	9	*Homo Sapiens*
transferase	P54760	ephrin type-B receptor 4	Easy	9	14	*Homo Sapiens*
hydrolase	P56658	adenosine deaminase	Easy	9	19	*Bos Taurus*
hydrolase	P56817	beta-secretase 1	Moderate	63	36	*Homo Sapiens*
hydrolase	P59071	phospholipase A2	Unfeasible	17	7	*Daboia Russellii Pulchella*
hydrolase	P61823	pancreatic ribonuclease A	Easy	9	21	*Bos Taurus*
transferase	P68400	casein kinase II subunit alpha	Hard	14	12	*Homo Sapiens*
hydrolase	P78536	ADAM 17	Moderate	15	29	*Homo Sapiens*

oxidoreductase	P80025	lactoperoxidase	Hard	6	17	*Bos Taurus*
oxidoreductase	Q00511	uricase	Moderate	8	28	*Aspergillus Flavus*
oxidoreductase	Q02127	dihydroorotate dehydrogenase. mitochondrial	Moderate	8	11	*Homo Sapiens*
transferase	Q04771	activin receptor type-1	Moderate	5	68	*Homo Sapiens*
hydrolase	Q07343	cAMP-specific 3'.5'-cyclic phosphodiesterase 4B	Moderate	14	61	*Homo Sapiens*
hydrolase	Q08499	cAMP-specific 3'.5'-cyclic phosphodiesterase 4D	Hard	14	1	*Homo Sapiens*
hydrolase	Q10714	angiotensin converting enzyme	Easy	6	15	*Drosophila Melanogaster*
isomerase	Q13526	peptidyl-prolyl cis-trans isomerase NIMA-interacting 1	Moderate	23	38	*Homo Sapiens*
transferase	Q16539	p38 MAP kinase	Moderate	75	33	*Homo Sapiens*
lyase	Q3JRA0	2-C-methyl-D-erythritol 2.4-cyclodiphosphate synthase	Hard	5	13	*Burkholderia Pseudomallei*
ligase	Q57834	tyrosyl-tRNA synthetase	Easy	7	16	*Methanocaldococcus Jannaschii*
oxidoreductase	Q581W1	pteridine reductase 1	Moderate	9	12	*Trypanosoma Brucei Brucei*
DNA-binding protein	Q92731	estrogen receptor beta	Moderate	18	55	*Homo Sapiens*
transferase	Q9BJF5	calmodulin-domain protein kinase1	Easy	12	17	*Toxoplasma Gondii*
hydrolase	Q9BZP6	acidic mammalian chitinase	Hard	5	16	*Homo Sapiens*
hydrolase	Q9L5C8	beta-lactamase CTX-M-9	Moderate	14	62	*Escherichia Coli*
hydrolase	Q9QYJ6	phosphodiesterase-10A	Moderate	10	57	*Rattus Norvegicus*
oxidoreductase	Q9T0N8	cytokinin dehydrogenase 1	Moderate	8	13	*Zea Mays*
hydrolase	Q9Y233	cAMP and cAMP-inhibited cGMP 3'. 5'-cyclic phosphodiesterase 10A	Moderate	9	69	*Homo Sapiens*

[a] UniProt ID. [b] Number of PDB entries belonging to each cluster. [c] Numbering from highest to lowest performance as obtained from $\log P_{o/w}^{ele}$ / $\log P_{o/w}^{cav}$ / HB (Figure 5). [*] Allosteric binding.

Table S2. Effect of weighting factors on the alignment accuracy upon combination of total hydrophobicity and HB fields. *Average* value (%) of successful overlays for all ligand-template pairs, and *best* value (%) obtained from the pose with the highest similarity score.

		Weight ($\log P_{o/w}$/HB**)**									
Subset	**System**	90/10		80/20		60/40		50/50		40/60	
		Average	*Best*	*Average*	*Best*	*Average*	*Best*	*Average*	*Best*	*Average*	*Best*
I	CDK2	10	66	10	68	11	71	11	73	11	70
	Elastase	14	0	14	0	14	0	18	0	18	0
	ER	33	84	33	84	31	84	31	84	29	84
	HIV (1)	18	64	21	60	23	64	24	67	23	64
	MAPK14	20	69	21	69	20	61	20	53	18	46
	Rhinovirus	72	100	70	100	69	100	69	100	75	100
	Thermolysin (1)	27	58	30	58	32	66	33	66	36	66
	Trypsin	55	71	55	57	55	71	55	71	55	71
	HIV (2)	25	50	25	50	32	40	35	50	34	50
	Total (155)	23.7	64.7	24.1	64.1	23.8	65.9	24.3	67.1	24.4	64.9
II	Chromanone	91	100	89	100	85	100	81	100	76	100
	Cruzain	68	96	69	96	69	96	68	96	68	96
	Dopamine	22	95	22	92	20	92	19	92	18	92
	GSK-3β	52	94	52	100	52	100	52	100	52	100
	Thermolysin (2)	23	90	24	89	33	90	33	91	33	93
	Total (255)	46.0	94	46.0	95.0	47.8	95.3	47.1	95.6	46.1	96.2
I+II	Total (410)	36.9	83	37.2	83.3	38.8	84.2	38.6	84.8	37.9	84.3

Table S3. Effect of weighting factors on the alignment accuracy upon combination of electrostatic and cavitation components of the molecular hydrophobicity. *Average* value (%) of successful overlays for all ligand-template pairs, and *best* value (%) obtained from the pose with the highest similarity score.

		Weight ($\log P_{o/w}^{ele} / \log P_{o/w}^{cav}$)					
Subset	**System**	50/50		40/60		20/80	
		Average	*Best*	*Average*	*Best*	*Average*	*Best*
I	CDK2	17	84	18	85	20	82
	Elastase	29	28	39	42	35	42
	ER	44	100	44	100	44	100
	HIV (1)	42	57	45	64	46	57
	MAPK14	18	53	20	61	24	69
	Rhinovirus	63	87	61	87	86	100
	Thermolysin (1)	30	50	30	58	28	66
	Trypsin	57	71	59	71	61	71
	HIV (2)	61	50	64	50	58	40
	Total (155)	32.4	70.0	34.1	73.6	36.1	72.5
II	Chromanone	93	100	94	100	94	100
	Cruzain	98	100	97	100	98	100
	Dopamine	42	100	42	100	40	100
	GSK-3β	76	100	76	100	74	100
	Thermolysin (2)	37	95	37	95	34	95
	Total (255)	64.2	98.5	64.3	98.5	62.6	98.5
I+II	Total (410)	52.21	87.8	52.9	89.1	52.7	88.7

Table S4. Effect of weighting factors on the alignment accuracy upon combination of electrostatic and cavitation components of the molecular hydrophobicity and HB fields. *Average* value (%) of succesful overlays for all ligand-template pairs.

		Weight ($\log P_{o/w}^{ele}$ / $\log P_{o/w}^{cav}$ / HB)							
Subset	**System**	25/65/10	35/55/10	45/45/10	20/60/20	30/50/20	40/40/20	25/45/30	35/35/30
I	CDK2	21	20	19	22	20	19	21	20
	Elastase	35	37	24	29	27	24	29	22
	ER	44	44	45	43	43	44	43	43
	HIV (1)	45	43	41	47	45	42	46	45
	MAPK14	23	22	20	22	21	20	22	22
	Rhinovirus	77	70	69	75	69	72	75	72
	Thermolysin (1)	31	33	31	34	35	32	33	33
	Trypsin	63	63	59	73	69	65	71	61
	HIV (2)	61	64	63	59	63	64	62	66
	Total (155)	36.2	35.5	33.6	36.9	35.5	34.3	36.4	35.2
II	Chromanone	94	94	93	93	93	92	93	92
	Cruzain	96	96	96	96	96	95	95	95
	Dopamine	42	43	45	43	43	44	42	44
	GSK-3β	76	77	77	76	77	76	75	75
	Thermolysin (2)	39	40	41	42	44	44	46	46
	Total (255)	64.7	65.4	65.9	65.6	66.5	66.1	65.6	66.4
I+II	Total (410)	53.9	54.1	53.7	54.8	54.8	54.1	55.2	54.6

Table S5. Effect of weighting factors on the alignment accuracy upon combination of electrostatic and cavitation components of the molecular hydrophobicity and HB fields. *Best* value (%) obtained from the pose with the highest similarity score

		Weight ($\log P_{o/w}^{ele} / \log P_{o/w}^{cav}$ / HB**)**							
Subset	**System**	25/65/10	35/55/10	45/45/10	20/60/20	30/50/20	40/40/20	25/45/30	35/35/30
I	CDK2	85	87	87	89	89	87	91	89
	Elastase	42	57	42	42	57	42	42	42
	ER	84	84	92	92	92	92	92	92
	HIV (1)	60	53	57	64	53	57	57	60
	MAPK14	84	61	76	84	84	76	84	76
	Rhinovirus	100	100	100	100	100	100	100	100
	Thermolysin (1)	75	58	50	66	75	66	75	66
	Trypsin	71	71	71	85	85	85	85	85
	HIV (2)	40	50	40	60	60	60	60	60
	Total (155)	74.8	72.3	73.1	78.9	78.3	76.3	79.1	77.5
II	Chromanone	100	100	100	100	100	100	100	100
	Cruzain	100	100	100	100	100	100	100	100
	Dopamine	100	100	100	100	100	100	100	100
	GSK-3β	100	100	100	100	100	100	100	100
	Thermolysin (2)	94	93	94	95	95	94	94	94
	Total (255)	98.3	98	98.2	98.5	98.5	98.3	98.3	98.3
I+II	Total (410)	89.4	88.3	88.7	91.13	90.9	89.9	91.0	90.42

Table S6. Effect of weighting factors on the alignment accuracy upon combination of electrostatic and steric fields. *Average* value (%) of successful overlays for all ligand-template pairs, and *best* value (%) obtained from the pose with the highest similarity score.

		Weight (Electrostatic/Steric)							
Subset	**System**	50/50		40/60		30/70		20/80	
		Average	*Best*	*Average*	*Best*	*Average*	*Best*	*Average*	*Best*
I	CDK2	17	80	16	78	16	78	16	84
	Elastase	35	42	35	42	31	57	31	71
	ER	36	69	36	69	32	69	33	76
	HIV (1)	46	67	45	67	44	67	43	67
	MAPK14	22	76	22	84	22	92	22	84
	Rhinovirus	69	100	72	100	66	100	63	100
	Thermolysin (1)	19	50	18	66	18	66	15	66
	Trypsin	61	85	65	85	61	85	57	71
	HIV (2)	64	70	61	60	58	60	57	60
	Total (155)	32.9	73	32.4	73.5	31.1	74.8	30.3	77
II	Chromanone	91	100	87	100	85	100	84	100
	Cruzain	95	100	96	100	95	100	95	100
	Dopamine	35	100	34	100	31	100	30	100
	GSK-3β	76	100	76	100	76	95	75	94
	Thermolysin (2)	22	94	21	95	19	95	18	95
	Total (255)	58.12	98.2	57.26	98.5	55.81	97.1	55.94	96
I+II	Total (410)	48.59	88.7	47.88	89.1	45.63	88.7	48,10	89.3

Figure S2. Representation of the accuracy (%) in molecular overlays obtained from $\log P_{o/w}^{ele}$/ $\log P_{o/w}^{cav}$ /HB fields versus the global hydrophobicity of ligands determined by using AlogP.

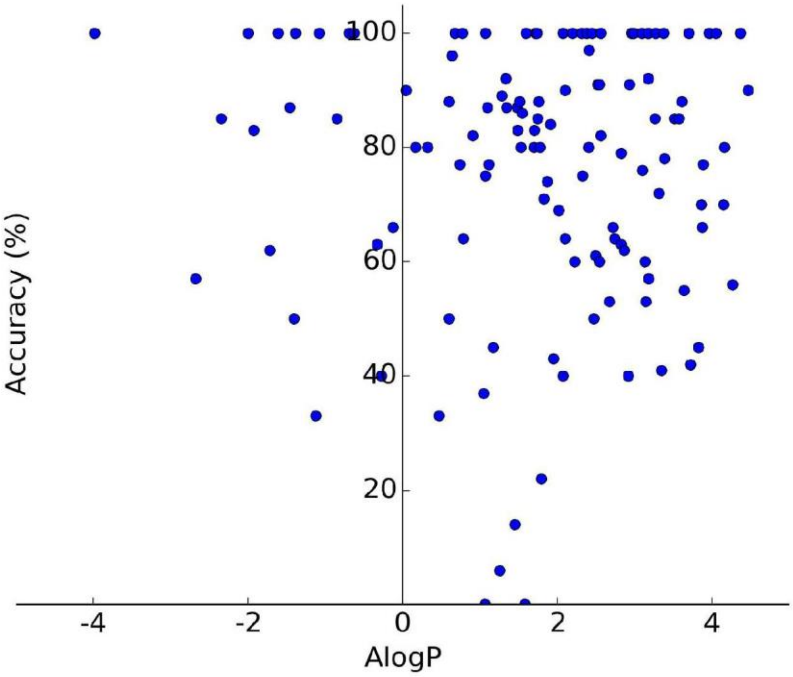
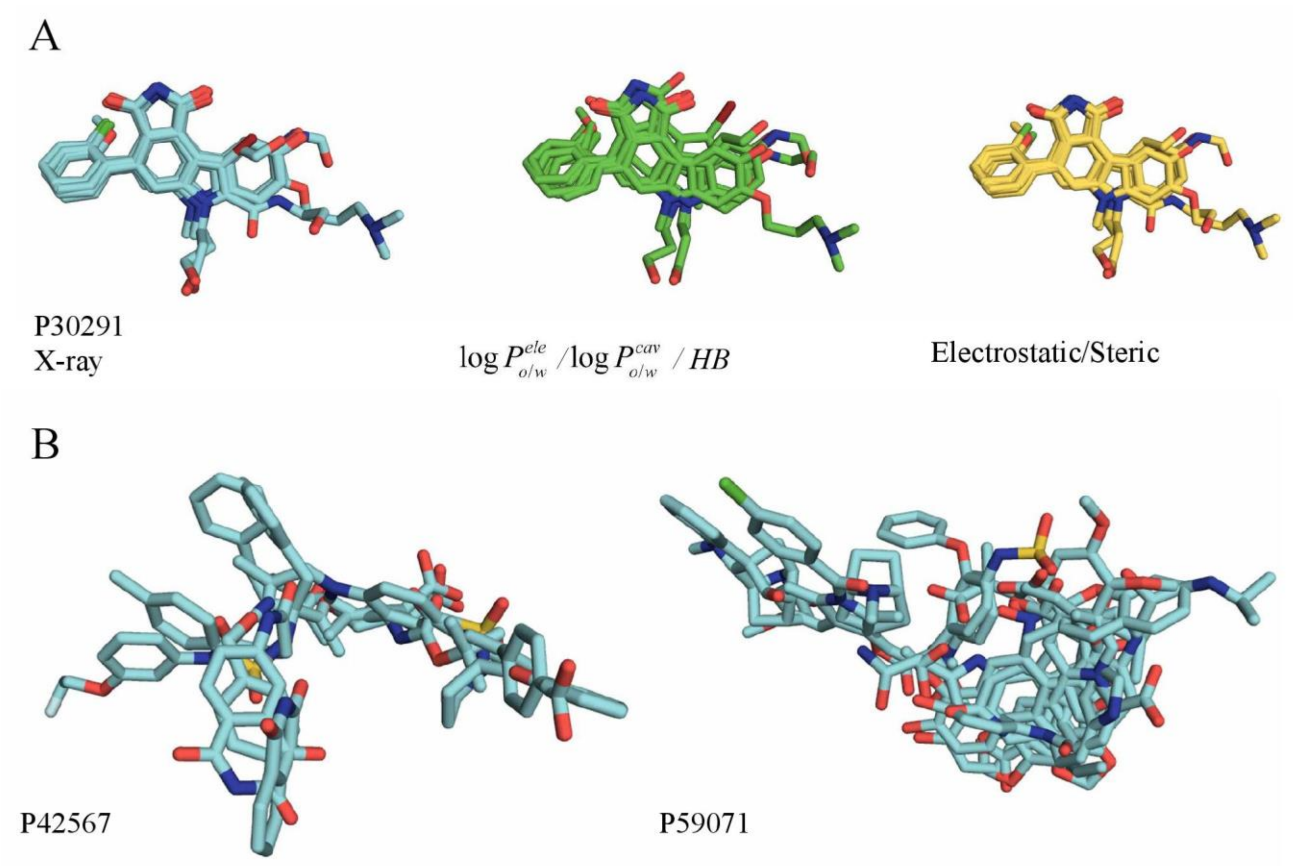


Figure S3. (A) Superposition of all crystallographic ligands of P30291 (Tyrosine-protein kinase), chosen as an example of easy target, and molecular alignments obtained from $\log P_{o/w}^{ele}$/ $\log P_{o/w}^{cav}$/HB and electrostatic/steric fields. Compared with the crystallographic structure all ligands have a RMSD ≤ 2 Å. (B) Superposition of all crystallographic ligands of P42567 (Epidermal growth factor receptor substrate 15; left) and P59071 (Basic phospholipase A2 VRV-PL-VIIIa; right), chosen as examples of unfeasible targets.



3.2 **PAPER 2:** *“Lipophilicity in drug design: an overview of lipophilicity descriptors in 3D-QSAR studies”*

Tiziana Ginex, † Javier Vazquez, †, ‡ Enric Gilbert, ‡ Enric Herrero, ‡ and Francisco J Luque, †

† Department of Nutrition, Food Sciences & Gastronomy, Faculty of Pharmacy & Food Sciences, Campus Torribera, Institute of Biomedicine (IBUB), & Institute of Theoretical & Computational Chemistry (IQTC-UB), University of Barcelona, Av. Prat de la Riba 171, Santa Coloma de Gramenet E-08921, Spain

‡ Pharmacelera, Plac¸a Pau Vila, 1, Sector 1, Edificio Palau de Mar, Barcelona 08039, Spain

For reprint orders, please contact: reprints@future-science.com

Future Medicinal Chemistry

Lipophilicity in drug design: an overview of lipophilicity descriptors in 3D-QSAR studies

Tiziana Ginex*,[1], Javier Vazquez[1,2], Enric Gilbert[2], Enric Herrero[2] & Francisco J Luque**,[1]

[1] Department of Nutrition, Food Sciences & Gastronomy, Faculty of Pharmacy & Food Sciences, Campus Torribera, Institute of Biomedicine (IBUB), & Institute of Theoretical & Computational Chemistry (IQTC-UB), University of Barcelona, Av. Prat de la Riba 171, Santa Coloma de Gramenet E-08921, Spain

[2] Pharmacelera, Plaça Pau Vila, 1, Sector 1, Edificio Palau de Mar, Barcelona 08039, Spain

*Author for correspondence: tiziana.ginex@ub.edu

**Author for correspondence: fjluque@ub.edu

The pharmacophore concept is a fundamental cornerstone in drug discovery, playing a critical role in determining the success of *in silico* techniques, such as virtual screening and 3D-QSAR studies. The reliability of these approaches is influenced by the quality of the physicochemical descriptors used to characterize the chemical entities. In this context, a pivotal role is exerted by lipophilicity, which is a major contribution to host–guest interaction and ligand binding affinity. Several approaches have been undertaken to account for the descriptive and predictive capabilities of lipophilicity in 3D-QSAR modeling. Recent efforts encode the use of quantum mechanical-based descriptors derived from continuum solvation models, which open novel avenues for gaining insight into structure–activity relationships studies.

First draft submitted: 30 August 2018; Accepted for publication: 4 February 2019; Published online: 25 February 2019

Keywords: 3D-QSAR • continuum solvation models • hydrophobic pharmacophore • lipophilicity • quantum mechanical-derived descriptors

The pharmacophore concept & its application in drug design

Almost all processes of life are determined by the recognition between biomolecules, a process dictated by the chemical complementarity between the interacting partners [1]. An effective characterization of the chemical features associated with the structure of both 'host' and 'guest' is necessary for disclosing the key molecular determinants implicated in the formation of the host–guest complex. In drug discovery studies addressing the interaction of small molecules (ligands) with macromolecular receptors, these determinants are generally encoded under the concept of pharmacophore. A simple and intuitive definition can be attributed to Paul Ehrlich, since this concept can be related to *"a molecular framework that carries (phoros) the essential features responsible for a drug's (pharmacon) biological activity"* [2]. Nevertheless, Ehrlich did not use the term *pharmacophore* in his papers, where the terms 'haptophore' and 'toxophore' were adopted [3]. Instead, the modern concept of pharmacophore evolved from the identification of 'chemical groups' to the definition as *"patterns of abstract features in space"* by Schueler [4], reflected in early models depicting key features for biological activity that must satisfy certain geometrical relationships [5,6], and the development of the first pharmacophore pattern recognition programs [7]. Thus, according to the International Union of Pure and Applied Chemistry (IUPAC), a pharmacophore *"does not represent a real molecule or a real association of functional groups, but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds toward their target structure,"* being the largest common denominator shared by a set of active molecules [8].

This evolution has been accompanied by the progressive refinements triggered by advances in molecular descriptors and computational methods seen in the last 30 years, since a variety of *in silico* techniques have exploited the pharmacophore concept. This is exemplified by virtual screening (VS) studies of large molecular databases performed to identify new promising compounds according to their similarity to a given privileged template, which should contain reference physicochemical features relevant for biological activity [9–11]. Molecular/chemical (global/local) similarity is a subjective concept since it depends on the specific details of the methodological ap-

newlands press

proach, the nature of the molecular features relevant for similarity assessment, and the definition of the similarity function [12]. A sensitive and effective estimation of molecular similarity is a fundamental pre-requisite for the identification of potential leads starting from a chemical reference, which represents the paradigm of VS.

Another successful application of the pharmacophore concept is linked to 3D-quantitative structure–activity relationships (3D-QSAR) [13], such as CoMFA [14], CoMSIA [15] and GRID/GOLPE [16]. These methods permit to identify a pharmacophore from the relationships between the biological activities of a set of aligned molecules and the projection of selected physicochemical descriptors into the surrounding space, leading to the disclosure of regions favorable or not to the bioactivity of compounds. 3D-QSAR approaches are also used to model ADME(T) properties in the attempt to predict whether a molecular candidate would be able to achieve its biological target [17]. Optimization of both ligand potency and ADME(T) profile is absolutely required to translate promising molecular candidates to successful low-dose therapeutics. However, the success of this operation is not trivial, since the final result depends on factors such as the quality of the input data, as well as the adequacy and level of description of the physicochemical parameters used in the analysis. In fact, Gleeson and collaborators [18] have observed the existence of a diametrically opposed relationship between descriptors that efficaciously model drug potency and ADME(T) properties, making more challenging the drug discovery process.

Lipophilicity in drug design

The relevance of lipophilicity in understanding the pharmacological profile of drug-like compounds is widely recognized [19], as a broad variety of biodistribution and toxicological processes are ultimately related to the differential solubility of solutes in aqueous and nonaqueous environments. This is illustrated by Lipinski's rule-of-five [20], which relates the drug-likeness of oral compounds with molecular weight, hydrogen bonding and lipophilicity. Being a key property for the prediction of ADME(T) properties, this has stimulated the development of experimental and computational approaches to quantify the lipophilicity of a (bio)organic molecule.

Experimentally, the lipophilicity of a molecule can be quantified by its partition coefficient (*P*), as this equilibrium thermodynamic property measures the ratio of concentrations of the compound between two immiscible solvents, generally water and *n*-octanol. In turn, the partition coefficient can be expressed in terms of the transfer free energy ($\Delta G_{tr}^{o/w}$) between the two solvents (Equation 1).

$$\Delta G_{tr}^{o/w} = -2.303\,RT\log P \qquad \text{(Eq. 1)}$$

Lipophilicity reflects the complex interplay between the intermolecular forces that dictate the differential solvation in the aqueous and organic phases. Accordingly, it can be factorized in terms of selected physicochemical properties of the compound that may be relevant for the preferential solvation in aqueous and nonaqueous solvents, as shown in Equation 2 [21], and references therein.

$$\log P = vV - \Lambda + I + IE \qquad \text{(Eq. 2)}$$

where v is a constant, V is the molar volume, which encompasses the ability of the solute to elicit nonpolar interactions, Λ is related to the polarity of the compound, and finally I and IE accounts for the solute capacity to form ionic interactions, which favor partitioning into the aqueous phase, and for the contribution due to intramolecular effects, respectively.

Let us note that lipophilicity and hydrophobicity, which are often used as equivalent concepts, are not strictly synonymous, the latter being in fact one of the contributions to molecular lipophilicity [22]. Thus, while hydrophobicity can be defined as the tendency of nonpolar groups of a molecule to aggregate in order to minimize the unfavorable exposition to the surrounding polar (water) solvent, lipophilicity is a measure of the affinity of the molecule for the nonpolar solvent in a biphasic system constituted by a polar and a nonpolar solvent.

Lipophilicity affects a number of pharmacokinetic parameters (Figure 1). Low lipophilicity is responsible of high aqueous solubility, which is a key factor for drug-likeness, but an excessively low lipophilicity could compromise the ability of the drug to achieve the biological target. On the opposite site, highly soluble compounds possess poor permeability through biological membranes, limiting absorption along the gastrointestinal tract or the transport across the blood–brain barrier. Therefore, optimal requirements for efficient solubility and permeability properties are inevitably enclosed in a very narrow range of lipophilicity. Another key aspect for drug-likeness is bioavailability, which is inversely correlated to low first-pass clearance. Once again, lipophilicity is crucial since high lipophilicity

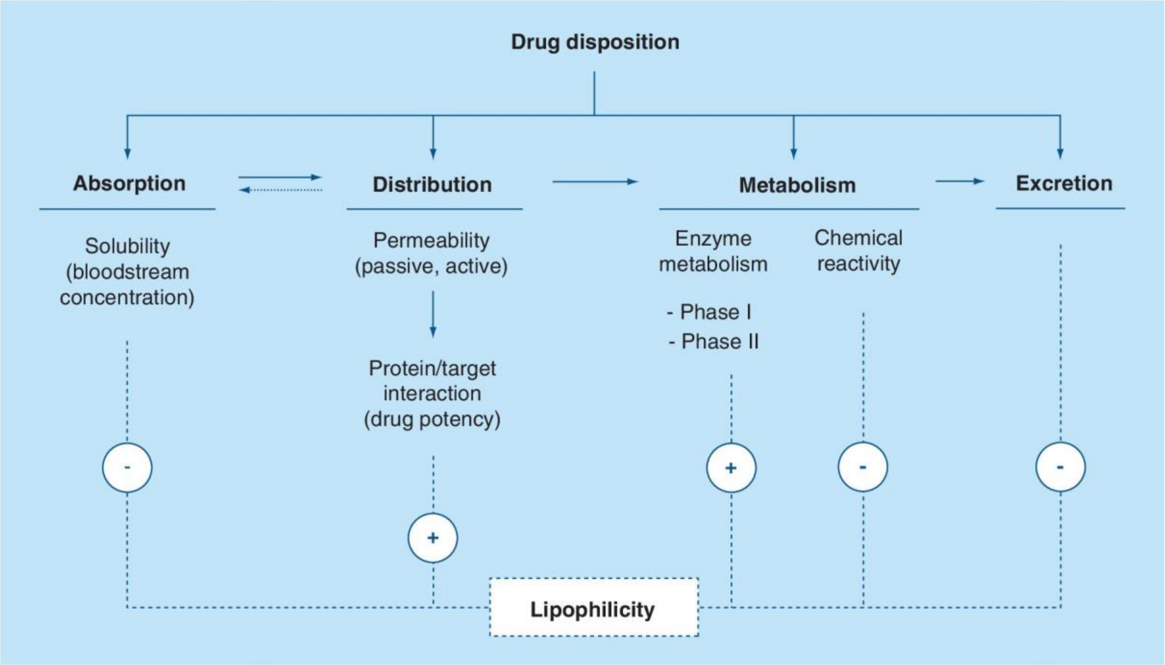


Figure 1. Schematic representation of the central role of lipophilicity in drug potency and pharmacokinetics profile. Direct (+) and inverse (-) correlation of lipophilicity with each of the main steps of ADME process are also highlighted.

is associated with high clearance and low metabolic stability. Overall, a careful handling of lipophilicity is required to optimize compound availability at the biological target.

On the other hand, lipophilicity has rarely been used as the primary descriptor in ligand–receptor recognition. Indeed, following the IUPAC recommendation for the definition of a pharmacophore, it is defined as *"the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure"* [8]. This definition hides the key role played by (de)solvation in the recognition and binding of a drug-like compound to its macromolecular target [23], especially keeping in mind that the maximal achievable affinity that can be attained for target binding sites is largely influenced by nonpolar desolvation [24]. This is consistent with the concept that favorable drug binding is largely driven not only by the global lipophilicity of a compound, but more importantly by the spatial distribution of polar and apolar regions along the chemical skeleton. Thus, while apolar regions determine the binding affinity with complementary lipophilic regions of the binding site, polar interactions would provide 'anchor points' contributing to ligand specificity and/or directionality in the binding pocket, as well as to modulate binding kinetics of the ligand [25–30].

Taken together, these data suggest that a concomitant optimization of both pharmacokinetic profile and drug potency have to be done to obtain successful drug products. This is encoded in the concept of lipophilicity efficiency (LipE), which provides a metric that normalizes the potency (generally measured as K_i or IC_{50}) of the ligand against a protein target for the lipophilicity of the compound [31–33]. This is achieved by substracting the *logP* (or the distribution coefficient for ionizable molecules, *logD*) from the negative logarithm of the potency (Equation 3).

$$\mathrm{lipE} = -\log(\mathrm{potency}) - \log\mathrm{P} \qquad \text{(Eq. 3)}$$

Lipophilicity efficiency can be useful to provide guidelines to study the simultaneous effects exerted by structural changes on potency and lipophilicity, which is central for drug design and lead optimization programs, thus giving support to the formulation of the 'lipophilic pharmacophore' concept.

From empirical fragment/atom-based approaches to 3D structure-based methods to estimate lipophilicity

Numerous efforts have been done to assess lipophilicity by means of experimental methods [34–36]. Similarly, a plethora of computational approaches for estimating log*P* have also been developed [37–42]. We limit ourselves to remark selected fundamental concepts, while the reader is addressed to the previously quoted reviews for detailed comparative analysis.

Within the framework of substructure-based methods for log*P* estimation, fragmental and atom-based techniques follow a general additive scheme as shown in Equation 4,

$$\mathrm{logP} = \sum_{i=1}^{n} a_i f_i + \sum_{j=1}^{m} b_j F_j \qquad \text{(Eq. 4)}$$

where *logP* is the sum of the weighted (a_i) contribution of each fragment/atom (f_i) and a correction factor (b_jF_j).

Fragmental methods are illustrated by the work of Leo, Hansch and Elkins [43] as well as Nys and Rekker [44]. The former relies on the concept of substituent constant, which encodes the lipophilicity contribution of a chemical group or atom when it replaces a hydrogen atom in a reference compound, and the theoretical estimation of log*Po/w* follows an additivity scheme, named cLOGP. This method permits to extrapolate the partition coefficients starting from a list of experimentally fitted fragmental contributions to lipophilicity. An arbitrary set of interfragmental rules was then used to compile a database library of fragment-weighted lipophilicity contributions. On the other hand, Nys and Rekker [44] introduced the concept of hydrophobic fragmental constant (*f*), which represents the lipophilicity contribution of a constituent part of a structure to the total lipophilicity of a given compound. Fragments range from atoms to heterocyclic rings, so that functional groups with direct contribution to resonance interactions were left intact, and are differentitated upon linkage to aliphatic and aromatic structures. The differences between experimental *logP* and the additive value estimated from the $\sum f$ approach was accounted for by correction rules, reflecting factors such as the presence of vicinal electronegative centers in the chemical structure, aromatic condensation, cross-conjugation or hydrogen-bonding [45].

An example of atom-based partitioning strategy was undertaken by Ghose and Crippen, who developed a procedure that combines lipophilicity contributions at an atomic level leading to the ALOGP method. This method encompassed a list of 120 atom types for carbon, hydrogen, oxygen, nitrogen, sulfur and halogens [46–48]. An alternative strategy is the XLOGP method [49], which is based on the summation of atomic contributions derived from experimental lipophilicity data of 1831 organic molecules, and includes correction factors for some intramolecular interactions.

In the last decades, the evolution of computer performances enabled the development of whole molecule-based strategies to predict the lipophilicity by taking into account the 3D-structure of compounds, and thus the effect of molecular conformation. Among all the available techniques, the molecular lipophilicity potential (MLP) [50] offers an empirical quantitative 3D-description of the lipophilicity potential from all the molecular fragments on the surrounding space of a compound. The MLP approach is then intended to model the lipophilic interactions between ligand and receptor as noted in Equation 5,

$$\mathrm{MLP_K} = \sum_{i=1}^{N} F_i\, f(d_{ik}) \qquad \text{(Eq. 5)}$$

where F_i is the lipophilic fragmental contribution and $f(d_{ik})$ is a distance function which depends on the separation between a given fragment (*i*) and any point on the molecular surface or volume (*k*).

Molecular fields derived from the MLP potential have found a wide range of pharmaceutical applications, including the prediction of skin permeation and distribution of new chemical entities [51], modeling of peptides and proteins [52,53], and structure–activity relationships studies [54].

The Hydrophobic INTeraction (HINT) method represents an alternative, promising strategy for the study of lipophilicity in biomolecular interactions [55,56]. This method exploits a scale of hydrophobic fragments constants at the atomic level by means of an adaptation of the CLOGP method, which are then used to evaluate a pairwise interaction energy term (b_{ij}) between atoms *i* and *j* in the interacting partners according to Equation 6,

$$b_{ij} = a_i S_i a_j S_j T_{ij} R_{ij} + r_{ij} \qquad \text{(Eq. 6)}$$

where a_i and S_i are respectively the hydrophobic constant and the accessible surface area of the atom *i*, T_{ij} is a logic function describing the character of interacting pairs (attraction or repulsion), and R_{ij} and r_{ij} denote functions of the distance between atoms i and j, the former following an exponential form and the latter a Lennard–Jones implementation.

Equation 5 encodes the formalism of the 'natural' HINT force-field, which has been used to explore a variety of applications in ligand–protein and protein–protein interactions [57–61].

Other approaches have relied on molecular properties derived from quantum mechanical treatments of molecules. An early attempt is the work by Roger and Cammarata [62,63], who related the logP of aromatic compounds with the charge density of both π and σ electron frameworks and the induced polarization. In a distinct approach, the BLOGP method relied on semiempirical AM1 calculations to derive geometrical and quantum chemical descriptors for the prediction of logP [64,65]. In a similar approach, Clark and coworkers performed AM1 and PM3 calculations to derive a series of descriptors, including electrostatic potentials, total dipole moments, mean polarizabilities, surfaces, volumes and charges, which were used in the prediction of partition coefficients [66,67].

These efforts can also be exemplified with the concept of heuristic MLP [68,69]. In this approach, the lipophilic/hydrophilic features of a compound are determined from the analysis of the electrostatic potential computed at the molecular surface. To this end, a dimensionless distance-dependent screening function is used to compare the local electron density at the surface of a given atom with the electrostatic potential generated on the rest of atoms. The screening function, which was derived from statistical mechanical treatment of polar solvent molecules as dipoles, accounts for the influence exerted by the atomic descriptors of the electrostatic potential from surrounding atoms. Ultimately, such a comparison leads to the definition of an atomic lipophilicity index, which can adopt positive or negative values, reflecting the lipophilic and hydrophilic nature, respectively, of such an atom.

Finally, a distinct approximation comes from the usage of solute–solvent correlation functions derived by using the reference interaction site model (RISM) as descriptors for QSAR studies. By using a classical statistical mechanics-based solvent model combined with machine learning, 1D solute–solvent correlation functions were used to predict Caco-2 cell permeabilities [70]. As an extension of this approach, Gussregen *et al.* proposed the Comparative Analysis of 3D-RISM Maps (CARMa) methodology [71]. In this computational strategy, the classical electrostatic and steric fields generally used in CoMFA are replaced by solute–solvent distribution functions determined from 3D-RISM computations, which are subsequently treated as descriptors to perform QSAR analysis. The method was validated using a set of serine protease inhibitors as a test system.

Even though CARMa uses a statistical mechanics solvent model, the electrostatic and steric effects implemented in CoMFA cannot be directly captured. This issue has been recently addressed by solving 3D-RISM equations for a solvent comprising CoMFA probes in aqueous solution, this extension being referred to as CARMa (electrolyte) [72]. The analysis performed for six protein–ligand systems reveals a small but consistent increase in prediction accuracy compared with CoMFA.

Lipophilicity from quantum mechanical continuum solvation methods

More elaborate methods for estimating the partition coefficients have been proposed in the framework of quantum mechanical (QM)-based continuum solvation models [73,74], which were developed with the aim of predicting the solvation free energy of solutes treating the solvent as a continuum polarizable medium. In spite of this rather crude approximation, these methods have proved to be a promising strategy that combines well established physical formalisms, a straightforward mathematical implementation and a reduced computational cost, while predicting solvation free energies of (bio)organic compounds with chemical accuracy after a careful parameterization against experimental data [75–77]. Since a broad review of these formalisms and their applications exceeds the aim of this review, we limit ourselves to stress a selected set of recent studies addressing the potential impact of QM-based continuum methods in drug design.

COSMO & COSMO-RS-based approaches

In this context, the Continuum Solvation Model for Real Solvents (COSMO-RS) has been recently utilized to evaluate the similarity between molecules within the so-called COSMO*sim* method [78]. This method relies on the conductor-like screening model (COSMO) calculations to derive the so-called σ-profile of a given compound. The σ-profile collects the set of polarization charge densities generated on the surface patches of the molecule immersed in the solvent, which is treated as an ideal conductor. The 1D histogram distribution of the σ values for the whole set of surface elements enclosed in the molecular surface gives rise to a characteristic signature of the solute, which can be used to measure a σ-profile-based similarity between compounds with application for the detection of bio-isosteric fragments or molecules. In order to enhance the computational efficiency, the σ-profile of a new compound can be replaced with a composition of partial σ-profiles taken from similar fragments of precalculated molecules stored in a database using COSMOfrag [79].

Since the σ-profile does not contain information about the spatial distribution of the polarization charge density, COSMOsim3D has been recently proposed to alleviate this limitation [80]. To this end, COSMOsim3D projects the surface charge density of each surface segment onto a regular 3D grid, so that each point of the grid has an associated local σ-profile. In other words, instead of generating a single 1D σ-profile for the entire molecule, COSMO*sim3D* creates a local 1D σ-profile at each position of a regular 3D grid. This process leads to a 4D histogram defined by the three Cartesian dimensions of the grid point and the local σ-profile as the fourth dimension. If calculated for two molecules, this strategy can be ultimately used to estimate their overall similarity. Furthermore, these local σ-profiles have been also used to generate molecular interactions fields for 3D-QSAR studies [81].

Fragmental lipophilicity model from the Miertus–Scrocco–Tomasi method: the Hyphar approach

The Miertus–Scrocco–Tomasi (MST) solvation model has been used to develop 3D-distribution patterns of lipophilicity, which in turn have been exploited in predicting molecular overlays and 3D-QSAR studies [82,83]. The MST model is a parametrized version of the polarizable continuum model developed by Tomasi and coworkers [84,85] at both semiempirical, Hartree–Fock and B3LYP levels [86–89] (for a review see [90]). From the solvation free energies in water and *n*-octanol, one can derive the *n*-octanol/water partition coefficient (Equation 1), which is a property of the whole molecule. Nevertheless, by decomposing the solvation free energy into atomic contributions, one can obtain the 3D profile of lipophilicity from the corresponding atomic contributions to the logP. For a molecule (M) containing *N* atoms, this is achieved by decomposing the logP (or the corresponding transfer free energy, $\Delta G^{o/w}_{tr,M}$) into electrostatic ($logP_{ele,i}$), cavitation ($logP_{cav,i}$) and van der Waals ($logP_{vw,i}$) components, which can be derived from the polar ($\Delta G^{o/w}_{ele,i}$) and nonpolar ($\Delta G^{o/w}_{cav,i}$, $\Delta G^{o/w}_{vW,i}$) contributions to the solvation free energy (Equations 7 & 8).

$$\Delta G^{o/w}_{tr,M} = \sum_{i=1}^{N} \Delta G^{o/w}_{tr,i} = \sum_{i=1}^{N} (\Delta G^{o/w}_{ele,i} + \Delta G^{o/w}_{cav,i} + \Delta G^{o/w}_{vW,i}) \qquad \text{(Eq. 7)}$$

$$\log P_M = \sum_{i=1}^{N} \log P_i = \sum_{i=1}^{N} (\log P_{ele,i} + \log P_{cav,i} + \log P_{vW,i}) \qquad \text{(Eq. 8)}$$

Partitioning of the electrostatic term into atomic contributions can be made resorting to a perturbation approximation of the coupling between the solute charge distribution and the solvent reaction field [91], leading to Equation 9,

$$\log P^{o/w}_{ele,i} = \frac{1}{2}\langle \Psi^{\circ} | \sum_{k_{k \in i}}^{k} = 1 \frac{qk^{w}}{|r^{w}_{k} - r|} - \sum_{I_{l \in i}}^{L} = 1 \frac{q1^{\circ}}{|r1^{\circ} - r|} | \Psi^{\circ} \rangle \qquad \text{(Eq. 9)}$$

where Ψ° is the solute wave function in the gas phase, and *K* and *L* stand for the total number of reaction field charges in water (q^{w}_{k}) and *n*-octanol ($\mathring{q}_{i}$), located at positions r^{w}_{k} and $\mathring{r}_{i}$.

The atomic decomposition of the cavitation and van der Waals terms takes advantage of the linear dependence with the solvent-exposed surface of the atoms in the molecule (Equations 10 & 11).

$$\log P^{o/w}_{cav,i} = \sum_{i=1}^{N} \frac{S_i}{S_T} \Delta G^{o/w}_{P,i} \qquad \text{(Eq. 10)}$$

$$\log P^{o/w}_{vW,i} = \sum_{i=1}^{N} S_i \Delta \xi^{o/w} \qquad \text{(Eq. 11)}$$

where $\Delta G_{p,i}^{o/w} = \Delta G_{P,i}^{w} - \Delta G_{P,i}^{o}$, $\Delta G_{P,i}$ being the cavitation free energy of atom i, $\Delta \xi^{o/w} = \xi^{w} - \xi^{o}$, with ξ_i being the atomic surface tension and S_i denotes the contribution of atom i to the total molecular surface (S_T).

In contrast to the COSMO-RS-based approaches, which rely on the concept of σ-profile (see above), the MST-derived applications use the atomic contributions to the thermodynamic components of the differential solvation free energy in water and *n*-octanol, which are encoded under the partition coefficient between these two solvents. Accordingly, they take into account the effect of specific chemical features of the molecule, such as the existence of specific tautomers or conformational species, or the formation of specific intramolecular interactions (i.e., hydrogen bond), in the computation of the 3D-distribution pattern of molecular lipophilicity.

These patterns have been exploited to predict the chemical similarity between compounds [92]. By using the MST-based hydrophobic descriptors $\log P_{eles,i}^{o/w}$ and $\log P_{cav,i}^{o/w}$, a computational procedure has been proposed to identify the molecular overlay that maximizes the lipophilic similarity. To this end, molecular similarity was achieved by comparing the hydrophobic fields generated by the molecules, which were prealigned following multipole expansions of the atomic lipophilic contributions. On the other hand, simple descriptors of the hydrogen-bond (HB) donor/acceptor character of atoms were used to complement the information about the chemical nature of polar atoms in a molecule (briefly, the current implementation assigns an arbitrary value of +1 to hydrogen atoms in HB donors, and -1 to N and O atoms that may act as acceptors). This choice obeys to the fact that the polar nature of hydrophilic groups cannot distinguish the HB donor/acceptor character, as this information is not implicitly encoded by the $\log P_{ele,i}^{o/w}$ term. Hydrophobic and HB properties are then projected into a 3D grid using the exponential function (Equation 12) implemented in CoMSiA [15], and then compared by means of the Tanimoto coefficient.

$$P_q = \sum_{i=1}^{N} w_i e^{-\alpha r_{iq}^2} \qquad \text{(Eq. 12)}$$

The method was implemented in PharmScreen software [83,93] and was successfully used to evaluate the molecular overlay for a collection of 121 molecular systems compiled by AstraZeneca, denoted as the AstraZeneca Overlays Validation Test Set [94]. This set contains molecular overlays experimentally characterized for 119 targets, which were grouped in four categories according to the expected difficulty in predicting the experimental overlay: easy, moderate, hard and unfeasible. The results pointed out that correct overlays were predicted for 94% (easy), 79% (moderate) and 54% (hard) of the cases. Moreover, the overall performance obtained from classical electrostatic/steric descriptors and from Hyphar ones was fairly similar for easy and moderate subsets, but the accuracy obtained with Hyphar for the subset of hard cases exceeded the performance obtained with electrostatic/steric properties. Finally, it was found that the similar performance of Hyphar and electrostatic/steric descriptors does not imply that they lead to identical overlays. Rather, the analysis of the predicted poses revealed that the degree of identity in molecular overlays was reduced with the increase in the difficulty of the target. Overall, these findings point out that Hyphar descriptors may be a valuable alternative for molecule superposition and VS of chemical libraries, especially for targets that may be challenging for predictive molecular similarity techniques.

On the other hand, the atom-centered MST-derived hydrophobic contributions have also been used as physicochemical descriptors to derive 3D-QSAR models using PharmQSAR [82]. MST/IEFPCM calculations were performed for five sets of compounds, including dopamine D2/D4 receptor antagonists, antifungal chromanones, glycogen synthase kinase-3 inhibitors, cruzain inhibitors and thermolysin inhibitors. The compounds in these sets covered a wide range of variance in selected physicochemical properties (molecular weight, hydrogen-bond donor/acceptor, clogP and number of rotatable bonds). The 3D-QSAR models obtained with the hydrophobic pharmacophore (HyPhar) were found to have a predictive accuracy comparable to standard CoMFA and CoMSiA techniques. Moreover, Hyphar descriptors were also valuable to discriminate the selectivity of compounds acting as inhibitors of thrombin, trypsin and factor Xa [83].

Overall, these findings support the usefulness of the MST-derived lipophilic descriptors as a valuable alternative to electrostatic/steric properties to carry out VS of chemical libraries for molecular similarity, as well as to derive 3D-lipophilic pharmacophores, thus providing valuable complementary information to gain insight into the molecular determinants of bioactivity.

A comparative analysis between Hyphar & electrostatic/steric properties

The strength of Hyphar descriptors in 3D-QSAR studies may be attributed to two major features. First, the concept of lipophilicity is very intuitive and widely accepted in medicinal chemistry. Second, the partitioning of lipophilicity, which reflects a property of the whole molecule, into atomic or fragmental contributions permits to obtain a graphical representation of the distribution pattern of polar and apolar regions adapted to the 3D-structure of a given compound. In turn, this paves the way to rationalize the recognition between a small compound and its macromolecular target from the complementarity between hydrophilic and lipophilic groups of the ligand and the polar and apolar nature of the side chains of residues that shape the binding pocket. As an additional remark, let us note that resorting to Hyphar descriptors benefits from the accurate description of the molecular charge distribution that can be attained by QM methods, which may take into account the influence arising from the chemical features of the bioactive compound, such as the ionization state, the preference for a tautomeric species, and the adoption of a given conformational state representative of the binding mode of the ligand.

Given the novelty of MST-based atomic lipophilicity contributions, it is nevertheless necessary to explore their suitability for 3D-QSAR studies. In this context, this section reports the results of a comparative analysis performed to calibrate the performance of Hyphar descriptors through comparison with electrostatic/steric ones. This analysis has been carried out using the comprehensive benchmark dataset compiled by Sutherland and coworkers [95], which comprises 113 ACE inhibitors, 111 AChE inhibitors, 147 ligands for BZR, 282 COX-2 inhibitors, 361 DHFR inhibitors, 66 GPB inhibitors, 74 THER inhibitors and 87 THR inhibitors.

Accordingly, the CoMFA/CoMSiA results reported in [95] were compared with the 3D-QSAR models obtained using Hyphar descriptors, which combine both 'polar' ($logP_{ele,i}$) and 'non-polar' ($logP_{cav,i}$) hydrophobic contributions (see above). To this end, the atomic electrostatic and nonelectrostatic components of the lipophilicity were used to generate the molecular fields through projection into a grid that encloses the set of aligned compounds using a similarity index function (see [82] for further details). For the sake of comparison, the original molecular geometries and protonation states of compounds were kept in this study. All the details about models generation, grid dimensions and points, training/test sets, and related activity ranges for the eight sets compiled by Sutherland are reported in Supplementary Tables 1–3. Only for the THERM dataset partition between training and test sets was made as indicated in [15].

As a preliminary step, the effect of the QM method selected to derive the hydrophobic contributions on the performance of the 3D-QSAR Hyphar models was evaluated for a subset of four systems (D2 inhibitors, antifungal chromanones, GSK3-β and cruzain inhibitors) taken from our previous study [82]. To this end, Hyphar descriptors were derived from continuum computations performed with the MST version parametrized for the semiempirical RM1 method [96], and alternatively with the version parametrized at the B3LYP/6-31G(d) level [89]. Comparison of the statistical parameters obtained for the subset of training and test compounds defined for each molecular system is shown in Table 1.

The results reveal that there is large resemblance in the overall performance of the 3D-QSAR models obtained from MST/RM1 and MST/B3LYP Hyphar descriptors for all datasets. This finding is remarkable, since 3D-QSAR models derived from the RM1 hydrophobic descriptors compare well with the performance obtained at the B3LYP level, but at a much lower computational cost, making the usage of semiempirical methods highly attractive for the study of large libraries of drug-like compounds. Accordingly, the computationally less demanding RM1 method seems to be a promising choice for 3D-QSAR studies with Hyphar parameters.

On the basis of these results, the benchmark dataset reported by Sutherland and coworkers [95] was examined using the MST/RM1 Hyphar descriptors. The 3D-QSAR Hyphar models were compared with the CoMFA/CoMSIA results reported in [95], which were obtained by using electrostatic potential-fitted charges at the MNDO level, but for the THER set, where Gasteiger–Marsili charges were used. For the sake of comparison, an additional model, denoted CoMFA (RM1), which exploits RM1 electrostatic-potential fitted partial charges in conjunction with an steric field obtained from the Lennard–Jones potential with a positively charged C.3 atom probe, was also examined. This model, therefore, is intended to explore the efficiency of RM1-based partial charges in defining electrostatic features of molecules at the atomic level.

Table 2 shows the statistical parameters of the 3D-QSAR models. In general, similar performances were obtained for the different 3D-QSAR models determined for molecules in the training test included in a given system, as noted in the large resemblance between the statistical values of the regression (r^2) and cross-validation (q^2) models. The same trend can be observed for the test set compounds, although a small improvement was found for CoMFA

Table 1. Statistical parameters of the 3D-QSAR HyPhar models obtained from Miertus–Scrocco–Tomasi/B3LYP and Miertus–Scrocco–Tomasi/RM1 calculations for the four sets of compounds.[†]

System	Training set				Test set		Nc[†]	Field (%)	
	r^2	q^2	*s*	*Spress*	r^2	*s*		Elec	Nonelec
D2									
MST/B3LYP	0.94	0.77	0.31	0.60	0.78	0.57	3	68.6	31.4
MST/RM1	0.93	0.74	0.28	0.65	0.71	0.63	3	70.9	29.1
Chromanones									
MST/B3LYP	0.77	0.51	0.49	0.29	0.81	0.20	3	34.3	65.7
MST/RM1	0.76	0.42	0.51	0.32	0.66	0.82	3	42.1	57.9
GSK3									
MST/B3LYP	0.91	0.80	0.12	0.19	0.79	0.21	3	54.5	45.5
MST/RM1	0.91	0.82	0.30	0.18	0.79	0.21	5	64.7	35.3
Cruzain									
MST/B3LYP	0.81	0.50	0.31	0.51	0.69	0.47	2	53.0	47.0
MST/RM1	0.91	0.65	0.31	0.44	0.70	0.46	3	58.4	41.6

[†] See [91] for a proper description of the molecular sets. Nc denotes the number of PLS components in the best 3D-QSAR model, and the terms Elec and Nonelec stand for the fraction (in percentage) of electrostatic ($logP_{ele,i}$) and nonelectrostatic ($logP_{cav,i}$) hydrophobic contributions to the final model.
MST: Miertus–Scrocco–Tomasic.

(RM1) and Hyphar models in GPB and THERM systems compared with reference CoMFA/CoMSiA models. In addition, a higher level of accuracy was also achieved by the models derived from RM1 calculations since the number of outliers in the test set was lower than in classical CoMFA/CoMSIA (Supplementary Table 4). On the other hand, both BZR and COX2 were confirmed to be challenging systems for QSAR modeling, as already noted by Sutherland and coworkers [95]. For instance, in case of COX2, part of the reason for the poor predictive behavior may probably be ascribed to the fact that training and test set cover different ranges of in the property space.

The predictive performance of the models was also examined by analyzing their capacity to discriminate between active and inactive compounds. To this end, for each molecular system the compounds in the test set were ranked according to their experimental potency: 'active/positive' (P) and 'inactive/negative' (N) were categorized by applying a threshold value of 6.0 (in pIC_{50}/pK_i units). Then, test set compounds with a predicted pIC_{50}/pK_i value larger than the threshold value were considered 'actives/positives' (TP), whereas compounds with a predicted pIC_{50}/pK_i value lower than the threshold were considered 'inactives/negatives' (TN). For each molecular system, the number of P, N, TP and TN compounds, as well as false positives (FP) and false negatives (FN) are compiled in Supplementary Table 5. In turn, these values were used to identify correctly negative (specificity or TNR; in green in Figure 2) and positive (sensitivity or TPR; in blue in Figure 2) compounds, and to reduce the false negative rate ('fall-out' or FPR; in red in Figure 2) by applying Equations. 13-15.

$$\text{Specificity(TNR)} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \qquad \text{(Eq. 13)}$$

$$\text{Sensitivity(TPR)} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \qquad \text{(Eq. 14)}$$

$$\text{Fall} - \text{out(FPR)} = \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{(\text{FP} + \text{TN})} = 1 - \text{TNR} \qquad \text{(Eq. 15)}$$

These parameters, which can vary from 0 to 1, can be considered a measure of the predictive performance of the model. According to this classification, a model can be considered good if it has high specificity/sensitivity and low fall-out values. Nevertheless, this analysis requires a balanced partition of active and inactive compounds in the set of compounds, a requirement that is not fulfilled in the case of BZR and GPB systems, since only one inactive and one active compound are present in these two sets, respectively. Accordingly, the results obtained for

Table 2. Statistical parameters obtained for CoMFA and CoMSiA models reported with the results determined by using COMFA (RM1) and Hyphar models in this study for the eight molecular systems (ACE, AChE, BZR, COX2, DHFR, GPB, THERM and THR).[†]

System	Training set				Test set		Nc[‡]	Field (%)		
	r^2	q^2	*S*	*Spress*	r^2	*S*		Ele	N-Ele	HB
ACE [#]										
CoMFA	0.80	0.68	1.04	–	0.49/0.55	1.54/1.47	3	–	–	–
CoMSiA	0.76	0.65	1.15	–	0.52/0.58	1.48/1.41	3	–	–	–
CoMFA (RM1)	0.82	0.67	0.42	1.37	0.54/0.61	1.45/1.32	3	29.4	70.6	–
Hyphar	0.75	0.64	0.51	1.43	0.42/0.62	1.62/1.35	2	28.8	53.5	17.7
AChE										
CoMFA	0.88	0.52	0.41	–	0.47/0.56	0.95/0.87	5	–	–	–
CoMSiA	0.86	0.48	0.45	–	0.44/0.60	0.98/0.81	6	–	–	–
CoMFA (RM1)	0.90	0.54	0.32	0.85	0.35/0.52	1.07/0.86	6	20.0	80.0	–
Hyphar	0.76	0.45	0.50	0.92	0.65	0.78	4	64.1	18.7	17.2
BZR										
CoMFA	0.61	0.32	0.41	–	0.00/0.18	0.97/0.81	3	–	–	–
CoMSiA	0.62	0.41	0.41	–	0.08/0.30	0.93/0.75	3	–	–	–
CoMFA (RM1)	0.60	0.36	0.64	0.53	0.21/0.21	0.81/0.80	3	30.5	69.5	–
Hyphar	0.67	0.37	0.58	0.54	0.00/0.02	0.91/0.86	6	48.8	16.7	34.5
COX2										
CoMFA	0.70	0.49	0.56	–	0.29/0.37	1.24/1.09	5	–	–	–
CoMSIA	0.69	0.43	0.56	–	0.03/0.22	1.44/1.20	6	–	–	–
CoMFA (RM1)	0.74	0.51	0.52	0.72	0.19/0.34	1.20/1.07	5	28.6	71.4	–
Hyphar	0.60	0.52	0.63	0.71	0.26/0.40	1.15/0.99	3	85.4	4.3	10.3
DHFR										
CoMFA	0.79	0.65	0.59	–	0.59/0.70	0.89/0.73	5	–	–	–
CoMSiA	0.76	0.63	0.62	–	0.52/0.63	0.96/0.81	5	–	–	–
RM1 CoMFA	0.81	0.67	0.44	0.73	0.42/0.55	1.04/0.91	4	17.7	82.3	–
Hyphar	0.72	0.63	0.53	0.78	0.53/0.56	0.94/0.89	5	36.2	38.8	25.0
GPB										
CoMFA	0.84	0.42	0.43	–	0.42/0.37	0.94/0.70	4	–	–	–
CoMSiA	0.78	0.43	0.50	–	0.46/0.34	0.90/0.82	4	–	–	–
CoMFA (RM1)	0.88	0.43	0.36	0.85	0.51	0.89	4	24.4	75.6	–
Hyphar	0.83	0.54	0.42	0.75	0.71	0.68	3	52.0	2.7	45.3
THERM										
CoMFA	0.94	0.51	0.55	1.54	0.60	1.26	7	–	–	–
CoMSiA	0.85	0.54	0.73	–	0.36/0.46	1.87/1.60	6	–	–	–
CoMFA (RM1)	0.90	0.46	0.33	1.57	0.51/0.66	1.39/1.18	5	25.5	74.5	–
Hyphar	0.84	0.49	0.41	1.51	0.67	1.13	4	37.9	25.5	36.6
THR[¶]										
CoMFA	0.86	0.59	0.36	–	0.54/0.73	1.59/0.56	4	–	–	–
CoMSiA	0.88	0.62	0.34	–	0.55/0.62	0.76/0.66	5	–	–	–
CoMFA (RM1)	0.89	0.59	0.33	0.64	0.45/0.58	0.86/0.82	5	16.0	84.0	–
Hyphar	0.87	0.64	0.37	0.59	0.53/0.56	0.79/0.74	4	37.5	41.7	20.8

[†] For test sets compounds, statistical parameters (r^2 and *S*) with (left) and without (right) outliers (i.e., compounds with residuals higher than 2.5-fold the standard deviation) are indicated. The number of outliers for each system is reported in Supplementary Table 4.

[‡] See [91] for a proper description of the molecular sets. Nc denotes the number of PLS components in the best 3D-QSAR model, and the terms Elec and Nonelec stand for the fraction (in percentage) of electrostatic (*logPele,i*) and nonelectrostatic (*logPcav,i*) hydrophobic contributions to the final model.

[#] mol0088 (original file name mol_17) was excluded because it contains iodine atom.

[¶] mol0088 (original file name 82) was excluded due to problems with the input geometry.

ACE: 113 angiotensin converting enzyme; AChE: 111acetylcholinesterase; BZR: 147 ligands for benzodiazepine receptors; CoMFA: Comparative molecular field analysis; CoMSiA: Molecular similarity indices in a comparative analysis; COX-2: 282 cyclooxygenase-2; DHFR: 361 dihydrofolatereductase; GPB: 66 glycogen phosphorylase b; THER: 74 thermolysin ; THR: 87 thrombine.

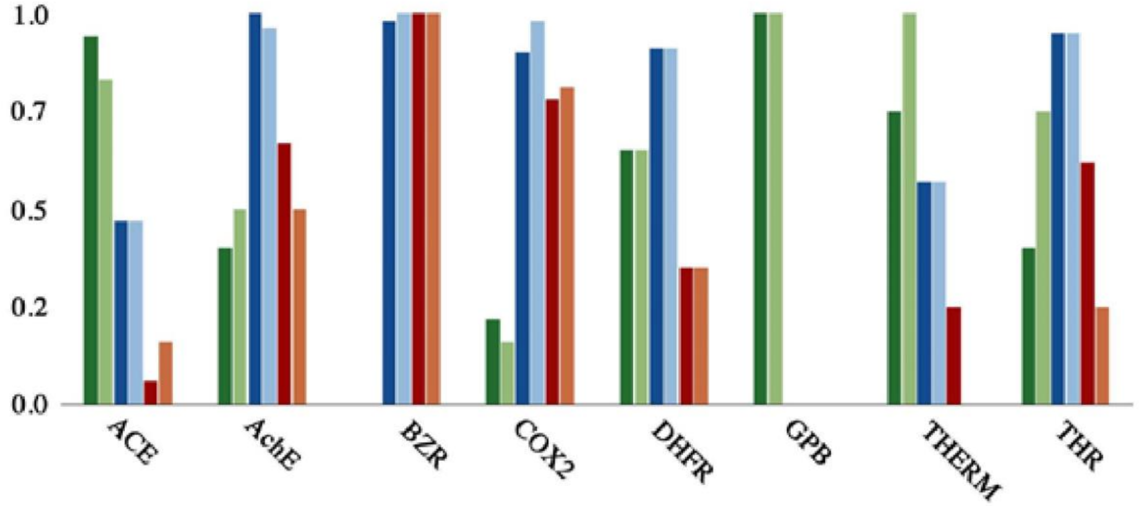


Figure 2. Specificity (in green), sensitivity (in blue) and fall-out (in red) for RM1 CoMFA (left) and H2 (right) models the test sets of the eight systems.
ACE: 113 angiotensin converting enzyme; AChE: 111acetylcholinesterase; BZR: 147 ligands for benzodiazepine receptors; COX2: 282 cyclooxygenase-2; DHFR: 361 dihydrofolatereductase; GPB: 66 glycogen phosphorylase b; THER: 74 thermolysin ; THR: 87 thrombine.

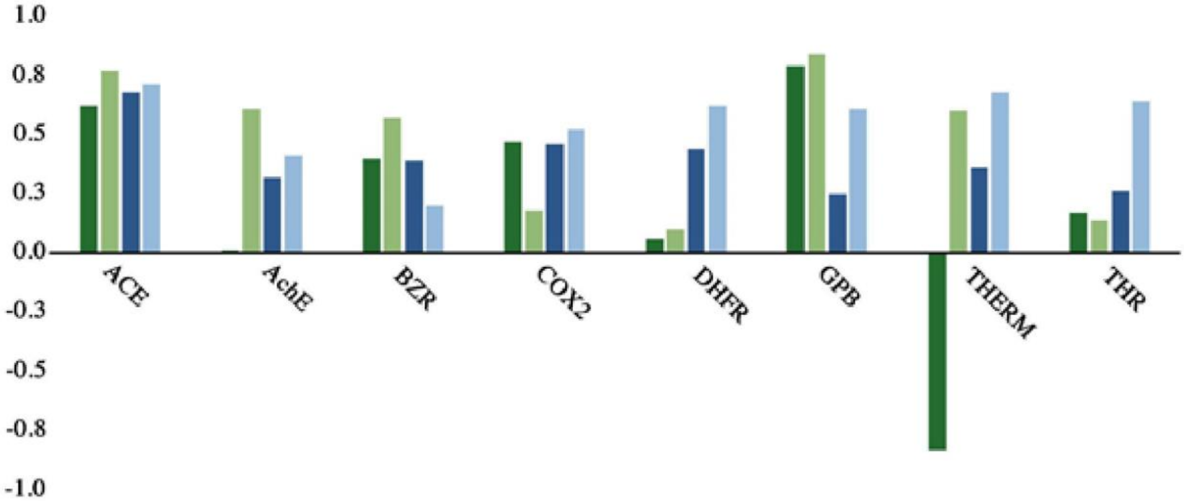


Figure 3. Spearman (Rs) coefficients for the first (Q1; in green), the second (Q2; in blue) and the third (Q3; in red) quartiles for RM1 CoMFA (left) and H2 (right) models.
ACE: 113 angiotensin converting enzyme; AChE: 111acetylcholinesterase; BZR: 147 ligands for benzodiazepine receptors; COX2: 282 cyclooxygenase-2; DHFR: 361 dihydrofolatereductase; GPB: 66 glycogen phosphorylase b; THER: 74 thermolysin; THR: 87 thrombine.

BZR and GPB should be excluded from the analysis. For the rest of molecular systems, both CoMFA (RM1) and Hyphar models exhibit generally similar trends (Figure 2). The Hyphar model has a slightly better performance in sensitivity/specificity and fall-out values for AchE, THERM and THR systems, whereas the opposite trend is found for CoMFA (RM1) in ACE and COX2.

Finally, the ability of CoMFA (RM1) and Hyphar models to rank the compounds according to their potency was also examined (Figure 3). To this end, the Spearman (*Rs*) coefficient for the first (Q1; in green), second (Q2; in blue) and third (Q3; in red) quartiles, which would encompass molecules with highest, medium and low activity/affinity, were determined for the test set compounds in each system. Although there is a notable resemblance in the general trends obtained for CoMFA (RM1) and Hyphar models, slightly better performances (higher *Rs* values) are observed for Hyphar models, especially for compounds of higher activity/affinity (Q1/Q2), whereas the differences are less pronounced for compounds in Q3, probably due to the larger noise associated with the biological activity low active compounds.

Overall, the results obained for the benchmark systems reveal that the Hyphar descriptors yield 3D-QSAR models with an overall performance that compares with the results obtained using standard CoMFA/CoMSiA. Hyphar models also seem to be more effective in locating (high sensibility) and ranking (high *Rs*) true positives, especially in regions of high and medium activity/affinity.

Final consideration & perspective

The concept of pharmacophore is essential to disclose the key features that dictate the interaction between ligand and receptor. Hence, it represents an important tool to identify guidelines valuable in computer-aided drug design, covering a variety of applications such as molecular similarity, VS, ligand optimization, scaffold hopping, as well as modeling of ADME(T) properties and target identification. The descriptive and predictive power of pharmacophores depends on the quality and adequacy of molecular properties used to disclose the hidden relationship between activity and chemical structure. In the last decades, several strategies were developed to derive descriptors capable of capturing the chemical features relevant for drug design, including the application of descriptors derived from QM methods coupled to continuum solvation models.

Although fundamental for the activity of drug-like compounds, inclusion of lipophilicity as a major descriptor has revealed more elusive, possibly due to the complexity of the chemical processes encompassed by this concept, or the difficulty to find a rigorous formalism to reduce it to atomic contributions since lipophilicity reflects a property of the whole molecule. In this context, it is worth stressing the efforts in deriving tools such as MLP [51] and HINT [55,56], where the molecular lipophilicity was treated by means of empirical atomic contributions, and hence enabling the analysis of the 3D-distribution of polar/apolar regions along the chemical scaffold to provide a novel interpretation to the molecular determinants responsible of biological activity.

QM-based continuum solvation methods are a promising strategy for deriving 3D-descriptors, such as COSMO-RS-based σ-profiles [78–81] or MST-derived 3D-lipophilicity patterns [82,83,92,97–99], which in turn may be exploited in computer-aided drug design. The set of studies reported up to now for a variety of benchmark datasets, covering both measurements of molecular similarity for aligned compound or the derivation of 3D-QSAR models, are encouraging. In general, the statistical performance of these QM-based descriptors compares well with the results obtained from classical approaches, generally combining electrostatic and steric fields, as illustrated in the comparative analysis reported here for the sets of compounds considered by Sutherland and coworkers [95]. At least in part, this may be due to the limitations of electrostatic/steric descriptors for describing enthalpy and entropy contributions to the binding affinity. On the other hand, QM-based approaches permit to account directly for the specific features of the bioactive species of the ligand, including effects attributable to ionization, tautomerism or the specific conformation, which may be advantageous compared with generic descriptors derived from empirical contributions. These computational approaches benefit from the usage of lipophilicity, a property widely used in drug design, easy to interpret by medicinal chemists and linked to a physicochemical property that can be measured experimentally. Through partitioning of the molecular lipophilicity into atomic contributions, novel fractional models that account for the 3D-lipophilicity pattern of compounds can then be exploited in computer-assisted drug design.

Overall, the analysis of structure–activity relationships in terms of the lipophilic/hydrophilic balance may provide a useful signature to complement studies performed with electrostatic/steric properties. In this sense, the QM MST-based hydrophobic descriptors are valuable in predicting molecular overlays and elucidating molecular similarity patterns. The higher descriptive quality of these descriptors could thus offer interesting clues in searching for novel bioactive compounds, especially for challenging targets.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/full/10.4155/fmc-2018-0435

Acknowledgments

The authors acknowledge J Muñoz-Muriedas (GSK group, Stevenage, UK) for valuable comments and suggestions.

Financial & competing interests disclosure

The authors thank the Ministerio de Economía y Competitividad (MINECO: SAF2017-88107-R, DI-14-06634, MDM-2017-0767) and the Generalitat de Catalunya (2017SGR1746, 2015-DI-052) for financial support. The Consorci de Serveis Universitaris de Catalunya (CSUC) is acknowledged for computational resources (Molecular Recognition project). J Vazquez is fellowship from the Generalitat de Catalunya. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

- All biological and biochemical processes are driven by the general concept of host–guest complementarity. Accordingly, an essential but effective description of the 'guest' is required for a successful prediction of 'host' recognition.
- The pharmacophore concept is a fundamental cornerstone in drug discovery, as it accounts for the common interaction features of a group of compounds toward their target structure, playing a critical role in determining the success of *in silico* techniques.
- Optimized descriptors able to model both pharmacokinetics and pharmacodynamics properties in drug design are not easily achievable, and the use of suboptimal physicochemical parameters may be a more effective strategy.
- Besides the relevance in predicting ADME(T) properties, lipophilicity exerts a pivotal role in accounting for the maximal achievable affinity that can be attained between ligand and receptor.
- The usage of lipophilicity descriptors may offer novel opportunities to disclose the underlying relationships between chemical features and biological activity. In this context, the availability of refined version of QM-based continuum solvation models may be an effective strategy for deriving novel descriptors well suited for drug design.
- In 3D-QSAR studies, the Miertus–Scrocco–Tomasi-derived Hyphar descriptors have been shown to provide models for structure–activity relationships with a predictive accuracy comparable to CoMFA/CoMSiA techniques based on electrostatic/steric parameters.
- The Hyphar descriptors are also a valuable alternative for molecule superposition and virtual screening of chemical libraries, especially for targets that may be challenging for predictive molecular similarity techniques.
- The availability of 'polar' and 'non-polar' fractional descriptors obtained from Miertus–Scrocco–Tomasi-based continuum solvation models may be valuable to explore the molecular determinants of bioactivity, providing complementary interpretations to classical descriptors in the rational design of novel compounds.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed. Engl.* 41, 2644–2676 (2002).
2. Khedkar SA, Malde AK, Coutinho EC, Srivastava S. Pharmacophore modeling in drug discovery and development: an overview. *Med. Chem.* 3, 187–197 (2007).
3. Güner OF, Bowen JP. Setting the record straight: the origin of the pharmacophore concept. *J. Chem. Inf. Model.* 54, 1269–1283 (2014).
4. Schueler FW. *Chemobiodynamics and Drug Design.* McGrawHill, NY, USA, (1960).
5. Beckett AH, Harper NJ, Clitherow JW. The impact of stereoisomerism in muscarinic activity. *J. Pharm. Pharmacol.* 15, 362–371 (1963).
6. Kier LB. Receptor mapping using molecular orbital theory. In: *Fundamental Concepts in Drug-Receptor Interactions.* Academic Press, NY, USA, 15–46 (1970).
7. Gund P, Wipke WT, Langridge R. Computer searching for molecular structure file for pharmacophoric patterns. In: *Computers in Chemical Research and Education (Volume 3).* Hadzi D, Zupan J (Eds). Elsevier Scientific, Amsterdam, The Netherlands, 5–33 (1973).
8. Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA. Glossary of terms used in medicinal chemistry (IUPAC recommendations 1998). *Pure Appl. Chem.* 70, 1129–1143 (1998).
9. Bender A, Glen RC. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* 2, 3204–3218 (2004).
10. Wolber G, Seidel T, Bendix F, Langer T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today* 13, 23–29 (2008).

11. Kaserer T, Beck KR, Akram M, Odermatt A, Schuster D. Pharmacophore models and pharmacophore-based virtual screening: concepts and applications exemplified on hydroxysteroid dehydrogensases. *Molecules* 20, 22799–22832 (2015).

12. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* 57, 3186–3204 (2013).

13. Verma J, Khedkar VM, Coutinho EC. 3D-QSAR in drug design – a review. *Curr. Top. Med. Chem.* 10, 95–115 (2010).

14. Cramer RD III, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* 110, 5959–5967 (1988).

15. Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* 37, 4130–4146 (1994).

16. Nilsson J, Wikström H, Smilde A *et al.* GRI D/GOLPE 3D quantitative structure–activity relationship study on a set of benzamides and naphthamides, with affinity for the dopamine D3 receptor subtype. *J. Med. Chem.* 40, 833–40 (1997).

17. Winiwarter S, Ridderström M, Ungell A-L, Andersson T, Zamora I. Use of molecular descriptors for absorption, distribution, metabolism, and excretion predictions. In: *Comprehensive Medicinal Chemistry II.* (Volume 5). Testa B, van de Waterbeemd H (Eds). Elsevier, Amsterdam, The Netherlands, 531–554 (2006).

18. Gleeson MP, Hersey A, Montanari D, Overington J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discov.* 10, 197 (2011).

19. Testa B, Carrupt PA, Gaillard P, Tsai RS. Intramolecular interactions encoded in lipophilicity: Their nature and significance. In: *Lipophilicity in Drug Action and Toxicology.* Pliska V, Testa B, van de Waterbeemd H (Eds). VCH, Weinheim, Germany, 49–71 (1996).

20. *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability.* van de Waterbeemd H, Lennernäs H, Artursson P (Eds). Wiley-VCH, Weinheim, Germany, (2003).

21. Caron G, Ermondi G, Scherrer RA. Lipophilicity, polarity, and hydrophobicity. In: *Comprehensive Medicinal Chemistry II.* Taylor JB, Triggle DJ (Eds). Elsevier Science, Oxford, UK, 5, 425–452 (2007).

22. Van de Waterbeemd H, Carter RE, Grassy G *et al.* Glossary of terms used in computational drug design (IUPAC Recommendations 1997). *Pure Appl. Chem.* 69, 1137–1152 (1997).

23. Spyrakis F, Ahmed MH, Bayden AS, Cozzini P, Mozzarelli A, Kellog GE. The roles of water in the protein matrix: a largely untapped resource for drug discovery. *J. Med. Chem.* 60, 6781–6827 (2017).

- • **This contribution provides an updated perspective on the roles of water molecules in protein structure, function and dynamics, with a particular focus on the applications in drug discovery and design.**

24. Cheng AC, Coleman RG, Smyth KT *et al.* Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* 25, 71–75 (2007).

- •• **Reports a model-based approach to predict druggable binding sites and estimate the maximal affinity acievable by a small compound that relies on the hydrophobic desolvation, and the nonpolar surface and curvatuve of the target binding site.**

25. Davis AM, Teague SJ. Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angew. Chem. Int. Ed. Engl.* 38, 736–749 (1999).

26. Hajduk PJ, Huth JR, Fesik SW. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* 48, 2518–2525 (2005).

27. Egner U, Hillig RC. A structural biology view of target drugability. *Expert Opin. Drug Discov.* 3, 391–401 (2008).

28. Schmidtke P, Barril X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* 53, 5858–5867 (2010).

29. Schmidtke P, Luque FJ, Murray JB, Barril X. Shielded hydrogen bonds as structural determinants of binding kinetics: application in drug design. *J. Am. Chem. Soc.* 133, 18903–18910 (2011).

30. Tsopelas F, Giaginis C, Tsantili-Kakoulidou A. Lipophilicity and biomimetic properties to support drug discovery. *Expert Opin. Drug Discov.* 12, 885–896 (2017).

31. Freeman-Cook KD, Hoffman RL, Johnson TW. Lipophilic efficiency: the most important efficiency metric in medicinal chemistry. *Future Med. Chem.* 5, 113–115 (2013).

32. Jopkins AL, Keserü GM, Leeson PD, Ress DC, Reynolds CH. The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discov.* 13, 105–121 (2014).

33. Johnson TW, Gallego RA, Edwards MP. Lipophilic efficiency as an important metric in drug design. *J. Med. Chem.* 61, 6401–6420 (2018).

- • **An updated overview of the role of lipophilic efficiency as a metric with increasing impact in guiding drug discovery.**

34. Chen Z, Weber SG. A high-throughput method for lipophilicity measurement. *Anal. Chem.* 79, 1043–1049 (2007).

35. Giaginis C, Tsantili-Kakoulidou A. Alternative measures of lipophilicity: from octanol-water partitioning to IAM retention. *J. Pharm. Sci.* 97, 2984–3004 (2008).

36. Andrés A, Rosés M, Ràfols C, Bosch E, Espinosa S, Segarra V, Huerta JM. Setup and validation of shake-flask procedures for the determination of partition coefficients (logD) from low drug amounts. *Eur. J. Phar. Sci.* 76, 181–191 (2015).

37. Mannhold R, Dross K. Calculation procedures for molecular lipophilicity: a comparative study. *Quant. Struct. Act. Relat.* 15, 403–409 (1996).

38. Ghose AK, Viswanadhan VN, Wendoloski JI. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* 19, 172–178 (1998).

39. Mannhold R, van de, Waterbeemd H. Substructure and whole molecule approaches for calculating logP. *J. Comput. Aided Mol. Des.* 15, 337–354 (2001).

40. Wang J-B, Cao D-S, Zhu M-F, Yun Y-H, Xiao N, Liang Y-Z. *In silico* evaluation of logD$_{7.4}$ abd comparison with other prediction methods. *J. Chemometrics*. 29, 389–398 (2015).

41. Chen HF. *In silico* logP prediction for a large data set with support vector machines, radial basis neutral networks and multiple linear regression. *Chem. Biol. Drug Des* 74, 142–147 (2009).

42. Mannhold R, Poda GI, Ostermann C, Tetko IV. Calculation of molecular lipophilicity: state-of-the-art and comparison of logP methods on more than 96,000 compounds. *J. Pharm. Sci.* 98, 861–893 (2009).

43. Leo A, Hansch C, Elkins D. Partition coefficients and their uses. *Chem. Rev.* 71, 525–616 (1971).

44. Nys GG, Rekker RF. The concept of hydrophobic fragmental constants (f values). II. Extension of its applicability to the calculation of lipophilicities of aromatic and heteroaromatic structures. *Eur. J. Med. Chem.* 9, 361–375 (1974).

45. Mannhold R, Rekker RF. The hydrophobic fragmental constant approach for calculating logP in octanol/water and aliphatic hydrocarbon/water systems. *Perspect. Drug Discovery Des.* 18, 1–18 (2000).

46. Ghose AK, Crippen GM. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure–activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* 27, 21–35 (1987).

47. Viswanadhan VN, Ghose AK, Revankar GR, Robins RK. An estimation of the atomic contribution to octanol-water partition coefficient and molar refractivity from fundamental atomic and structural properties: its uses in computer-aided drug design. *Math. Comput. Model.* 14, 505–510 (1990).

48. Wildman SA, Crippen GM. Prediction of physicochemical properties by atomic contributions. *J. Chem. Inf. Comput. Sci.* 39, 868–873 (1999)

49. Wang R, Fu Y, Lai L. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Model.* 37, 615–621 (1997).

50. Ottaviani G, Martel S, Carrut P-A. *In silico* and *in vitro* filters for the fast estimation of skin permeation and distribution of new chemical entities. *J. Med. Chem.* 50, 742–748 (2007).

51. Gaillard P, Carrupt PA, Testa B, Boudon A. Molecular lipophilicity potential, a Tool in 3D QSAR: method and applications. *J. Comput. Aided Mol. Des.* 8, 83–96 (1994).

52. Laguerre M, Saux M, Dubost J, Carpy A. MLPP: a program for the calculation of molecular lipophilicity potential in proteins. *Pharm. Pharmacol. Commun.* 3, 217–222 (1997).

53. Efremov RG, Chugunov AO, Pyrkov TV, Priestle JP, Arseniev AS, Jacoby E. Molecular lipophilicity in protein modeling and drug design. *Curr. Med. Chem.* 14, 393–415 (2016).

54. Bitam S, Hamadache M, Hanini S. QSAR model for prediction of the therapeutic potency of N-benzylpiperidine derivatives as AChE inhibitors. *SAR QSAR Environ. Res.* 28, 471–489 (2017).

55. Kellogg GE, Semus SF, Abraham DJ. HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput. Aided Mol. Des.* 5, 454–552 (1991).

56. Kellogg GE, Abraham DJ. Hydrophobicity: is Log P(o/w) more than the sum of its parts? *Eur. J. Med. Chem.* 35, 651–661 (2000).

57. Fornabaio M, Spyrakis F, Mozzarelli A, Cozzini P, Abraham DJ, Kellogg GE. Simple, intuitive calculations of free energy of binding for protein–ligand complexes. 3. The free energy contribution of structural water molecules in HIV-1 protease complexes. *J. Med. Chem.* 47, 4507–4516 (2004).

58. Amadasi A, Spyrakis F, Cozzini P, Abraham DJ, Kellogg GE, Mozzarelli A. Mapping the energetics of water-protein and water-ligand interactions with the ‘natural’ HINT forcefield: predictive tools for characterizing the roles of water in biomolecules. *J. Mol. Biol.* 358, 289–309 (2006).

59. Marabotti A, Spyrakis F, Facchiano A *et al.* Energy-based prediction of amino acid-nucleotide base recognition. *J. Comput. Chem.* 29, 1955–1969 (2008).

60. Amadasi A, Surface JA, Spyrakis F, Cozzini P, Mozzarelli A, Kellogg GE. Robust classification of ‘relevant’ water molecules in putative protein binding sites. *J. Med. Chem.* 51, 1063–1067 (2008).

61. Ahmed MH, Spyrakis F, Cozzini P *et al.* Bound water at protein-protein interfaces: partners, roles and hydrophobic bubbles as a conserved motif. *PLoS ONE* 6, e24712 (2011).

62. Rogers KS, Cammarata A. A molecular orbital description of the partitioning of aromatic compounds between polar and non-polar phases. *Biochim. Biophys. Acta* 193, 22–29 (1969).

63. Rogers KS, Cammarata A. Superdelocalizability and charge density. A correlation with partition coefficients. *J. Med. Chem.* 12(4), 692–693 (1969).

90. Luque FJ, Curutchet C, Muñoz-Muriedas J *et al.* Continuum solvation models: dissecting the free energy of solvation. *Phys. Chem. Chem. Phys.* 5, 3827–3836 (2003).

91. Luque FJ, Bofill JM, Orozco M. New strategies to incorporate the solvent polarization in self-consistent reaction field and free-energy perturbation simulations. *J. Chem. Phys.* 103, 10183–10191 (1995).

92. Vázquez J, Deplano A, Herrero A. *et al*. Development and validation of molecular overlays derived from 3D hydrophobic similarity with PharmScreen. *J. Chem. Inf. Model.* 58, 1596–1609 (2018).

•• **A comparative analysis of electrostatic/steric and QM-based lipophilicity (Hyphar) descriptors for predicting molecular overlays from 3D similarity measurements.**

93. PharmScreen – PharmQSAR, Pharmacelera, Barcelona, Spain. www.pharmacelera.com

94. Giangreco I, Cosgrove DA, Packer MJ. An extensive and diverse set of molecular overlays for the validation of pharmacophore programs. *J. Chem. Inf. Model.* 53, 852–866 (2013).

95. Sutherland JJ, O'Brien LA, Weaver DF. A comparison of methods for modeling quantitative structure–activity relationships. *J. Med. Chem.* 47, 5541–5554 (2004).

96. Forti F, Barril X, Luque FJ, Orozco M. Extension of the MST continuum solvation model to the RM1 semiempirical Hamiltonian. *J. Comput. Chem.* 29, 578–587 (2008).

97. Muñoz J, Barril X, Hernandez B, Orozco M, Luque FJ. Hydrophobic similarity between molecules: A MST-based hydrophobic similarity index. *J. Comput. Chem.* 23, 554–563 (2002).

98. Muñoz-Muriedas J, Perspicace S, Bech N, Guccione S, Orozco M, Luque FJ. Hydrophobic molecular similarity from MST fractional contributions to the octanol/water partition coefficient. *J. Comput. Aided Mol. Des.* 19, 401–419 (2005).

99. Muñoz-Muriedas J, Barril X, Lopez JM, Orozco M, Luque FJ. A hydrophobic similarity analysis of solvation effects on nucleic acid bases. *J. Mol. Model.* 13, 357–365 (2007).

Supplementary Material

Lipophilicity in drug design: Novel lipophilicity-based descriptors for 3D-QSAR studies

Computational details of 3D-QSAR (CoMFA, CoMSIA, Hyphar) models.

Table S1. Details about pre-compiled sets in Ref. S1.

Table S2. Details of grid and field projections for sets reported in Table 1.

Table S3. Grid dimensions for CoMFA (RM1) and Hyphar models calculated in this study.

Table S4. Number of outliers in the test set.

Table S5. Number of positives (P), negatives (N), true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) for each molecular system. A threshold of 6.0 to the pIC_{50}/pK_i value was applied to classify compounds of each test set.

References

Computational details of 3D-QSAR (CoMFA, CoMSIA, Hyphar) models.

CoMFA/CoMSiA models determined by Sutherland and coworkers [S1] relied on molecular geometries obtained by energy minimization with the MMFF94S force field in Sybyl [S2]. Electrostatic potential-fitted (ESP) charges were determined using the semiempirical MNDO method [S3], but for compounds in the THER set, where Gasteiger-Marsili partial charges [S4] were adopted. Aligned compounds were enclosed in a 2 Å-spaced grid with boundaries set to 4 Å around the molecules. The "minimum σ" value for removing descriptors with low variance was set to 2.0 for CoMFA and 1.0 for CoMSiA.

The Hyphar model was derived using PharmQSAR® and the same molecular alignment defined in ref. S1. Atomic "polar" ($logP_{ele,i}$) and "non-polar" ($logP_{cav,i}$) hydrophobic contributions were obtained from MST continuum solvation calculations coupled to the semiempirical Hamiltonian RM1 [S5]. In addition, a third field that accounts for the HB donor/acceptor character of polar atoms was included as described in Ref. S6. The aligned molecules were enclosed in a 1 Å-spaced grid with boundaries set to 4 Å around the molecules. The similarity index function implemented in CoMSiA [S7] was applied to project the atomic contributions. PLS statistical analysis based on the NIPALS algorithm was performed to evaluate models robustness. In this regard, each projected field was stored in a *M*x*Ng* matrix (*M* is the number of molecules and *Ng* is the number of grid points). Field values were then centered, scaled to unit variance and columns with a standard deviation lower than a certain threshold (typically 0.1-0.01) were excluded. The best Hyphar model, in term of number of PLS components, was identified in accordance to the lowest standard deviation error in prediction (*Spress*) corrected by the degree of freedom of the model and the predictive ability of the model for the test set (r^2). The effect of outliers on the predictive performance for the test set was also examined.

Finally, for the sake of comparison, PharmQSAR® was also used to derived an additional CoMFA model, named CoMFA (RM1), which was obtained by combining the electrostatic field determined from the RM1 ESP partial charges in conjunction with an steric field obtained from the Lennard-Jones potential interaction energies with a positively charged C.3 atom probe.

Table S1. Details about pre-compiled sets in Ref. S1.

System [*]	**Training Set**	**Test Set (Internal validation)**	**Activity Range (Training set)** [c]	**Activity Range (Test set)** [c]
Angiotensin converting enzyme (ACE)	76	37 [a]	2.14 – 9.88	2.70 – 9.94
Acetylcholinesterase (AChE)	74	37	4.28 – 9.52	4.27 – 9.22
Benzodiazepine (BZR)	98	49	6.34 – 8.92	5.52 – 8.85
Cyclooxygenase-2 (COX-2)	188	94	4.03 – 9.00	4.03 – 8.70
Dihydrofolate reductase (DHFR)	237	124	3.30 – 9.81	3.57 – 9.40
Glycogen phosphorylase b (GPB)	44	22	1.30 – 5.50	1.40 – 6.80
Thermolysin (THERM) [**]	59	15	0.52 – 10.17	2.51 – 7.73
Thrombine (THR)	59	28 [b]	4.57 – 8.48	4.36 – 8.38

[*] Compounds defined as "inactives" in Ref. S1 were excluded from this study.

[**] Partition in training-test sets made according to Ref. S7.

[a] **mol0088** with original file name "**mol_17**", was excluded because it contains iodine atom.

[b] **mol0088** with original file name "**82"** was excluded from the calculations due to problems with the input geometry.

[c] Experimental data for ACE, AChE, BZR, COX-2, and DHFR sets are in pIC$_{50}$ units. Experimental data for GPB, THERM and THR sets are in pK$_i$ units.

Table S2. Details of grid and field projections for sets reported in Table 1.

Set	**Grid step size**	**Box Extension**	**Radius probe** [*]	**Grid points**
ACE	1.0	4.0	1.52	18009
AChE	1.0	4.0	1.52	17472
BZR	1.0	4.0	1.52	11440
COX2	1.0	4.0	1.52	10488
DHFR	1.0	4.0	1.52	12144
GPB	1.0	4.0	1.52	10488
THERM	1.0	4.0	1.52	18720
THR	1.0	4.0	1.52	14300

[*] All grids use C.3 probe atom with charge +1, a grid step size of 1.0 Å, a box extension of 4.0 Å, and a probe radius of 1.52 Å.

Table S3. Grid dimensions for CoMFA (RM1) and Hyphar models calculated in this study.

	ACE	**AChE**	**BZR**	**COX2**	**DHFR**	**GPB**	**THERM**	**THR**
x min	-10.46	-10.62	-11.42	14.03	4.72	-10.68	-9.30	-5.66
x max	15.54	12.38	13.58	37.03	26.72	12.32	15.70	19.34
y min	-14.92	51.29	-8.97	14.56	-4.29	-8.07	-17.76	-16.93
y max	13.08	76.29	12.03	32.56	16.71	9.93	11.24	4.07
z min	-10.21	54.07	-8.14	4.65	-1.03	-7.24	-12.30	-0.79
z max	11.79	81.07	10.86	26.65	21.98	14.76	10.70	23.21

Table S4. Number of outliers in the test set.

ACE	AChE	BZR	COX2	DHFR	GPB	THERM	THR
CoMFA/CoMSIA [*]							
1	1	3	5	6	1	1	1
CoMFA (RM1)							
3	2	1	5	5	0	3	1
Hyphar							
5	0	1	5	1	0	0	1

[*] Number of outliers for CoMFA/CoMSIA models taken from Ref. S1.

Table S5. Number of positives (P), negatives (N), true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) for each molecular system. A threshold of 6.0 to the pIC_{50}/pK_i value was applied to classify compounds of each test set.

	P	**N**	**TP**	**TN**	**FP**	**FN**
ACE						
CoMFA(RM1)	19	18	9	17	1	10
Hyphar	19	18	9	15	3	10
AChE						
CoMFA (RM1)	28	9	28	4	6	0
Hyphar	28	9	28	5	5	0
BZR						
CoMFA (RM1)	45	1	44	0	1	1
Hyphar	45	1	45	0	1	0
COX2						
CoMFA (RM1)	61	32	55	7	25	6
Hyphar	61	32	60	5	27	1
DHFR						
CoMFA (RM1)	59	65	54	42	23	5
Hyphar	59	65	54	42	23	5
GPB						
CoMFA (RM1)	1	21	0	21	0	1
Hyphar	1	21	0	21	0	1
THERM						
CoMFA (RM1)	7	8	4	6	2	3
Hyphar	7	8	6	8	0	1

References

S1 Sutherland JJ, O'Brien LA, Weaver DF. A comparison of methods for modeling quantitative structure−activity relationships. *J. Med. Chem.* 47, 5541-5554 (2004).

S2 Sybyl 8.1, Tripos Inc., St. Louis, MO (2008).

S3 Besler BH, Merz KM, Kollman PA Atomic charges derived from semi-empirical methods. *J. Comput. Chem.* 11, 431–439 (1990).

S4 Gasteiger J, Marsili M. Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. *Tetrahedron* 36, 3219–3228 (1980).

S5 Ginex T, Muñoz-Muriedas J, Herrero E, Gibert E, Cozzini P, Luque FJ. Development and validation of hydrophobic molecular fields derived from the quantum mechanical IEF/PCM-MST solvation models in 3D-QSAR. *J. Comput. Chem.* 37, 1147-1162 (2016).

S6 Vázquez J, Deplano A, Herreo A, Ginex T, Gibert E, Rabal O, Oyarzabal J, Herrero E, Luque FJ. Development and validation of molecular overlays derived from three-dimensional hydrophobic similarity with PharmScreen. *J. Chem. Inf. Model.* 58, 1596-1609 (2018).

S7 Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* 37, 4130-4146 (1994).

3.3 **PAPER 3:** *"Similarity assessment of Lipophilic Distribution: A Boost for Structure-Based Methods"*

Manuscript under preparation

Javier Vazquez, †, ‡ Enric Herrero, ‡ and Francisco J Luque, †

† Department of Nutrition, Food Sciences & Gastronomy, Faculty of Pharmacy & Food Sciences, Campus Torribera, Institute of Biomedicine (IBUB), & Institute of Theoretical & Computational Chemistry (IQTC-UB), University of Barcelona, Av. Prat de la Riba 171, Santa Coloma de Gramenet E-08921, Spain

‡ Pharmacelera, Plaça Pau Vila, 1, Sector 1, Edificio Palau de Mar, Barcelona 08039, Spain

Similarity assessment of lipophilic distribution: A boost for structure-based methods

Javier Vazquez, [†,‡] Enric Herrero,[†] and F. Javier Luque[‡]

[†] *Pharmacelera, Plaça Pau Vila, 1, Sector 1, Edificio Palau de Mar, Barcelona 08039, Spain*
[‡] *Department of Nutrition, Food Science and Gastronomy, Faculty of Pharmacy and Food Sciences, Institute of Biomedicine (IBUB), and Institute of Theoretical and Computational Chemistry (IQTC-UB), University of Barcelona, Av. Prat de la Riba 171, Santa Coloma de Gramenet E-08921, Spain*

Abstract

In structure-based (SB) virtual screening (VS), a scoring function is usually applied to rank a database of screened compounds. Docking programs are generally successful in reproducing the experimental binding modes, but the scoring functions still present serious limitations to provide an accurate estimate of the binding affinity. The combination of SB and ligand-based (LB) 3D similarity may be a promising strategy to increase hit rates in VS. Here, we propose a combined method to solve the limitations of both VS approximations that balances both the docking score with the similarity between compounds and a reference ligand. In this work, the similarity is determined through an atom-based description of the 3D distribution lipophilicity map determined calculations performed with the MST continuum solvation model. Different strategies have been explored to combine the information provided by docking and similarity measurements to obtain an improved ranking score. For a benchmarking of 44 data sets, including 41 targets, the proposed methods increase the identification of actives compounds in the early stage (ROCe%) and total (AUC) performance of VS compared to pure LB and SB methods in isolation.

Key words

Virtual screening, Compound ranking, Molecular docking, Binding mode, 3D similarity, Protein–ligand interactions

INTRODUCTION

Structure-based (SB) and ligand-based (LB) approaches have been widely used in virtual screening (VS) processes in computer-aided drug design.[1,2] SB techniques encompass methods that exploit the structural information of the macromolecular target, enabling the study of the binding mode of drug-like compounds. Thus, they rely on the availability of precise three-dimensional information of the structural arrangement of atoms in the target protein, particularly regarding the geometrical and physicochemical properties of the residues that shape the ligand binding cavity. This information can be determined experimentally (X-ray crystallography, nuclear magnetic resonance, or cryo-electron microscopy), or through computational methods (homology modeling or molecular dynamics)[3,4]. The most accepted and extensively applied SB technique is molecular docking, which predicts the preferred orientation of a drug-like compound, often supplemented with pharmacophoric constraints, and the search for hits in VS of fragment and compound libraries. On the other hand, LB refers to a diverse group of strategies, which primarily disclose similarity relationships between molecular descriptors without the need for structural information of the target. The similarity principle property (SPP)[5] relies on the concept that similar compounds should have similar properties. Under this framework, a wide variety of methods have been developed with the aim to find structure-activity relationships, derive pharmacophores that may rationalize the activity of compounds, and the application of similarity measurements to the search of novel chemical scaffolds .[6–8]

Many efforts have been dedicated to improving the accuracy and predictive power of both LB- and SB methods, which are limited by several challenges. On one side, besides the lack of precise structural information of the target, LB methods are limited by the quality of the descriptors used to characterize the chemical features of compounds, the consistency and chemical diversity of the training set, and the mathematical formalism that underlies the measurements of similarity between molecules. On the other hand, SB methods may be affected by the limited accuracy of the 3D geometrical data, the involvement of different conformational states, often induced by specific ligands, of the target protein, or the assistance of structural waters in mediating ligand binding. Even in the case of well-defined structural models of the target protein, the predictive power of SB techniques may be affected by the use of oversimplified scoring functions, which provide a rough approximation to the balance between enthalpic and entropic contributions to the ligand-target interaction, and the exhaustiveness of the sampling search, which may lead to a substantial computational cost for VS applications.[9]

In this context, the combination of LB and SB methods may be a valuable synergistic strategy to exploit the structural and chemical information available for the target biological system, and to minimize the bias due to the intrinsic deficiencies of both methods.[10–13] In fact, the combination of LB and SB methods has been reviewed by Drwal and Griffith,[2] who classified the combined approaches into three categories: sequential, parallel, and hybrid approaches. The sequential approach splits the screening process into various steps to overcome the expensive computational cost of the SB approach. Accordingly, a prefiltering step is performed at the beginning of the VS using less expensive LB techniques, and the retrieved hits are subsequently evaluated using molecular docking.[14–16] In the parallel approach, LB and SB methods are run independently, and then the results are merged to obtain a mixed ranking.[10,17–19] Finally, hybrid approaches integrate 3D ligand information and SB in one independent system. The most common approach consists of translating protein-ligand interactions from SB tools into pharmacophore features to be used in LB algorithms. These approach has been demonstrated successfully in VS[20] and for profiling purposes[21]. Alternatively, in other hybrid protocols the similarity between the docked compounds (SB) and a known crystallographic ligand (LB) is computed to re-score the docking outcome.[12,22] This protocol directly addresses the docking scores limitations, but only a few validation studies have been reported.[23–25]

Most of these methods were born to improve the discrimination capacity between active and inactive molecules. To attain this objective, an accurate scoring and ranking function must be defined. However, the scored output produced by docking tools is not always the best way to select compounds and rank them.[26,27] This work aims to explore parallel and hybrid approaches throughout the combination of molecular docking and 3D similarity based on the comparison of fractional descriptions of the 3D lipophilicity distribution pattern of molecules derived from quantum mechanical (QM) continuum solvation models.[28–31] In particular, we evaluate the suitability of a hybrid method that takes advantage of data fusion[18] techniques to re-rank the docked poses with 3D similarity.

METHODS

Test dataset. The performance of VS methods is highly sensitive to the set of compounds. Therefore, a big and diverse number of receptors should be taken into account in order to include cases were different tools work better. For our purposes here, the Directory of Useful Decoys (DUD; http://dud.docking.org/)[32] has been used. Although DUD is suitable to address the weaknesses of docking methods, LB methods can easily account for the differences between actives and inactives[33]. For this reason, the subset of DUD proposed by Good and Opera called DUD_LIB_VS_1.0[34–36] has been chosen, as this specific dataset was conceived with the aim of avoiding an overestimation of the performance of LB methods. A lead-like filter and clustering algorithm was applied to eliminate large molecules with inappropriate physicochemical properties and to reduce the artificial bias between structural analogs and actives during the enrichment test.[37,38] Additionally, four sets taken from DEKOIS V2.0 (http://www.dekois.com)[39], which was specially compiled to evaluate combined LB and SB methods,[23] were also considered.

The DUD_LIB_VS_1.0 set contains known actives and mimetic[40] decoys for 40 target proteins downloaded from the DUD website. The second set is made up of the DHFR, GR, HIV1PR, and VEGFR2 benchmarking sets directly extracted from DEKOIS V2.0, a subset previously used to test combinatorial approaches.[41] In this benchmark, each different set has the same size and the same number of active ligands, selected from BindingDB.[42] For ease of reading, DUD_LIB_VS_1.0 will be referred to as BS1 (benchmarking set 1), and the subset of DEKOIS 2.0 as BS2 (benchmarking set 2). Since there is an overlap between BS1 and BS2 targets, a suffix (BS1/BS2) is added to each target name. A detailed description of the original datasets is provided in tables S1 of the Supporting Information.

Ligand Preparation. In this study, two complementary aspects of ligand-receptor interactions were analyzed: the (de)solvation contribution using the MST-derived lipophilicity descriptors implemented in PharmScreen,[43] and the ligand fit into the binding site, which was examined using Glide.[44–46]

To obtain an initial 3D conformation, the geometry of all ligands was minimized using the semiempirical Hamiltonian RM1[47,48] using a locally modified version of MOPAC.[49] The hydrophobic descriptors used in the LB method were obtained by using the RM1-parametrized version of the MST solvation model.[48] The parameterization of MST/RM1 provides accurate estimates of the solvation free energy for neutral molecules. However, the treatment of ionic compounds is more delicate due to the high dependence between the scaling factor used to modulate de electrostatic boundary between solute and solvent and the nature of the ionizable group.[48] Therefore, all compounds were modeled considering a neutral state. MolVS[50], a standardization tool written in Python using the RDKit[51] chemistry framework, was applied to neutralize the protonation states. Tautomerism was not modified. To explore the conformation space, 100 conformations for each database ligand were calculated using RDKit[51]. For the SB approach, tautomerism and the ionization state from the original sets were not modified.

Protein Preparation. A target was picked for each ligand set to perform the SB analysis. For BS1, all 40 targets were obtained from the DUD Web site (DUD release 2). The original waters and co-factors retained for each protein target were maintained. For BS2, the targets were obtained from the Protein Data Bank (PDB)[52]. The protocol to preserve waters and co-factors in this set was the one reported by Anighoro and Bajorath[23]. All the structures were prepared using the "Protein Preparation Wizard" module in Maestro. A detailed description of the targets is provided in Supporting Information Table S1.

Query Preparation. The query structures chosen to perform the similarity search in the LB analysis are the same as reported in previous works.[23,35,36,53] To achieve LB similarity in BS1, the queries selected were the same as proposed by Huang et al.[36] and used later in the validation of LB tools[54,55] The structures were downloaded from DUD Web site (DUD release 2). For BS2, the same co-crystallized ligands used by Anighoro and Bajorath[23] were extracted from the PDB (Table S1).

Ligand-Based VS. PharmScreen was used as the LB virtual screening tool with all settings left at the default configuration. This methodology exploits the partitioning of lipophilicity into atomic contributions within the framework of continuum solvation models in conjunction with the hydrogen bond distribution.[43] The calibration of the field weighting was achieved using a training set that consists of 14 molecular systems[56–58]. The largest accuracy was reported for weighting factors of 15 (electrostatic contribution to logP), 55 (non-electrostatic contribution to LogP), and 30 (Hydrogen bond).[43]

Structure-Based VS. Both HTVS and SP modes were used in Glide as the SB virtual screening tool. The receptor grids were centered on the molecule selected as the query in the LB method. Grid dimensions were defined as default except for BS2 where the sizes conform to the indications set in the reference paper[23]. The general van der Waals radius-scaling factor was reduced (default: 1.0, modified to 0.9) to decrease the number of rejected molecules. The remaining settings in the grid generation were left at the default values. With the same objective, the cutoff of Coulomb-van der Waals energy and H-bond score was virtually disallowed for the docking job (default: 0.0, modified to 1000). Even with these changes, some molecules were rejected in docking calculations with Glide and they were excluded in the analysis of the results obtained for BS1 and BS2. A list of discarded molecules is reported in Supporting Information Table S1.

Protocols applied to Combine LB and SB methods. Three different protocols were tested for the combination of LB and SB methods, namely, parallel ranking, rescoring ranking, and consensus ranking.

The first protocol pertains to the parallel combination category,[59] where the output rankings obtained from separate LB and SB screenings are merged to create the final ranking. With the aim to treat both methods with equal parity, the first molecule of SB ranking and the first molecule of LB ranking will occupy the first and second position (or vice versa) of the final parallel ranking (PR). Between this pair of molecules, the first will be the compound with the lower sum of both ranking positions. Accordingly, the molecules ranked second for each method would be re-ranked third and fourth, and so on until all molecules are reordered.

The other two protocols fall into the hybrid category.[59] The rescoring ranking (RR) protocol generates a ranking based on the scores of a 3D similarity method[23] using the alignment obtained by the SB approach. Finally, the consensus ranking (CR) is based on the combination of the SB (docking) ranking and the RR, following the protocol formulated in parallel ranking.

The Tanimoto coefficient and Tversky coefficient[60] were used to score the docked poses in both RR and CR. None of these methods require any prior knowledge or input other than the results from the single methods and, thus, are directly applicable.

Performance Evaluation. Receiver Operator Characteristic (ROC) curves and Area Under the ROC Curve (AUC) were used as the metrics to assess the performance of the three SB+LB re-ranking strategies.[61–63] On the other hand, the ROC enrichment factor (ROCe, Eq. 1) captures the performance at a given percentage of the poses at the top of the ranking,.

$$R\mathrm{OCe\ X\%} = \frac{\frac{\mathrm{N}^{\mathrm{X\%}}_{\mathrm{actives\ selected}}}{\mathrm{N\ total\ actives}}}{\frac{\mathrm{N}^{\mathrm{X\%}}_{\mathrm{decoys\ selected}}}{\mathrm{N\ total\ decoys}}} = \frac{\frac{\mathrm{TP}}{\mathrm{TP+FN}}}{\frac{\mathrm{FP}}{\mathrm{TN+FP}}} = \frac{\mathrm{sensitivity}}{1-\mathrm{specificity}} \qquad (1)$$

ROCe values at false positive rates of 0.5%, 1.0%, 2.0%, 5.0% are reported as suggested by Jan and Nicholls.[64] Both metrics –AUC and ROCe– were used to validate the combination strategies.

In addition, chemotype clustering analyses was included in our evaluation throughout the awROCe values[65], Eq.2. This parameter was determined taking into account the same percentages adopted for ROCe.

$$awR\mathrm{OCe\ X\%} = \frac{\dfrac{\sum_{j}^{N_{clusters}} \sum_{i}^{N_j} w_{ij} a_{ij}^{X\%}}{\mathrm{N\ clusters}}}{\dfrac{\mathrm{N}_{\mathrm{decoys\ selected}}^{\mathrm{X\%}}}{\mathrm{N\ total\ decoys}}} \qquad (2)$$

where $w_{ij} = \frac{1}{N_j}$ is the weight of the i^{th} structure from the j^{th} cluster, N_j is the number of structures in a given cluster, $a_{ij}^{X\%}$ is 1 or 0 depending on whether the ith structure of the j^{th} cluster already (respectively) appeared or not in the chosen fraction of the dataset. With this solution the value of the true positive hit is weighted depending on the cluster to which it belongs to and on the number of molecules in the cluster.

RESULTS AND DISCUSSION

The results obtained from the VS are reported in Table 1, which shows the average ROCe in the top 0.5%, 1%, 2% and 5% and the average AUC (results for individual sets are provided in SI, Tables S2 and S3). The comparison of these parameters permits to assess the performance of the three combination strategies (PR, RR, and CR) considered in this study. The results in Table 1 are obtained using the SP mode of Glide considering the two similarity metrics (Tanimoto and Tversky). The analysis of the results obtained for Glide using the HTVS mode, where the poses are expected to be less accurate, are reported in SI (Table S2) and discussed later. Furthermore, the awROCe metric was employed to evaluate the impact of structural analogues in early enrichment.

Table 1. AUC and ROCe metrics for PharmScreen, Glide SP, and the three combination strategies (PR, RR, CR).

					GLOBAL SIMILARITY		**PARTIAL SIMILARITY**	
		Pharm Screen	Glide SP	PR	RR	CR	RR	CR
BS1	ROCE 0.5	33.5	28.3	37.3	32.9	32.2	**43.1**	**43.1**
	ROCE 1	21.2	18.8	24.0	22.0	22.7	27.5	**27.9**
	ROCE 2	12.5	12.3	15.4	13.0	14.5	17.0	**17.5**
	ROCE 5	6.6	6.8	8.2	6.7	8.0	8.5	**9.5**
	AUC	0.66	0.74	0.77	0.71	0.76	0.76	**0.8**
BS2	ROCE 0.5	17.7	34.6	30.8	38.6	**41.4**	32.8	39.0
	ROCE 1	10.5	20.6	22.0	22.6	**30.6**	16.4	28.0
	ROCE 2	8.4	12.6	14.4	13.3	**18.1**	11.3	17.1
	ROCE 5	6.6	7.1	8.6	7.1	**9.4**	6.2	9.0
	AUC	0.72	0.72	**0.81**	0.73	0.77	0.73	0.78

Assessment of the combination strategies derived using global similarity. For BS1, the three combination strategies lead to a slightly higher performance compared to either PharmScreen or Glide (Table 1) when the global similarity measurement is used. Different trends are, however, observed for the BS2 dataset, where the combined approach improves both ROCe and AUC, especially for the CR.

Even though the average trends using global similarity do not show a remarkable difference in the overall performance, significant differences can be found for individual members of the dataset (see SI Tables S2 and S3), as can be found in the results obtained for BS2. In this case, CR leads to a remarkable improvement in both ROCe and AUC compared to PharmScreen and Glide. On the other hand, RR performs better than PR in recovering actives in the initial stages of the VS. Furthermore, the improvement found PR is challenged by the higher computational cost required for this method.

The analysis of the average trends masks the occurrence of significant improvements observed for individual targets. This is illustrated by the behavior observed for DHFR_BS2. In this target, a narrow and deep pocket defines a single binding mode, which is shown in Figure 1. Moreover, the query and 24 hits share in their structure a pyrido[2-3]pyrimidine ring, which is able to form 3 hydrogen bonds with 3 amino acids (Glu30A, Ile7A, and Val115A)[66] settled at the bottom of the pocket, thus favoring the definition of a unique specific binding mode. Thus, the query and most of the docked hits show a high overlap (see Figure 1). These conditions are ideal for the application of the RR protocol. Let us note that for DHFR_BS1, which shares the biological target with DHFR_BS2, RR also performs better than the rest of the methods, as noted in the fact that ROCe 0.5% increased from 25.6 for Glide SP to 76.9 for the RR method (SI Table S2).

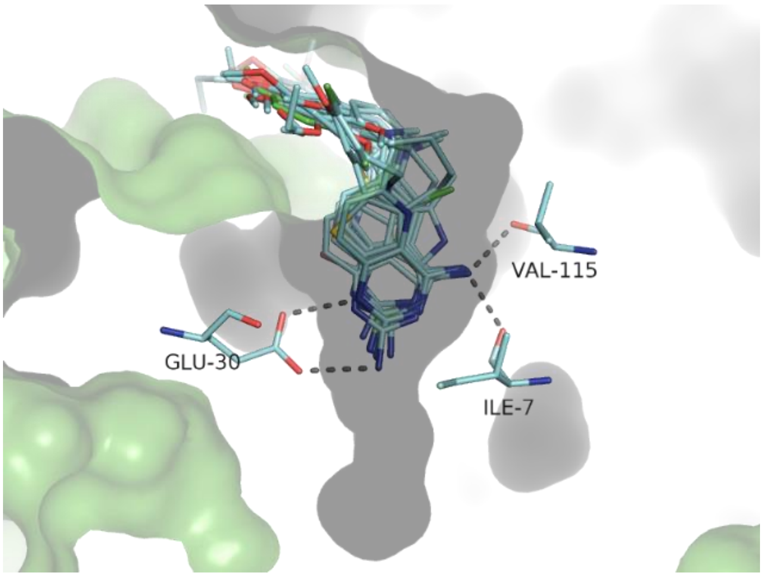


Figure 1. Binding mode of DHFR (PDB code: 1kmv, green), co-crystallized reference molecule (green), docked molecules with a pyrido[2-3]pyrimidine ring using Glide (cyan).

Conversely, Trypsin_BS1 presents an open and superficial pocket, where a significant part of the crystallized ligand is exposed to the solvent (Figure 2). In this case, the ligands exhibit a higher diversity in their binding mode to the pocket, and hence RR, which computes the similarity from the crystallized ligand, performs worse. Nevertheless, the challenges posed by the existence of multiple binding modes are solved by the use of CR, which appears to be a suitable strategy to correct the limitations of either docking and similarity measurements.

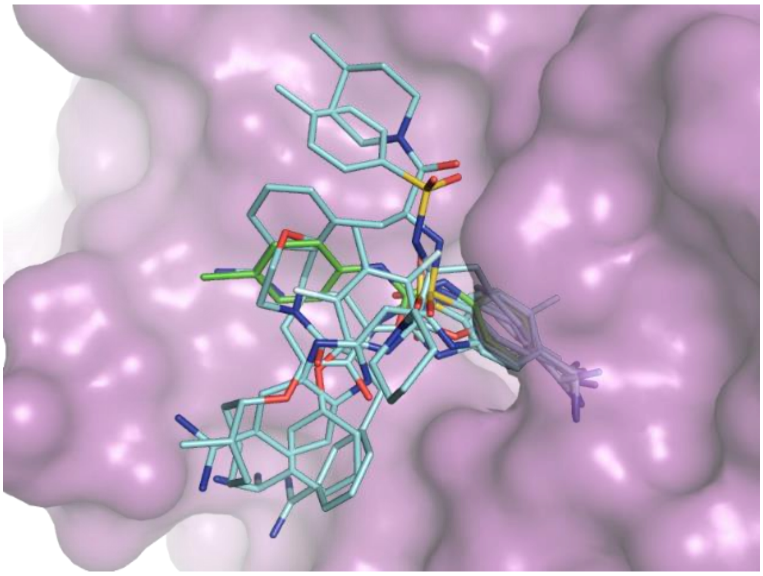


Figure 2. Binding mode of beta-trypsin (PDB code: 1bju, purple), co-crystallized reference molecule (green), docked molecules using Glide (cyan).

Although the preceding results give support to the adoption of the RR method, there are cases where this strategy leads to a negligible improvement in the re-ranking of compounds. This is exemplified by COMT_BS1, which also exhibits a large solvent-exposed pocket (Figure 3A). In this case, RR is not able to account for the existence of multiple binding modes. Two causes have been identified: (1) the results obtained with Glide are not high enough to enhance RR, and (2) although the molecules of this set share the same binding mode, the different size between five actives (ZINC03814485, ZINC00392003, ZINC03814484, ZINC00021789, and ZINC00330141) and the crystallographic reference penalize the 3D similarity evaluation (Figure 3B). None of them is re-ranked above 5% of the ranking.

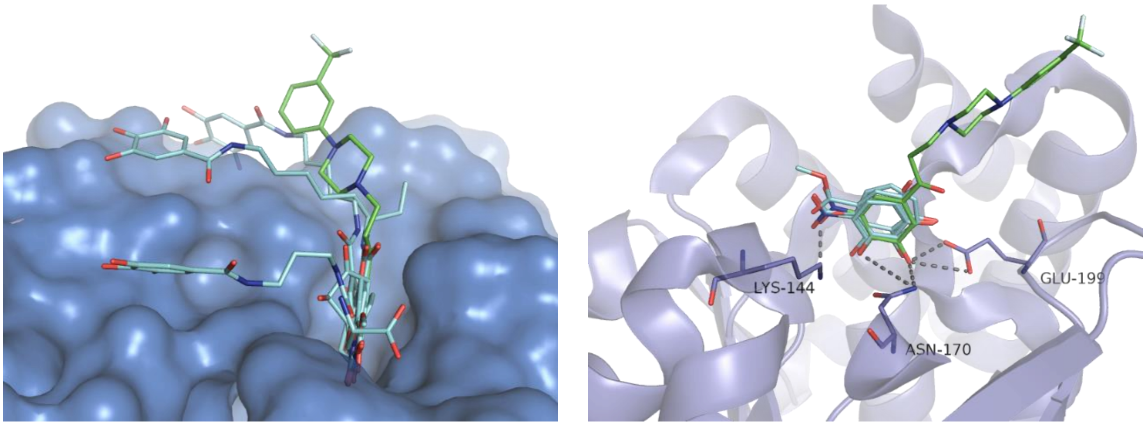


Figure 3. Left, binding mode of catechol O-metiltransferasa (PDB code: 1h1d, blue), co-crystallized reference molecule (green), docked molecules using Glide (cyan). Right, reference molecule (green, co-crystallized structure BIA) and 5 docked ligands of COMT set (ZINC03814485, ZINC00392003, ZINC03814484, ZINC00021789, and ZINC00330141) by Glide (cyan) in the

binding site of catechol O-metiltransferasa (blue). 5 possible hydrogen bonds are reported between the reference and lysine 144, asparagine 170 and glutamine 199.

In summary, the results point out that CR is the combination strategy that recovers more actives. However, the performance may be limited by the constraints imposed by measurements of global similarity against the reference compound, besides the potential influence exerted by the occurrence of different binding modes for the set of ligands.

Influence of partial similarity on the performance of hybrid approaches. The usage of a partial similarity measurement, such as Tversky coefficient, can offer a suitable tradeoff to alleviate the impact of the preceding problem and improve the performance of hybrid approaches.

Table 1 shows the ROCe and AUC values obtained from the application of a partial similarity coefficient (the results for all sets are reported in SI Table S3). The RR leads to a notable increase in both ROCe and AUC when the Tversky coefficient is used for most BS1 sets, even improving the behavior observed for PR but at a much lower computational expensiveness. As an example, let us note that 4 out of 5 hits (ZINC03814485, ZINC00392003, ZINC03814484, and ZINC00330141) showed in Figure 3 are rescored within 1% of the ranking using partial similarity measurements. In addition, Figure 4 shows the hits found in the ROCe 0.5% using the Tversky coefficient for SRC and PPAR_gamma, which is contrast with the lack of actives in these early enrichments when the global similarity (Tanimoto) metrics is considered. Accordingly, the performance of the CR also exhibits an improved performance, outperforming the ranking obtained from Glide SP in both BS1 and BS2 datasets. With this enhancement, CR coupled to partial similarity measurements becomes the best performing protocol for recovering actives among the combination strategies.

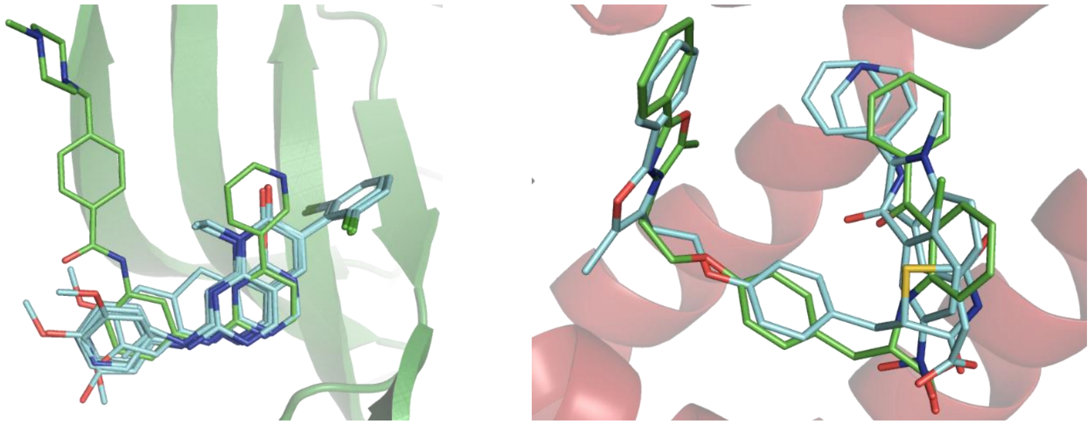


Figure 4. Left, binding mode of tyrosine kinase (PDB code: 2src, green), co-crystallized reference molecule (cyan), and docked hits using Glide in the ROCe 0.5% (cyan). Right, binding mode of peroxisome proliferator-activated receptor gamma (PDB code: 1fm9, green), co-crystallized reference molecule (cyan), and docked hits using Glide in the ROCe 0.5% (cyan).

As a final remark, it is worth examining in more detail the results obtained for BS2. Compared to the results reported using global similarity, slightly lower average values are obtained using partial similarity, as noted in the decrease of ROCe 1% from30.6% to 28.0. This is primarily due to the results obtained for HIV1PR_BS2 and GR_BS2, whereas there is generally an improvement for DHFR_BS2 and VEGFR2_BS2 (SI Table S3 and S4). A detailed analysis showed that some decoys positioned in the lower part of the ranking discarded by Tanimoto coefficient are shifted to the top when similarity is evaluated with the Tversky coefficient. Thus, for HIVPR_BS2, 16 out of 18 first decoys (ROCe 1%) using local similarity are ranked lower than position 95 with global similarity. Similarly, for GR_BS2, 7 out of 14 first decoys (ROCe 1%) using Tversky coefficient are ranked lower than position 70 with Tanimoto metrics. This behavior, which primarily arises from the comparison of molecules with notable differences in their size, is shown in Figure 5, which displays two representative decoys for RR using the local similarity for HIV-1 protease and glucocorticoid receptor.

Globally, CR in combination with local similarity measurements outperforms all other methods, although RR may occasionally give slightly improved results for specific targets.

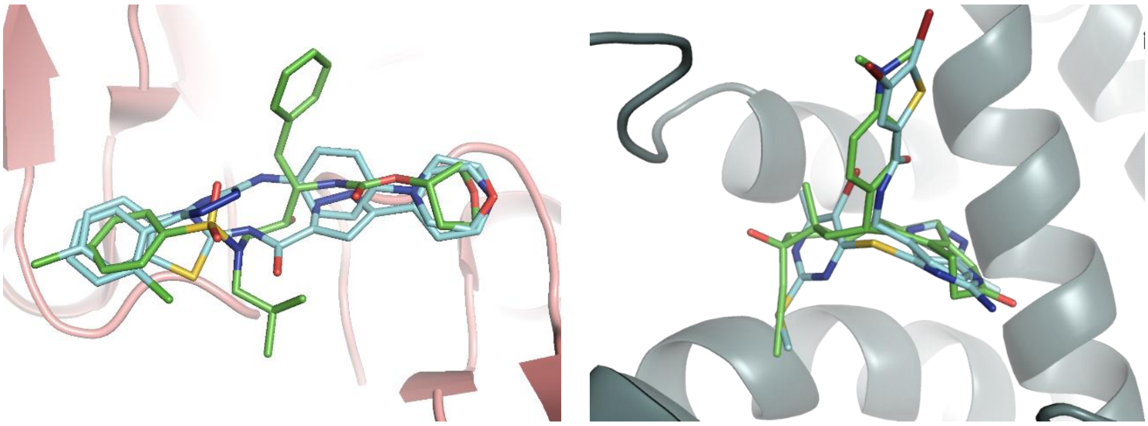


Figure 5. Left, binding mode of HIV-1 protease (PDB code: 3nu3, green), co-crystallized reference molecule (cyan), and docked hits using Glide in the ROCe 0.5% (cyan). Right, binding mode of glucocorticoid receptor (PDB code: 1nhz, green), co-crystallized reference molecule (cyan), and docked hits using Glide in the ROCe 0.5% (cyan).

Robustness assessment. To further analyze the performance consistency, heatmaps of the hierarchical position of each approach among the others for ROCe 1% and AUC are reported in this section (the remaining metrics are presented in SI Figures S1 and S2). Figure 6 and 7 shows the comparison of the combination methods using local similarity, which returns an increase in the performance against pure PharmScreen and Glide SP methods. Figure 6, corroborates the better performance of CR, showing higher robustness in addition to higher performance. In particular, Figure 7, which compares CR directly with PharmScreen and Glide SP, shows that CR is never ranked third for ROCe 1% and only in one case for AUC, giving support to the combined use of LB and SB rankings. Thus, CR improves clearly the results offering a balanced alternative to standard VS methods.

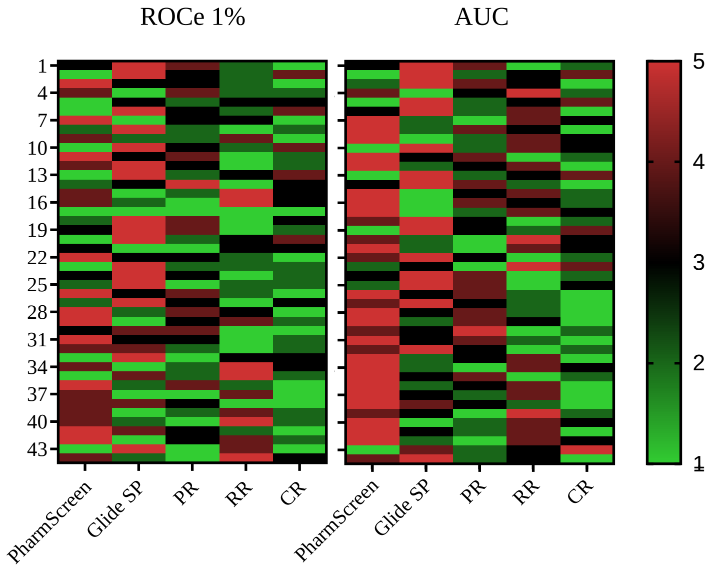


Figure 6. Heatmap of the hierarchical position of all methods for ROCe 1 % and AUC. The color scale is indicative of the position, being the first green and the fifth red.

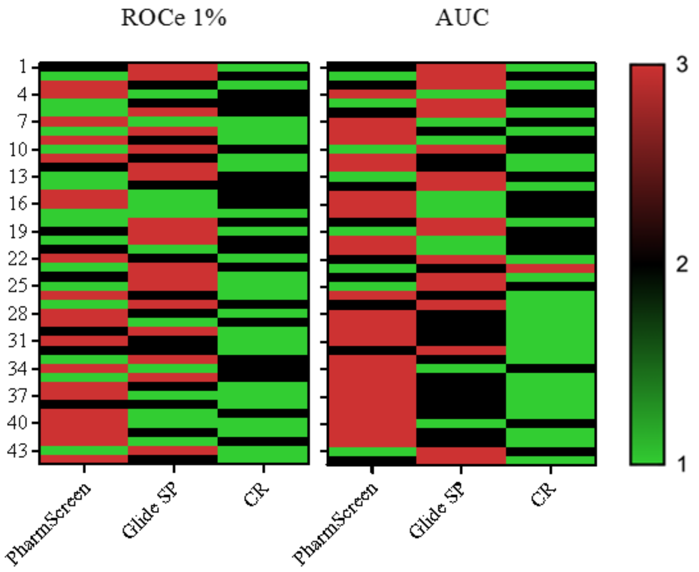


Figure 7. Heatmap of the hierarchical position of CR with local similarity against PharmScreen and Glide SP performance for ROCe 1% and AUC. The method ranked first is shown in green, the second in black and the third in red.

Finally, the effect of global versus local similarity measurements in CR is shown in Figure 8 (see SI Figure S3 for the analysis of the remaining metrics). Contrary to what is observed in the average performance value of Table1, for BS2 (last 4 sets), only 2 sets perform better using Tanimoto coefficient than Tversky metrics (ROCe 1% and AUC).

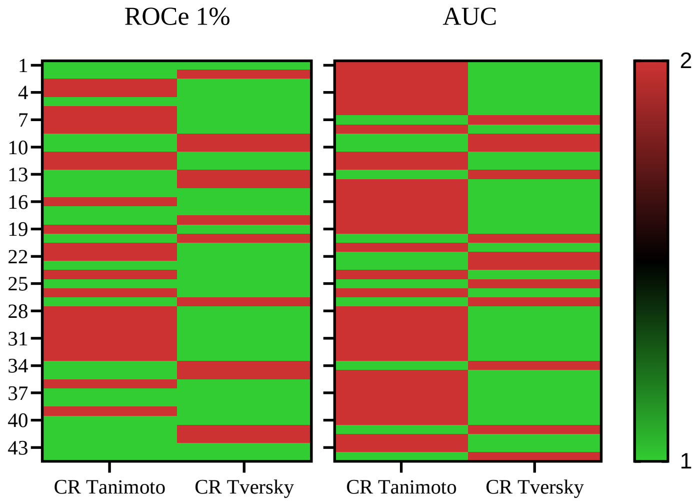


Figure 8. Heatmap of the hierarchical position of CR using global (Tanimoto) and local (Tversky) similarity measurements for ROCe 0.5% and AUC. The method ranked first is shown in green and the second in red.

Comparison of Combined HTVS methods against Glide SP. Finding a balance between computational efficiency and accuracy in predictions is a relevant aspect to be considered in VS. In this section, the influence of using the Glide HTVS score on the performance of combined protocols is evaluated and compared with the results discussed above for Glide SP.

In general, the combined methods derived from Glide SP perform better (see SI Tables S2 and S3). However, similar trends can also be observed for Glide HTVS. If a global similarity coefficient is used (Tn), PR and CR return a better performance and robustness (SI Tables S4) than the other methods for BS2. However, unlike SP, since the lower values of Glide HTVS affect CR, the performance values reported by RR are the highest for BS2. When local similarity is used, CR is found to be the best option in terms of performance for BS1. However, CR performance falls below PR for BS2 in all average metrics.

Finally, the performance of combined HTVS approaches is compared with Glide SP to determine if similar results are obtained with a considerable time reduction. Table 2 suggests that combined methods emerged from HTVS in combination with partial similarity overcome Glide SP results.

Table 2. AUC and ROCe metrics for Glide SP and the three combination strategies (PR, RR, CR) derived from Glide HTVS.

				GLOBAL SIMILARITY		**PARTIAL SIMILARITY**	
		Glide SP	PR	RR	CR	RR	CR
ALL DATA SETS	ROCE 0.5	32.0	35.6	30.2	31.7	39.3	**40.1**
	ROCE 1	20.2	23.9	18.6	20.7	**24.5**	24.3
	ROCE 2	12.1	**15.3**	11.3	13.2	14.0	15.2
	ROCE 5	6.6	**7.9**	6.0	7.0	7.8	7.8
	AUC	0.7	**0.8**	0.7	0.7	0.7	0.7

Chemotype diversity analysis. To ensure the chemical diversity of hits founded, a weighting scheme based on the ROC metric following ligand clustering is applied[65]. Since this analysis requires a clustered set, DUD_LIB_VS_1.0,[67] BS1 is the only benchmark employed in this section. Table 3 shows the average of awROCe at different percentages for PharmScreen, Glide SP, and the best combined method, CR with local similarity. The results point out that the latter achieved the best overall performance in chemotype enrichment, in line with ROCe values. In addition, the heatmaps of robustness for the awROCe 1% and 5% are reported in Figure 10. Individual values and heatmaps at all percentages are shown in SI Tables S4 and S5. For none of the percentages studied more than two sets are classified in third position using CR with local similarity (Figure 10 and SI Figure S5). The improvement of chemotype enrichment (awROCe) using CR corroborates the synergy between LB and SB, apart from the bias of similar chemotypes existence.

Table 3. awROCe metrics for Glide SP and the three combination strategies (CR) derived from Glide SP.

		PharmScreen	Glid SP	CR
BS1	awROCe 0.5	26.4	29.3	41.4
	awROCe 1	17.4	19.4	26.0
	awROCe 2	10.3	11.8	11.6
	awROCe 5	5.9	6.6	8.9

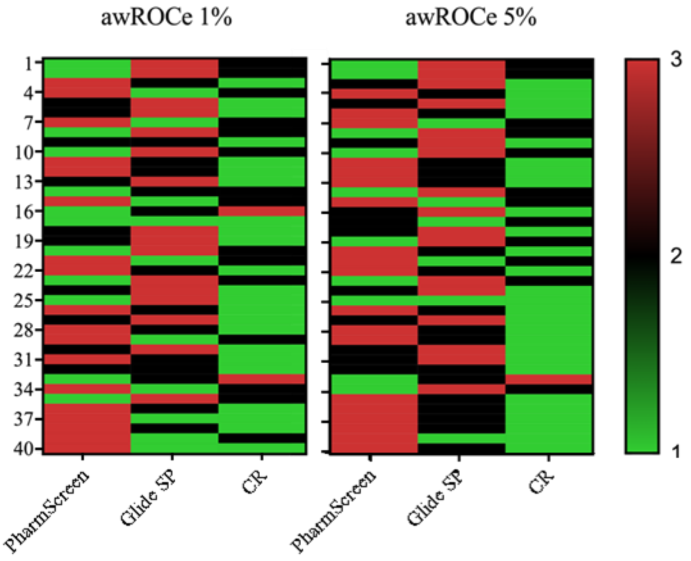


Figure 10. Heatmap of the hierarchical position of Consensus Ranking using Tv among PharmScreen and Glide SP performance for awROCe 1% and 5%. The method ranked first is shown in green, the second in black and the third in red.

CONCLUSIONS

Since (de)solvation is fundamental for the establishment of the ligand-receptor complex, it can be expected that ligands docked in the same pocket share lipophilic characteristics which are complementary to the residues that shape the binding pocket, even if there are several binding modes. Thus, lipophobicity similarity is hypothesized as a valid scoring function for discerning between active and inactive compounds.

In this work, we have explored three alternatives to combine topological distribution of LB-lipophilic similarity and SB approaches. The fusion of 3D similarity and docking output was based on the idea that deficiencies in one method would be compensated for by others, inspired by the "consensus scoring"[68] in the docking field. To address the proposed approaches, a proof-of-concept investigation was carried out. For 44 data sets, including 41 targets, 3D similarity and docking score performance was compared against the combined methods.

The results show that combined protocols are a valuable tool in VS. Combined ranking reduces the dependency on single VS method performance as well as having the potential to outperform the best single method used. We show that, on average, 44 data sets investigated herein, combined methods recover more active compounds than individual LB and SB tools. Among the proposed protocols, CR using partial similarity has the best average performance in recovering actives in the data sets, but both RR and PR also have good performance.

An essential feature of the combined methods introduced herein is that 3D similarity calculations are independent of the generation of docking poses. Hence, any existing ranking can also be re-evaluated based on 3D similarity calculations relative to experimental binding modes.

These findings support the usefulness of $\text{LogP}_{\text{ele}}/\text{LogP}_{\text{cav}}/\text{HB}$ as driver descriptors in molecular similarity studies in promoting their use in virtual screening campaigns in combination with SB techniques.

REFERENCES

(1) Sperandio, O.; Miteva, M.; Villoutreix, B. Combining Ligand- and Structure-Based Methods in Drug Design Projects. *Curr. Comput. Aided-Drug Des.* **2008**, *4* (3), 250–258.

(2) Drwal, M. N.; Griffith, R. Combination of Ligand- and Structure-Based Methods in Virtual Screening. *Drug Discov. Today Technol.* **2013**, *10* (3), e395–e401.

(3) Lavecchia, A.; Di Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.* **2013**, *20* (23), 2839–2860.

(4) Macalino, S. J. Y.; Gosu, V.; Hong, S.; Choi, S. Role of Computer-Aided Drug Design in Modern Drug Discovery. *Arch. Pharm. Res.* **2015**, *38* (9), 1686–1701.

(5) Klopmand, G. Concepts and Applications of Molecular Similarity, by Mark A. Johnson and Gerald M. Maggiora, Eds., John Wiley &amp; Sons, New York, 1990, 393 Pp. *J. Comput. Chem.* **1992**, *13* (4), 539–540.

(6) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B.-T. Molecular Similarity and Diversity in Chemoinformatics: From Theory to Applications. *Mol. Divers.* **2006**, *10* (1), 39–79.

(7) Villoutreix, B. O.; Renault, N.; Lagorce, D.; Sperandio, O.; Montes, M.; Miteva, M. A. Free

Resources to Assist Structure-Based Virtual Ligand Screening Experiments. *Curr. Protein Pept. Sci.* **2007**, *8* (4), 381–411.

(8) Banegas-Luna, A.-J.; Cerón-Carrasco, J. P.; Pérez-Sánchez, H. A Review of Ligand-Based Virtual Screening Web Tools and Screening Algorithms in Large Molecular Databases in the Age of Big Data. *Future Med. Chem.* **2018**, *10* (22), 2641–2658.

(9) Hein, M.; Zilian, D.; Sotriffer, C. A. Docking Compared to 3D-Pharmacophores: The Scoring Function Challenge. *Drug Discov. Today Technol.* **2010**, *7* (4), e229–e236.

(10) Wilson, G. L.; Lill, M. A. Integrating Structure-Based and Ligand-Based Approaches for Computational Drug Design. *Future Med. Chem.* **2011**, *3* (6), 735–750.

(11) Borges, N. M.; Rodrigues Sartori, G.; Ribeiro, J. F. R.; Rocha, J. R.; Martins, J. B. L.; Montanari, C. A.; Gargano, R. Similarity Search Combined with Docking and Molecular Dynamics for Novel HAChE Inhibitor Scaffolds. *J. Mol. Model.* **2018**, 24–41.

(12) Anighoro, A.; Bajorath, J. A Hybrid Virtual Screening Protocol Based on Binding Mode Similarity. In *Rational Drug Design*; Humana Press, New York, NY, 2018; pp 165–175.

(13) da Silva Figueiredo Celestino Gomes, P.; Da Silva, F.; Bret, G.; Rognan, D. Ranking Docking Poses by Graph Matching of Protein–Ligand Interactions: Lessons Learned from the D3R Grand Challenge 2. *J. Comput. Aided. Mol. Des.* **2018**, *32* (1), 75–87.

(14) Banoglu, E.; Çalışkan, B.; Luderer, S.; Eren, G.; Özkan, Y.; Altenhofen, W.; Weinigel, C.; Barz, D.; Gerstmeier, J.; Pergola, C.; et al. Identification of Novel Benzimidazole Derivatives as Inhibitors of Leukotriene Biosynthesis by Virtual Screening Targeting 5-Lipoxygenase-Activating Protein (FLAP). *Bioorg. Med. Chem.* **2012**, *20* (12), 3728–3741.

(15) Smith, J. R.; Evans, K. J.; Wright, A.; Willows, R. D.; Jamie, J. F.; Griffith, R. Novel Indoleamine 2,3-Dioxygenase-1 Inhibitors from a Multistep in Silico Screen. *Bioorg. Med. Chem.* **2012**, *20* (3), 1354–1363.

(16) Drwal, M. N.; Agama, K.; Wakelin, L. P. G.; Pommier, Y.; Griffith, R. Exploring DNA Topoisomerase I Ligand Space in Search of Novel Anticancer Agents. *PLoS One* **2011**, *6* (9), e25150.

(17) Swann, S. L.; Brown, S. P.; Muchmore, S. W.; Patel, H.; Merta, P.; Locklear, J.; Hajduk, P. J. A Unified, Probabilistic Framework for Structure- and Ligand-Based Virtual Screening. *J. Med. Chem.* **2011**, *54* (5), 1223–1232.

(18) Svensson, F.; Karlén, A.; Sköld, C. Virtual Screening Data Fusion Using Both Structure- and Ligand-Based Methods. *J. Chem. Inf. Model.* **2012**, *52* (1), 225–232.

(19) Tan, L.; Geppert, H.; Sisay, M. T.; Gütschow, M.; Bajorath, J. Integrating Structure- and Ligand-Based Virtual Screening: Comparison of Individual, Parallel, and Fused Molecular Docking and Similarity Search Calculations on Multiple Targets. *ChemMedChem* **2008**, *3* (10), 1566–1571.

(20) Larsson, M.; Fraccalvieri, D.; Andersson, C. D.; Bonati, L.; Linusson, A.; Andersson, P. L. Identification of Potential Aryl Hydrocarbon Receptor Ligands by Virtual Screening of Industrial Chemicals. *Environ. Sci. Pollut. Res.* **2018**, *25* (3), 2436–2449.

(21) 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* (4), 943–955.

(22) Anighoro, A.; Bajorath, J. A Hybrid Virtual Screening Protocol Based on Binding Mode Similarity. In *Methods in molecular biology (Clifton, N.J.)*; 2018; Vol. 1824, pp 165–175.

(23) Anighoro, A.; Bajorath, J. Three-Dimensional Similarity in Molecular Docking: Prioritizing Ligand Poses on the Basis of Experimental Binding Modes. *J. Chem. Inf. Model.* **2016**, *56* (3), 580–587.

(24) Anighoro, A.; Bajorath, J. Binding Mode Similarity Measures for Ranking of Docking Poses: A

Case Study on the Adenosine A2A Receptor. *J. Comput. Aided. Mol. Des.* **2016**, *30* (6), 447–456.

(25) Anighoro, A.; Bajorath, J. Compound Ranking Based on Fuzzy Three-Dimensional Similarity Improves the Performance of Docking into Homology Models of G-Protein-Coupled Receptors. *ACS Omega* **2017**, *2* (6), 2583–2592.

(26) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein - Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.

(27) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Eneregies for Protein- Ligand Bindning. *J. Med. Chem.* **2006**, *49*, 5880−5884.

(28) Curutchet, C.; Orozco, M.; Luque, F. J. Solvation in Octanol: Parametrization of the Continuum MST Model. *J. Comput. Chem.* **2001**, *22* (11), 1180–1193.

(29) Muñoz-Muriedas, J.; Perspicace, S.; Bech, N.; Guccione, S.; Orozco, M.; Luque, F. J. Hydrophobic Molecular Similarity from MST Fractional Contributions to the Octanol/Water Partition Coefficient. *J. Comput. Aided. Mol. Des.* **2005**, *19* (6), 401–419.

(30) Vázquez, J.; Deplano, A.; Herrero, A.; Ginex, T.; Gibert, E.; Rabal, O.; Oyarzabal, J.; Herrero, E.; Luque, F. J. Development and Validation of Molecular Overlays Derived from Three-Dimensional Hydrophobic Similarity with PharmScreen. *J. Chem. Inf. Model.* **2018**, *58* (8), 1596–1609.

(31) Ginex, T.; Vazquez, J.; Gilbert, E.; Herrero, E.; Luque, F. J. Lipophilicity in Drug Design: An Overview of Lipophilicity Descriptors in 3D-QSAR Studies. *Future Med. Chem.* **2019**, fmc-2018-0435.

(32) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking Benchmarking Sets for Molecular Docking. *Society* **2006**, *49* (23), 6789–6801.

(33) Irwin, J. J. Community Benchmarks for Virtual Screening. *J. Comput. Aided. Mol. Des.* **2008**, *22* (3–4), 193–199.

(34) Good, A. C.; Oprea, T. I. Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: A Help or Hindrance in Tool Selection? *J. Comput. Aided. Mol. Des.* **2008**, *22* (3–4), 169–178.

(35) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal Assignment Methods for Ligand-Based Virtual Screening. *J. Cheminform.* **2009**, *1*, 14.

(36) Niu Huang; Brian K. Shoichet.; Irwin, J. J. Benchmarking Sets for Molecular Docking. **2006**.

(37) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2002**, *41* (5), 1308–1315.

(38) Edward J. Barker.; Eleanor J. Gardiner.; Valerie J. Gillet.; Paula Kitts.; Morris, J. Further Development of Reduced Graphs for Identifying Bioactive Compounds. **2003**.

(39) Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 – A Public Library of Challenging Docking Benchmark Sets. *J. Chem. Inf. Model.* **2013**, *53* (6), 1447–1462.

(40) Nicholls, A. What Do We Know and When Do We Know It? *J. Comput. Aided. Mol. Des.* **2008**, *22* (3–4), 239–255.

(41) Anighoro, A.; Bajorath, J. Three-Dimensional Similarity in Molecular Docking: Prioritizing Ligand Poses on the Basis of Experimental Binding Modes. *J. Chem. Inf. Model.* **2016**, *56* (3), 580–587.

(42) Chen, X.; Liu, M.; Gilson, M. BindingDB: A Web-Accessible Molecular Recognition Database. *Comb. Chem. High Throughput Screen.* **2001**, *4* (8), 719–725.

(43) Vázquez, J.; Deplano, A.; Herrero, A.; Ginex, T.; Gibert, E.; Rabal, O.; Oyarzabal, J.; Herrero, E.; Luque, F. J. Development and Validation of Molecular Overlays Derived from Three-Dimensional

Hydrophobic Similarity with PharmScreen. *J. Chem. Inf. Model.* **2018**, *58* (8), 1596–1609.

(44) Schrodinger. Glide. LLC: New York.

(45) 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.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749.

(46) 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. *J. Med. Chem.* **2004**, *47* (7), 1750–1759.

(47) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. RM1: A Reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.* **2006**, *27* (10), 1101–1111.

(48) Forti, F.; Barril, X.; Luque, F. J.; Orozco, M. Extension of the MST Continuum Solvation Model to the RM1 Semiempirical Hamiltonian. *J. Comput. Chem.* **2008**, *29* (4), 578–587.

(49) Luque, F. J.; Orozco, M. Version Locally Modified MOPAC 6.0. University of Barcelona 2008.

(50) MolVS 0.1.1 https://pypi.org/project/MolVS/ (accessed Apr 11, 2018).

(51) Landrum, G. RDKit: Open-Source Cheminformatics. 2006.

(52) RCSB Protein Data Bank http://www.pdb.org.

(53) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. *J. Chem. Inf. Model.* **2008**, *48* (11), 2108–2117.

(54) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen : Virtual Screening Using Molecular Fields . Application to the DUD Data Set FieldScreen : Virtual Screening Using Molecular Fields . Application to the DUD Data Set. *J. Chem. Inf. Model* **2008**, *48* (11), 2108–2117.

(55) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal Assignment Methods for Ligand-Based Virtual Screening. *J. Cheminform.* **2009**, *1* (1), 1–23.

(56) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41* (23), 4502–4520.

(57) Chen, Q.; Higgs, R. E.; Vieth, M. Geometric Accuracy of Three-Dimensional Molecular Overlays. *J. Chem. Inf. Model.* **2006**, *46* (5), 1996–2002.

(58) Ginex, T.; Muñoz-Muriedas, J.; Herrero, E.; Gibert, E.; Cozzini, P.; Luque, F. J. Development and Validation of Hydrophobic Molecular Fields Derived from the Quantum Mechanical IEF/PCM-MST Solvation Models in 3D-QSAR. *J. Comput. Chem.* **2016**, *37* (13), 1147–1162.

(59) Drwal, M. N.; Griffith, R. Combination of Ligand- and Structure-Based Methods in Virtual Screening. *Drug Discov. Today Technol.* **2013**, *10*, e395–e401.

(60) Crisman, T. J.; Bender, A.; Milik, M.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Fejzo, J.; Hommel, U.; Davies, J. W.; Glick, M. "Virtual Fragment Linking": An Approach To Identify Potent Binders from Low Affinity Fragment Hits. *J. Med. Chem.* **2008**, *51* (8), 2481–2491.

(61) Truchon, J.F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47* (2), 488–508.

(62) Zhao, W.; Hevener, K. E.; White, S. W.; Lee, R. E.; Boyett, J. M. A Statistical Framework to Evaluate Virtual Screening. *BMC Bioinformatics* **2009**, *10*, 225.

(63) Sheridan, R. P. Alternative Global Goodness Metrics and Sensitivity Analysis: Heuristics to Check the Robustness of Conclusions from Studies Comparing Virtual Screening Methods. *J. Chem. Inf. Model.* **2008**, *48* (2), 426–433.

(64) Jain, A. N.; Nicholls, A. Recommendations for Evaluation of Computational Methods. *J. Comput.*

Aided. Mol. Des. **2008**, *22* (3–4), 133–139.

(65) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal Assignment Methods for Ligand-Based Virtual Screening. *J. Cheminform.* **2009**, *1* (1), 14.

(66) Klon, A. E.; Héroux, A.; Ross, L. J.; Pathak, V.; Johnson, C. A.; Piper, J. R.; Borhani, D. W. Atomic Structures of Human Dihydrofolate Reductase Complexed with NADPH and Two Lipophilic Antifolates at 1.09 a and 1.05 a Resolution. *J. Mol. Biol.* **2002**, *320* (3), 677–693.

(67) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.

(68) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42* (25), 5100–5109.

Supporting Information

Similarity assessment of lipophilic distribution: A boost for structure-based methods.

Javier Vazquez,[†, ‡] Enric Herrero,[†] and F. Javier Luque[‡]

[†] *Pharmacelera, Plaça Pau Vila, 1, Sector 1, Edificio Palau de Mar, Barcelona 08039, Spain*

[‡] *Department of Nutrition, Food Science and Gastronomy, Faculty of Pharmacy and Food Sciences, Institute of Biomedicine (IBUB), and Institute of Theoretical and Computational Chemistry (IQTC-UB), University of Barcelona, Av. Prat de la Riba 171, Santa Coloma de Gramenet E-08921, Spain*

Table S1. List of targets included in BS1 and BS2. In brackets the number of hits discarded.

	Target	PDB code	Decoys	Ligands	Discarded htvs	Discarded Sp	Discarded Htvs + sp	Total molecules
BS1	ace	1o86	1796	46	10	1	10	1832
	ache	1eve	3859	99	32	15	34(2)	3924
	ada	1stw	927	23	4	2	6	944
	alr2	1ah3	986	26	11	2	12	1000
	ampc	1xgj	786	21	10	3	11	796
	ar	1xq2	2848	69	111	38	118(1)	2799
	cdk2	1ckp	2070	47	31	4	33(1)	2084
	comt	1h1d	468	11	1	0	1	478
	cox1	1p4g	910	23	0	0	0	933
	cox2	1cx2	12606	212	215	54	238(2)	12580
	dhfr	3dfr	8350	190	82	14	94	8446
	egfr	1m17	15560	365	106	21	116	15809
	er_agonist	1l2i	2568	63	44	25	59	2572
	er_antagonist	3ert	1058	18	55	5	59(3)	1017
	fgfr1	1agw	3462	71	30	12	33	3500
	fxa	1f0r	2092	64	19	19	19	2137
	gart	1c2t	155	8	2	0	2	161
	gpb	1a8i	2135	52	24	6	32(1)	2155
	gr	1m2z	2585	32	225	225	225	2392
	hivrt	1rt1	1494	34	47	17	50	1478
	hivpr	1hpx	9	4	0	0	0	13
	hmga	1hw8	1423	25	21	2	22	1426
	hsp90	1uy6	975	23	14	5	15	983
	inha	1p44	2707	57	20	1	21	2743
	mr	2aa2	636	13	38	6	42	607
	na	1a4g	1713	49	15	5	16	1746
	p38	1kv2	6779	137	16	3	18	6898
	parp	1efy	1350	31	12	2	12	1369
	pde5	1xp0	1698	26	69	11	73	1651
	pdgfrb	model	5603	124	60	12	63(2)	5664
	pnp	1b8o	1036	25	19	33	47(1)	1014
	ppar_gamma	1fm9	40	6	2	0	2	44
	pr	1sr7	920	22	35	14	39(1)	903
	rxr_alpha	1mvc	575	18	67	7	70	523
	sahh	1a7a	1346	33	44	26	66	1313
	src	2src	5679	98	27	8	34	5743
	thrombin	1ba8	1148	23	12	0	12	1159
	tk	1kim	891	22	10	4	11	902
	trypsin	1bju	718	9	13	3	13	714
	vegfr2	1vr2	2712	48	73	13	78(1)	2682
BS2	dhfr	1kmv	1200	40	24	7	24(1)	1216
	vegfr2	1nhz	1200	40	102	23	105(15)	1135
	gr	3nu3	1200	40	29	9	29(4)	1211
	hiv1pr	3vo3	1200	40	104	7	104(3)	1136

Table S2. ROCE and AUC values for all data sets included in BS1 and BS2 using Glide in HTVS mode. Parallel ranking (PR), rescoring ranking (RR) and consensus ranking (CR). Tv: Tversky coefficient.

ROCe 0.5%								
	Target	Pha	Glide	PR	RR	CR	CR tv	CR tv
BS1	ace	17.4	17.4	21.7	26.1	21.7	34.8	30.4

	ache	57.7	0.0	28.9	35.1	22.7	26.8	20.6
	ada	0.0	8.7	0.0	8.7	8.7	8.7	8.7
	alr2	7.7	15.4	15.4	7.7	15.4	15.4	23.1
	ampc	76.2	0.0	38.1	9.5	9.5	9.5	0.0
	ar	53.7	29.9	41.8	35.8	38.8	29.9	38.8
	cdk2	0.0	34.8	8.7	8.7	26.1	17.4	26.1
	comt	54.5	0.0	54.5	0.0	0.0	54.5	54.5
	cox1	0.0	8.7	8.7	26.1	17.4	17.4	17.4
	cox2	145.5	60.3	111.0	99.5	101.4	93.8	95.7
	dhfr	0.0	6.4	4.3	12.8	9.6	27.7	23.4
	egfr	15.9	15.9	23.0	20.3	19.7	35.6	35.1
	er_agonist	92.1	47.6	57.1	98.4	57.1	85.7	76.2
	er_antagonist	40.0	13.3	13.3	80.0	26.7	80.0	40.0
	fgfr1	8.5	33.8	33.8	2.8	25.4	0.0	25.4
	fxa	6.3	12.5	12.5	0.0	6.3	3.1	9.4
	gart	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	gpb	47.1	7.8	35.3	70.6	51.0	78.4	43.1
	gr	68.8	43.8	50.0	50.0	50.0	81.3	81.3
	hivrt	41.2	29.4	52.9	35.3	47.1	41.2	47.1
	hivpr	50.0	50.0	100.0	0.0	0.0	50.0	100.0
	hmga	40.0	80.0	80.0	88.0	104.0	80.0	104.0
	hsp90	34.8	0.0	34.8	26.1	26.1	26.1	26.1
	inha	80.7	3.5	52.6	7.0	10.5	7.0	7.0
	mr	107.7	123.1	138.5	123.1	138.5	138.5	138.5
	na	4.1	69.4	36.7	57.1	69.4	81.6	102.0
	p38	5.8	0.0	4.4	14.6	11.7	13.1	13.1
	parp	6.5	71.0	45.2	6.5	38.7	45.2	71.0
	pde5	23.1	15.4	30.8	0.0	15.4	46.2	38.5
	pdgfrb	13.1	8.2	13.1	11.5	11.5	31.1	27.9
	pnp	8.3	0.0	8.3	16.7	16.7	41.7	25.0
	ppar_gamma	33.3	0.0	33.3	33.3	33.3	133.3	33.3
	pr	28.6	0.0	19.0	19.0	9.5	9.5	9.5
	rxr_alpha	88.9	122.2	144.4	144.4	133.3	100.0	144.4
	sahh	24.2	0.0	6.1	36.4	12.1	12.1	6.1
	src	4.1	22.7	20.6	4.1	16.5	12.4	20.6
	thrombin	0.0	8.7	8.7	0.0	8.7	8.7	8.7
	tk	9.1	9.1	9.1	18.2	18.2	18.2	18.2
	trypsin	22.2	88.9	44.4	0.0	44.4	44.4	88.9
	vegfr2	21.3	46.8	55.3	12.8	42.6	8.5	38.3
	average	33.5	27.6	37.4	31.2	32.9	41.2	42.9
	St_desv	34.6	33.1	35.1	36.6	33.9	36.2	37.5
	max	145.5	123.1	144.4	144.4	138.5	138.5	144.4
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BS2	**dhfr**	25.6	0.0	20.5	30.8	25.6	41.0	15.4
	vegfr2	0.0	16.7	11.1	5.6	11.1	16.7	16.7
	gr	22.2	5.6	16.7	16.7	11.1	11.1	11.1
	hiv1pr	23.1	7.7	23.1	30.8	30.8	7.7	15.4
	average	17.7	7.5	17.8	20.9	19.7	19.1	14.6
	St_desv	11.9	6.9	5.2	12.2	10.1	15.1	2.4
	max	25.6	16.7	23.1	30.8	30.8	41.0	16.7
	min	0.0	0.0	11.1	5.6	11.1	7.7	11.1
All data sets	**average**	32.0	25.8	35.6	30.2	31.7	39.2	40.4
	St_desv	33.4	32.1	34.0	35.2	32.6	35.3	36.7
	max	145.5	123.1	144.4	144.4	138.5	138.5	144.4
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ROCe 1%								
	Target	**Pha**	**Glide**	**PR**	**RR**	**CR**	**CR tv**	**CR tv**
BS1	**ace**	15.2	13.0	13.0	13.0	13.0	21.7	19.6
	ache	47.4	1.0	25.8	21.6	16.5	18.6	11.3
	ada	0.0	4.3	4.3	13.0	8.7	8.7	8.7

	alr2	3.8	11.5	7.7	3.8	7.7	7.7	11.5
	ampc	38.1	0.0	28.6	4.8	4.8	4.8	4.8
	ar	31.3	28.4	35.8	19.4	23.9	28.4	22.4
	cdk2	2.2	17.4	15.2	6.5	17.4	8.7	19.6
	comt	27.3	18.2	36.4	0.0	0.0	27.3	36.4
	cox1	4.3	4.3	4.3	21.7	8.7	8.7	13.0
	cox2	76.6	37.8	70.3	52.6	54.1	50.2	52.6
	dhfr	0.5	3.7	3.2	10.6	7.4	18.6	15.4
	egfr	9.0	15.9	15.1	14.5	14.8	19.7	22.2
	er_agonist	74.6	49.2	63.5	55.6	58.7	57.1	58.7
	er_antagonist	60.0	33.3	26.7	46.7	46.7	60.0	46.7
	fgfr1	7.0	18.3	19.7	1.4	16.9	0.0	16.9
	fxa	3.1	7.8	9.4	1.6	4.7	3.1	6.3
	gart	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	gpb	45.1	7.8	25.5	43.1	35.3	49.0	29.4
	gr	37.5	21.9	31.3	28.1	25.0	43.8	40.6
	hivrt	23.5	14.7	26.5	17.6	23.5	23.5	26.5
	hivpr	25.0	25.0	50.0	0.0	0.0	25.0	50.0
	hmga	24.0	40.0	40.0	44.0	56.0	48.0	52.0
	hsp90	21.7	0.0	17.4	13.0	13.0	13.0	13.0
	inha	40.4	1.8	40.4	3.5	5.3	3.5	5.3
	mr	53.8	69.2	69.2	76.9	69.2	76.9	69.2
	na	2.0	36.7	30.6	34.7	44.9	57.1	55.1
	p38	3.6	0.0	2.9	10.2	7.3	8.0	6.6
	parp	3.2	51.6	35.5	16.1	35.5	38.7	48.4
	pde5	11.5	11.5	19.2	0.0	7.7	23.1	19.2
	pdgfrb	8.2	4.1	6.6	6.6	5.7	18.9	15.6
	pnp	4.2	8.3	4.2	12.5	8.3	29.2	12.5
	ppar_gamma	16.7	0.0	16.7	16.7	16.7	66.7	16.7
	pr	19.0	0.0	14.3	9.5	9.5	4.8	4.8
	rxr_alpha	50.0	66.7	72.2	72.2	72.2	61.1	72.2
	sahh	24.2	9.1	12.1	30.3	15.2	18.2	6.1
	src	5.2	14.4	13.4	3.1	11.3	9.3	13.4
	thrombin	0.0	17.4	4.3	0.0	8.7	17.4	13.0
	tk	4.5	4.5	9.1	9.1	9.1	9.1	9.1
	trypsin	11.1	77.8	55.6	22.2	55.6	33.3	66.7
	vegfr2	12.8	25.5	29.8	6.4	25.5	8.5	21.3
	average	21.2	19.3	25.1	19.1	21.6	25.7	25.8
	St_desv	21.1	20.4	20.0	19.9	20.2	20.9	20.5
	max	76.6	77.8	72.2	76.9	72.2	76.9	72.2
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BS2	**dhfr**	12.8	0.0	12.8	15.4	15.4	20.5	15.4
	vegfr2	2.8	11.1	5.6	5.6	8.3	8.3	8.3
	gr	11.1	8.3	11.1	13.9	8.3	5.6	5.6
	hiv1pr	15.4	7.7	15.4	19.2	15.4	15.4	7.7
	average	10.5	6.8	11.2	13.5	11.9	12.4	9.2
	St_desv	5.5	4.8	4.2	5.8	4.1	6.8	4.3
	max	15.4	11.1	15.4	19.2	15.4	20.5	15.4
	min	2.8	0.0	5.6	5.6	8.3	5.6	5.6
All data sets	**average**	20.2	18.2	23.9	18.6	20.7	24.5	24.3
	St_desv	20.4	19.8	19.5	19.1	19.5	20.3	20.1
	max	76.6	77.8	72.2	76.9	72.2	76.9	72.2
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ROCe 2%								
	Target	**Pha**	**Glide**	**PR**	**RR**	**CR**	**CR tv**	**CR tv**
BS1	**ace**	10.9	6.5	10.9	8.7	7.6	14.1	12.0
	ache	28.9	1.5	16.5	13.9	10.3	10.8	9.3
	ada	0.0	4.3	2.2	6.5	6.5	4.3	6.5
	alr2	1.9	5.8	5.8	1.9	5.8	3.8	7.7

	ampc	21.4	0.0	19.0	2.4	2.4	2.4	2.4
	ar	18.7	20.1	23.9	13.4	17.2	16.4	23.1
	cdk2	2.2	13.0	9.8	3.3	9.8	5.4	10.9
	comt	13.6	13.6	22.7	0.0	9.1	13.6	18.2
	cox1	2.2	4.3	4.3	10.9	10.9	6.5	6.5
	cox2	39.5	24.2	39.7	27.5	28.9	26.3	28.5
	dhfr	0.5	3.5	2.1	8.2	5.6	14.4	10.1
	egfr	7.0	11.1	12.2	9.3	10.8	11.2	12.7
	er_agonist	40.5	30.2	39.7	31.7	32.5	31.0	34.1
	er_antagonist	43.3	23.3	36.7	30.0	30.0	33.3	30.0
	fgfr1	6.3	9.2	10.6	0.7	9.2	1.4	8.5
	fxa	1.6	4.7	5.5	1.6	3.9	2.3	3.9
	gart	0.0	12.5	0.0	0.0	0.0	0.0	0.0
	gpb	27.5	3.9	18.6	24.5	19.6	25.5	20.6
	gr	23.4	10.9	18.8	15.6	14.1	23.4	21.9
	hivrt	14.7	10.3	14.7	10.3	11.8	11.8	14.7
	hivpr	12.5	12.5	25.0	0.0	0.0	12.5	25.0
	hmga	18.0	22.0	22.0	26.0	28.0	28.0	30.0
	hsp90	13.0	0.0	10.9	6.5	6.5	6.5	6.5
	inha	21.1	0.9	20.2	1.8	2.6	5.3	2.6
	mr	30.8	42.3	38.5	42.3	42.3	42.3	42.3
	na	1.0	18.4	18.4	25.5	25.5	30.6	31.6
	p38	4.0	0.0	1.8	6.6	5.1	4.4	4.0
	parp	1.6	33.9	22.6	11.3	27.4	24.2	33.9
	pde5	5.8	11.5	9.6	5.8	5.8	17.3	13.5
	pdgfrb	4.1	2.9	4.1	3.7	3.3	10.2	9.4
	pnp	2.1	12.5	6.3	8.3	8.3	14.6	14.6
	ppar_gamma	8.3	0.0	8.3	8.3	8.3	33.3	8.3
	pr	9.5	7.1	7.1	9.5	4.8	2.4	2.4
	rxr_alpha	27.8	33.3	36.1	36.1	36.1	30.6	36.1
	sahh	15.2	7.6	13.6	18.2	16.7	25.8	12.1
	src	3.1	9.3	9.8	3.1	7.2	5.7	8.8
	thrombin	0.0	10.9	8.7	0.0	8.7	8.7	13.0
	tk	4.5	2.3	4.5	4.5	4.5	4.5	4.5
	trypsin	5.6	38.9	38.9	16.7	44.4	16.7	44.4
	vegfr2	6.4	12.8	18.1	5.3	14.9	7.4	12.8
	average	12.5	12.3	16.0	11.5	13.7	14.7	15.9
	St_desv	12.2	11.1	11.6	11.0	11.7	11.1	11.9
	max	43.3	42.3	39.7	42.3	44.4	42.3	44.4
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BS2	**dhfr**	9.0	1.3	6.4	11.5	7.7	10.3	10.3
	vegfr2	1.4	8.3	6.9	6.9	5.6	4.2	5.6
	gr	9.7	5.6	8.3	8.3	8.3	4.2	5.6
	hiv1pr	13.5	3.8	9.6	9.6	9.6	9.6	9.6
	average	8.4	4.8	7.8	9.1	7.8	7.1	7.7
	St_desv	5.1	3.0	1.4	2.0	1.7	3.3	2.5
	max	13.5	8.3	9.6	11.5	9.6	10.3	10.3
	min	1.4	1.3	6.4	6.9	5.6	4.2	5.6
All data sets	**average**	12.1	11.6	15.2	11.3	13.1	14.0	15.2
	St_desv	11.8	10.8	11.3	10.5	11.3	10.8	11.6
	max	43.3	42.3	39.7	42.3	44.4	42.3	44.4
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ROCe 5%								
	Target	**Pha**	**Glide**	**PR**	**RR**	**CR**	**CR tv**	**CR tv**
BS1	**ace**	5.7	3.0	6.1	3.5	3.5	7.0	5.7
	ache	13.2	2.9	11.8	7.2	6.6	7.0	5.8
	ada	3.5	2.6	1.7	3.5	5.2	2.6	3.5
	alr2	0.8	3.8	2.3	0.8	2.3	1.5	3.1
	ampc	8.6	1.0	8.6	2.9	1.9	1.9	1.0
	ar	9.9	12.8	12.2	7.5	11.0	9.6	13.1

	cdk2	1.7	7.4	6.5	1.7	5.7	3.0	6.5
	comt	5.5	7.3	9.1	1.8	5.5	9.1	9.1
	cox1	3.5	3.5	5.2	6.1	7.0	5.2	5.2
	cox2	17.1	12.2	16.9	11.8	12.9	11.2	12.9
	dhfr	0.3	2.9	1.7	5.6	4.8	8.3	6.8
	egfr	3.8	6.0	7.4	4.9	5.8	5.6	6.2
	er_agonist	17.1	13.3	17.5	16.2	15.2	14.0	15.2
	er_antagonist	17.3	13.3	17.3	14.7	14.7	14.7	17.3
	fgfr1	4.2	3.9	5.4	0.8	3.7	0.6	3.9
	fxa	0.9	2.2	2.5	1.3	2.2	2.2	2.5
	gart	5.0	7.5	2.5	2.5	5.0	2.5	5.0
	gpb	12.9	3.1	12.9	11.0	10.2	10.2	10.2
	gr	12.5	5.0	9.4	6.9	6.9	10.6	9.4
	hivrt	7.1	6.5	8.2	5.3	5.9	7.1	7.1
	hivpr	5.0	5.0	10.0	0.0	0.0	5.0	10.0
	hmga	9.6	10.4	10.4	12.8	12.0	12.8	14.4
	hsp90	7.8	1.7	5.2	3.5	2.6	3.5	2.6
	inha	9.5	0.4	8.8	0.7	1.1	3.2	2.5
	mr	15.4	16.9	16.9	18.5	16.9	18.5	16.9
	na	0.8	9.0	7.8	12.7	13.1	15.5	13.9
	p38	2.2	0.7	1.9	3.5	2.8	2.5	2.5
	parp	1.3	14.8	13.5	8.4	13.5	13.5	16.1
	pde5	3.1	8.5	6.9	4.6	6.9	9.2	9.2
	pdgfrb	1.6	1.3	1.6	1.6	1.6	5.2	4.8
	pnp	3.3	5.0	5.8	3.3	5.8	5.8	6.7
	ppar_gamma	3.3	0.0	3.3	6.7	3.3	13.3	6.7
	pr	3.8	4.8	6.7	3.8	6.7	2.9	3.8
	rxr_alpha	13.3	14.4	15.6	15.6	14.4	15.6	15.6
	sahh	9.1	9.7	9.1	10.9	10.3	13.9	13.3
	src	2.7	5.6	6.0	1.9	5.2	4.5	6.2
	thrombin	2.6	5.2	4.3	2.6	4.3	5.2	6.1
	tk	7.3	2.7	3.6	8.2	3.6	4.5	3.6
	trypsin	8.9	15.6	15.6	6.7	17.8	8.9	17.8
	vegfr2	3.0	6.8	7.2	3.0	6.8	3.4	6.0
	average	6.6	6.5	8.1	6.1	7.1	7.5	8.2
	St_desv	5.0	4.7	4.8	4.8	4.7	4.8	4.9
	max	17.3	16.9	17.5	18.5	17.8	18.5	17.8
	min	0.3	0.0	1.6	0.0	0.0	0.6	1.0
BS2	**dhfr**	4.6	4.1	4.6	5.6	5.6	4.6	4.6
	vegfr2	5.0	4.4	5.0	4.4	6.1	2.8	3.9
	gr	6.7	2.8	6.1	5.0	4.4	1.7	2.8
	hiv1pr	10.0	2.3	6.9	6.9	3.8	4.6	4.6
	average	6.6	3.4	5.7	5.5	5.0	3.4	4.0
	St_desv	2.5	1.0	1.1	1.1	1.0	1.5	0.9
	max	10.0	4.4	6.9	6.9	6.1	4.6	4.6
	min	4.6	2.3	4.6	4.4	3.8	1.7	2.8
All data sets	**average**	6.6	6.2	7.9	6.1	6.9	7.1	7.8
	St_desv	4.8	4.6	4.6	4.6	4.6	4.7	4.8
	max	17.3	16.9	17.5	18.5	17.8	18.5	17.8
	min	0.3	0.0	1.6	0.0	0.0	0.6	1.0
AUC								
	Target	**Pha**	**Glide**	**PR**	**RR**	**CR**	**CR tv**	**CR tv**
BS1	**ace**	0.53	0.52	0.58	0.57	0.52	0.63	0.61
	ache	0.77	0.65	0.76	0.72	0.73	0.76	0.77
	ada	0.79	0.59	0.73	0.59	0.67	0.63	0.66
	alr2	0.59	0.71	0.70	0.50	0.64	0.51	0.65
	ampc	0.78	0.52	0.74	0.58	0.56	0.56	0.56
	ar	0.91	0.82	0.93	0.87	0.90	0.90	0.93
	cdk2	0.46	0.66	0.75	0.55	0.65	0.55	0.64

	comt	0.38	0.73	0.72	0.49	0.71	0.69	0.79
	cox1	0.55	0.61	0.59	0.54	0.58	0.58	0.60
	cox2	0.96	0.87	0.97	0.84	0.88	0.82	0.88
	dhfr	0.52	0.72	0.65	0.75	0.74	0.78	0.77
	egfr	0.53	0.63	0.67	0.57	0.60	0.61	0.62
	er_agonist	0.97	0.95	0.97	0.95	0.95	0.92	0.95
	er_antagonist	0.96	0.73	0.93	0.85	0.84	0.88	0.87
	fgfr1	0.47	0.59	0.58	0.48	0.54	0.47	0.55
	fxa	0.33	0.52	0.44	0.44	0.50	0.48	0.52
	gart	0.59	0.86	0.83	0.77	0.83	0.73	0.80
	gpb	0.87	0.64	0.87	0.78	0.76	0.80	0.78
	gr	0.89	0.67	0.84	0.83	0.79	0.88	0.85
	hivrt	0.70	0.67	0.79	0.75	0.72	0.77	0.73
	hivpr	0.28	0.42	0.53	0.19	0.36	0.42	0.56
	hmga	0.89	0.82	0.90	0.89	0.93	0.92	0.93
	hsp90	0.83	0.61	0.80	0.59	0.67	0.61	0.65
	inha	0.67	0.52	0.64	0.45	0.49	0.54	0.59
	mr	0.97	0.93	0.96	0.98	0.98	0.98	0.98
	na	0.57	0.82	0.80	0.91	0.90	0.94	0.93
	p38	0.51	0.54	0.51	0.49	0.52	0.44	0.49
	parp	0.47	0.95	0.90	0.87	0.94	0.92	0.96
	pde5	0.50	0.82	0.74	0.73	0.78	0.82	0.82
	pdgfrb	0.28	0.37	0.33	0.36	0.35	0.59	0.54
	pnp	0.77	0.70	0.79	0.60	0.70	0.64	0.70
	ppar_gamma	0.49	0.60	0.48	0.62	0.56	0.79	0.71
	pr	0.59	0.69	0.69	0.35	0.69	0.44	0.65
	rxr_alpha	0.87	0.92	0.95	0.94	0.95	0.93	0.95
	sahh	0.85	0.86	0.84	0.87	0.88	0.92	0.92
	src	0.35	0.66	0.65	0.59	0.65	0.67	0.67
	thrombin	0.62	0.80	0.78	0.60	0.78	0.74	0.80
	tk	0.77	0.76	0.79	0.84	0.82	0.86	0.83
	trypsin	0.85	0.92	0.97	0.85	0.92	0.85	0.94
	vegfr2	0.58	0.74	0.72	0.67	0.77	0.68	0.75
	average	0.66	0.70	0.75	0.67	0.72	0.72	0.75
	St_desv	0.21	0.15	0.16	0.19	0.17	0.17	0.14
	max	0.97	0.95	0.97	0.98	0.98	0.98	0.98
	min	0.28	0.37	0.33	0.19	0.35	0.42	0.49
BS2	**dhfr**	0.60	0.74	0.74	0.62	0.72	0.66	0.73
	vegfr2	0.62	0.69	0.73	0.65	0.72	0.64	0.71
	gr	0.86	0.49	0.82	0.63	0.60	0.57	0.54
	hiv1pr	0.81	0.32	0.75	0.66	0.61	0.67	0.62
	average	0.72	0.56	0.76	0.64	0.66	0.64	0.65
	St_desv	0.13	0.19	0.04	0.02	0.06	0.05	0.09
	max	0.86	0.74	0.82	0.66	0.72	0.67	0.73
	min	0.60	0.32	0.73	0.62	0.60	0.57	0.54
All data sets	**average**	0.66	0.69	0.75	0.67	0.71	0.71	0.74
	St_desv	0.20	0.15	0.15	0.18	0.16	0.16	0.14
	max	0.97	0.95	0.97	0.98	0.98	0.98	0.98
	min	0.28	0.32	0.33	0.19	0.35	0.42	0.49

Table S3. ROCE and AUC values for all data sets included in BS1 and BS2 using Glide in SP mode. Parallel ranking (PR), rescoring ranking (RR) and consensus ranking (CR). Tv: Tversky coefficient.

ROCe 0.5%								
	Target	**Pha**	**Glide**	**PR**	**RR**	**CR**	**CR tv**	**CR tv**
BS1	**ace**	17.4	8.7	17.4	39.1	21.7	34.8	26.1
	ache	57.7	0.0	26.8	72.2	28.9	51.5	22.7
	ada	0.0	8.7	0.0	0.0	0.0	26.1	17.4
	alr2	7.7	7.7	7.7	7.7	7.7	15.4	15.4
	ampc	76.2	0.0	38.1	9.5	9.5	9.5	9.5

	ar	53.7	14.9	35.8	41.8	35.8	38.8	35.8
	cdk2	0.0	34.8	17.4	8.7	21.7	21.7	26.1
	comt	54.5	0.0	54.5	0.0	0.0	54.5	36.4
	cox1	0.0	17.4	8.7	17.4	17.4	8.7	17.4
	cox2	145.5	16.3	74.6	152.2	71.8	136.8	68.9
	dhfr	0.0	6.4	4.3	37.2	16.0	50.0	33.0
	egfr	15.9	5.5	17.0	32.3	23.6	78.9	49.3
	er_agonist	92.1	44.4	76.2	98.4	73.0	98.4	88.9
	er_antagonist	40.0	0.0	13.3	26.7	13.3	13.3	13.3
	fgfr1	8.5	50.7	39.4	2.8	31.0	0.0	31.0
	fxa	6.3	12.5	12.5	0.0	6.3	3.1	9.4
	gart	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	gpb	47.1	0.0	31.4	94.1	43.1	82.4	43.1
	gr	68.8	43.8	50.0	50.0	50.0	81.3	75.0
	hivrt	41.2	11.8	35.3	41.2	41.2	23.5	23.5
	hivpr	50.0	100.0	100.0	0.0	0.0	50.0	50.0
	hmga	40.0	80.0	80.0	96.0	96.0	112.0	120.0
	hsp90	34.8	0.0	34.8	34.8	34.8	34.8	34.8
	inha	80.7	49.1	80.7	73.7	70.2	91.2	84.2
	mr	107.7	61.5	92.3	92.3	92.3	107.7	92.3
	na	4.1	77.6	32.7	24.5	44.9	65.3	106.1
	p38	5.8	0.0	4.4	7.3	5.8	5.8	2.9
	parp	6.5	38.7	32.3	12.9	25.8	45.2	77.4
	pde5	23.1	30.8	38.5	7.7	30.8	23.1	46.2
	pdgfrb	13.1	14.8	14.8	14.8	14.8	21.3	21.3
	pnp	8.3	16.7	16.7	16.7	16.7	41.7	33.3
	ppar_gamma	33.3	33.3	66.7	0.0	0.0	100.0	66.7
	pr	28.6	9.5	28.6	9.5	19.0	19.0	19.0
	rxr_alpha	88.9	155.6	100.0	88.9	100.0	55.6	100.0
	sahh	24.2	30.3	42.4	30.3	48.5	6.1	24.2
	src	4.1	22.7	20.6	0.0	16.5	35.1	39.2
	thrombin	0.0	17.4	8.7	0.0	8.7	0.0	8.7
	tk	9.1	9.1	9.1	72.7	54.5	72.7	54.5
	trypsin	22.2	66.7	88.9	0.0	66.7	0.0	66.7
	vegfr2	21.3	34.0	38.3	4.3	29.8	8.5	34.0
	average	33.5	28.3	37.3	32.9	32.2	43.1	43.1
	St_desv	34.6	32.6	29.4	37.4	27.7	36.6	30.7
	max	145.5	155.6	100.0	152.2	100.0	136.8	120.0
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BS2	**dhfr**	25.6	25.6	35.9	76.9	76.9	76.9	71.8
	vegfr2	0.0	55.6	16.7	22.2	27.8	22.2	38.9
	gr	22.2	11.1	16.7	16.7	22.2	16.7	22.2
	hiv1pr	23.1	46.2	53.8	38.5	38.5	15.4	23.1
	average	17.7	34.6	30.8	38.6	41.3	32.8	39.0
	St_desv	11.9	20.0	17.9	27.2	24.7	29.6	23.2
	max	25.6	55.6	53.8	76.9	76.9	76.9	71.8
	min	0.0	11.1	16.7	16.7	22.2	15.4	22.2
	average	32.0	28.9	36.7	33.5	33.0	42.7	42.0
	St_desv	33.4	31.5	28.4	36.4	27.3	35.4	29.9
	max	145.5	155.6	100.0	152.2	100.0	136.8	120.0
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ROCe 1%								
	Target	**Pha**	**Glide**	**PR**	**RR**	**CR**	**CR tv**	**CR tv**
BS1	**ace**	15.2	4.3	13.0	19.6	19.6	17.4	19.6
	ache	47.4	1.0	23.7	38.1	28.9	34.0	21.6
	ada	0.0	4.3	4.3	21.7	4.3	13.0	17.4
	alr2	3.8	11.5	3.8	3.8	3.8	7.7	7.7
	ampc	38.1	4.8	28.6	4.8	4.8	4.8	4.8
	ar	31.3	13.4	25.4	26.9	19.4	28.4	22.4

	cdk2	2.2	19.6	10.9	4.3	13.0	10.9	19.6
	comt	27.3	0.0	27.3	0.0	0.0	36.4	27.3
	cox1	4.3	8.7	8.7	17.4	17.4	4.3	13.0
	cox2	76.6	15.3	59.3	77.0	64.1	74.2	56.9
	dhfr	0.5	4.8	3.2	25.0	16.0	32.4	21.8
	egfr	9.0	7.1	10.7	25.8	16.7	44.9	38.9
	er_agonist	74.6	38.1	61.9	63.5	57.1	60.3	55.6
	er_antagonist	60.0	40.0	26.7	53.3	53.3	66.7	40.0
	fgfr1	7.0	26.8	23.9	1.4	21.1	1.4	21.1
	fxa	3.1	7.8	9.4	1.6	4.7	1.6	6.3
	gart	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	gpb	45.1	0.0	21.6	52.9	35.3	56.9	27.5
	gr	37.5	21.9	31.3	28.1	25.0	43.8	40.6
	hivrt	23.5	5.9	20.6	20.6	20.6	17.6	11.8
	hivpr	25.0	50.0	50.0	0.0	0.0	25.0	25.0
	hmga	24.0	40.0	40.0	56.0	52.0	56.0	64.0
	hsp90	21.7	0.0	17.4	17.4	17.4	17.4	17.4
	inha	40.4	24.6	40.4	38.6	35.1	47.4	43.9
	mr	53.8	38.5	61.5	53.8	53.8	53.8	53.8
	na	2.0	40.8	30.6	34.7	46.9	53.1	59.2
	p38	3.6	0.0	2.9	3.6	3.6	5.8	2.9
	parp	3.2	41.9	19.4	12.9	25.8	38.7	45.2
	pde5	11.5	26.9	19.2	3.8	15.4	15.4	23.1
	pdgfrb	8.2	7.4	7.4	7.4	7.4	10.7	10.7
	pnp	4.2	8.3	8.3	12.5	8.3	37.5	16.7
	ppar_gamma	16.7	16.7	33.3	0.0	0.0	50.0	33.3
	pr	19.0	9.5	19.0	4.8	9.5	14.3	14.3
	rxr_alpha	50.0	83.3	77.8	77.8	72.2	27.8	66.7
	sahh	24.2	18.2	21.2	21.2	24.2	9.1	21.2
	src	5.2	21.6	10.3	1.0	8.2	21.6	29.9
	thrombin	0.0	8.7	8.7	0.0	8.7	0.0	8.7
	tk	4.5	4.5	9.1	45.5	40.9	40.9	40.9
	trypsin	11.1	55.6	44.4	0.0	33.3	11.1	44.4
	vegfr2	12.8	19.1	23.4	4.3	19.1	6.4	19.1
	average	21.2	18.8	24.0	22.0	22.7	27.5	27.9
	St_desv	21.1	18.6	18.5	22.7	19.4	20.9	17.9
	max	76.6	83.3	77.8	77.8	72.2	74.2	66.7
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BS2	**dhfr**	12.8	15.4	25.6	38.5	48.7	38.5	46.2
	vegfr2	2.8	30.6	16.7	11.1	27.8	11.1	27.8
	gr	11.1	5.6	11.1	13.9	11.1	8.3	11.1
	hiv1pr	15.4	30.8	34.6	26.9	34.6	7.7	26.9
	average	10.5	20.6	22.0	22.6	30.6	16.4	28.0
	St_desv	5.5	12.3	10.3	12.6	15.6	14.8	14.3
	max	15.4	30.8	34.6	38.5	48.7	38.5	46.2
	min	2.8	5.6	11.1	11.1	11.1	7.7	11.1
	average	20.2	18.9	23.8	22.1	23.4	27.5	27.3
	St_desv	20.4	18.0	17.8	21.9	19.0	20.3	17.3
	max	76.6	83.3	77.8	77.8	72.2	74.2	66.7
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ROCe 2%								
	Target	**Pha**	**Glide**	**PR**	**RR**	**CR**	**RR tv**	**CR tv**
BS1	**ace**	10.9	3.3	9.8	12.0	10.9	9.8	10.9
	ache	28.9	0.5	15.5	21.6	18.6	20.6	14.4
	ada	0.0	6.5	2.2	10.9	8.7	15.2	8.7
	alr2	1.9	9.6	3.8	1.9	3.8	5.8	7.7
	ampc	21.4	2.4	19.0	4.8	4.8	2.4	4.8
	ar	18.7	10.4	17.2	16.4	14.2	19.4	14.9
	cdk2	2.2	18.5	10.9	2.2	10.9	8.7	14.1
	comt	13.6	0.0	13.6	0.0	0.0	18.2	13.6

	cox1	2.2	4.3	6.5	10.9	10.9	6.5	6.5
	cox2	39.5	14.4	38.5	39.2	39.2	38.8	36.6
	dhfr	0.5	5.3	2.4	16.0	13.0	21.5	16.8
	egfr	7.0	5.9	8.1	16.2	14.8	24.4	24.8
	er_agonist	40.5	26.2	38.9	38.9	34.9	34.9	34.9
	er_antagonist	43.3	30.0	43.3	30.0	30.0	33.3	36.7
	fgfr1	6.3	14.8	15.5	2.1	13.4	0.7	13.4
	fxa	1.6	4.7	5.5	1.6	3.9	2.3	3.9
	gart	0.0	12.5	0.0	0.0	0.0	0.0	0.0
	gpb	27.5	0.0	15.7	31.4	23.5	32.4	21.6
	gr	23.4	10.9	18.8	17.2	14.1	23.4	21.9
	hivrt	14.7	11.8	11.8	10.3	10.3	11.8	7.4
	hivpr	12.5	25.0	25.0	0.0	0.0	12.5	12.5
	hmga	18.0	22.0	22.0	30.0	28.0	28.0	32.0
	hsp90	13.0	0.0	10.9	8.7	8.7	8.7	8.7
	inha	21.1	13.2	20.2	19.3	18.4	24.6	22.8
	mr	30.8	23.1	30.8	30.8	30.8	34.6	30.8
	na	1.0	23.5	19.4	24.5	30.6	35.7	36.7
	p38	4.0	0.4	1.8	4.0	1.8	5.1	2.9
	parp	1.6	29.0	14.5	9.7	19.4	22.6	33.9
	pde5	5.8	15.4	15.4	3.8	13.5	17.3	15.4
	pdgfrb	4.1	3.7	4.1	3.7	3.7	5.7	5.3
	pnp	2.1	6.3	4.2	12.5	6.3	22.9	14.6
	ppar_gamma	8.3	8.3	16.7	0.0	0.0	25.0	16.7
	pr	9.5	4.8	9.5	2.4	4.8	7.1	9.5
	rxr_alpha	27.8	41.7	41.7	38.9	41.7	19.4	36.1
	sahh	15.2	9.1	13.6	18.2	18.2	22.7	18.2
	src	3.1	16.0	11.3	4.1	10.3	18.0	20.1
	thrombin	0.0	6.5	4.3	0.0	4.3	6.5	4.3
	tk	4.5	2.3	4.5	22.7	22.7	22.7	20.5
	trypsin	5.6	38.9	33.3	0.0	22.2	5.6	33.3
	vegfr2	6.4	12.8	13.8	4.3	12.8	6.4	13.8
	average	12.5	12.3	15.3	13.0	14.4	17.0	17.5
	St_desv	12.2	10.7	11.4	12.2	11.1	10.9	10.9
	max	43.3	41.7	43.3	39.2	41.7	38.8	36.7
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BS2	**dhfr**	9.0	10.3	14.1	19.2	26.9	20.5	26.9
	vegfr2	1.4	18.1	16.7	6.9	18.1	5.6	16.7
	gr	9.7	2.8	5.6	9.7	8.3	5.6	5.6
	hiv1pr	13.5	19.2	21.2	17.3	19.2	13.5	19.2
	average	8.4	12.6	14.4	13.3	18.1	11.3	17.1
	St_desv	5.1	7.7	6.6	5.9	7.6	7.2	8.8
	max	13.5	19.2	21.2	19.2	26.9	20.5	26.9
	min	1.4	2.8	5.6	6.9	8.3	5.6	5.6
	average	12.1	12.4	15.3	13.0	14.8	17.0	17.3
	St_desv	11.8	10.3	11.0	11.7	10.9	10.7	10.6
	max	43.3	41.7	43.3	39.2	41.7	38.8	36.7
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ROCe 5%								
	Target	**Pha**	**Glide**	**PR**	**RR**	**CR**	**RR tv**	**CR tv**
BS 1	**ace**	5.7	1.7	6.1	5.2	5.7	7.4	6.1
	ache	13.2	0.8	11.1	10.1	8.9	9.1	8.7
	ada	3.5	2.6	2.6	5.2	7.8	7.0	8.7
	alr2	0.8	5.4	3.8	1.5	4.6	3.1	5.4
	ampc	8.6	1.0	8.6	1.9	2.9	1.0	1.9
	ar	9.9	10.1	9.9	7.2	8.7	8.1	9.9
	cdk2	1.7	10.4	8.7	2.2	8.7	3.9	8.7
	comt	5.5	3.6	7.3	0.0	1.8	7.3	7.3
	cox1	3.5	2.6	4.3	6.1	7.0	3.5	4.3

	cox2	17.1	9.8	16.7	15.9	16.4	15.9	16.2
	dhfr	0.3	6.3	2.6	8.7	8.0	12.4	11.2
	egfr	3.8	4.8	5.3	8.7	8.8	10.6	11.6
	er_agonist	17.1	13.3	16.8	16.2	16.5	16.5	15.6
	er_antagonist	17.3	13.3	17.3	12.0	13.3	14.7	16.0
	fgfr1	4.2	6.5	7.3	2.0	6.2	1.1	5.9
	fxa	0.9	2.2	2.5	1.3	2.2	2.2	2.5
	gart	5.0	7.5	5.0	0.0	5.0	0.0	5.0
	gpb	12.9	0.8	11.0	14.5	12.5	15.7	12.5
	gr	12.5	5.0	9.4	7.5	7.5	11.3	9.4
	hivrt	7.1	7.1	8.8	5.3	7.1	7.1	8.8
	hivpr	5.0	10.0	10.0	0.0	0.0	5.0	5.0
	hmga	9.6	9.6	10.4	14.4	12.8	14.4	13.6
	hsp90	7.8	2.6	7.0	3.5	4.3	3.5	4.3
	inha	9.5	6.3	8.8	8.8	7.7	10.9	9.8
	mr	15.4	15.4	15.4	15.4	15.4	15.4	15.4
	na	0.8	11.4	9.4	13.5	14.3	17.1	17.1
	p38	2.2	1.2	1.9	3.4	1.8	3.1	2.3
	parp	1.3	16.1	11.6	9.7	13.5	15.5	16.8
	pde5	3.1	8.5	6.9	3.8	6.9	10.0	10.0
	pdgfrb	1.6	1.6	1.6	1.6	1.5	3.3	2.6
	pnp	3.3	4.2	4.2	11.7	6.7	15.0	10.8
	ppar_gamma	3.3	3.3	6.7	0.0	3.3	10.0	10.0
	pr	3.8	2.9	3.8	1.9	1.9	3.8	3.8
	rxr_alpha	13.3	16.7	17.8	16.7	17.8	12.2	17.8
	sahh	9.1	10.3	8.5	10.9	9.1	13.9	13.3
	src	2.7	8.5	7.4	4.3	7.6	8.7	11.1
	thrombin	2.6	4.3	2.6	2.6	2.6	5.2	5.2
	tk	7.3	3.6	4.5	11.8	10.0	10.9	10.0
	trypsin	8.9	15.6	15.6	0.0	15.6	2.2	15.6
	vegfr2	3.0	6.4	7.7	3.0	7.2	3.8	8.1
	average	6.6	6.8	8.2	6.7	8.0	8.5	9.5
	St_desv	5.0	4.7	4.5	5.3	4.8	5.2	4.6
	max	17.3	16.7	17.8	16.7	17.8	17.1	17.8
	min	0.3	0.8	1.6	0.0	0.0	0.0	1.9
BS2	**dhfr**	4.6	7.7	8.2	8.2	11.8	9.2	12.8
	vegfr2	5.0	7.8	7.8	5.6	8.9	4.4	8.3
	gr	6.7	2.2	4.4	6.1	4.4	2.8	3.3
	hiv1pr	10.0	10.8	13.8	8.5	12.3	8.5	11.5
	average	6.6	7.1	8.6	7.1	9.4	6.2	9.0
	St_desv	2.5	3.6	3.9	1.5	3.6	3.1	4.2
	max	10.0	10.8	13.8	8.5	12.3	9.2	12.8
	min	4.6	2.2	4.4	5.6	4.4	2.8	3.3
	average	6.6	6.9	8.2	6.7	8.1	8.6	9.4
	St_desv	4.8	4.5	4.4	5.1	4.6	5.0	4.5
	max	17.3	16.7	17.8	16.7	17.8	17.1	17.8
	min	0.3	0.8	1.6	0.0	0.0	0.0	1.9
AUC								
	Target	**Pha**	**Glide**	**PR**	**RR**	**CR**	**RR tv**	**CR tv**
BS1	**ace**	0.53	0.45	0.49	0.61	0.57	0.70	0.69
	ache	0.77	0.62	0.76	0.72	0.72	0.73	0.72
	ada	0.79	0.61	0.75	0.77	0.78	0.78	0.79
	alr2	0.59	0.80	0.75	0.56	0.74	0.55	0.76
	ampc	0.78	0.39	0.69	0.58	0.53	0.58	0.54
	ar	0.91	0.85	0.92	0.85	0.90	0.90	0.93
	cdk2	0.46	0.79	0.85	0.66	0.79	0.62	0.77
	comt	0.38	0.78	0.69	0.53	0.68	0.70	0.83
	cox1	0.55	0.66	0.64	0.57	0.63	0.61	0.63
	cox2	0.96	0.89	0.96	0.91	0.93	0.90	0.93
	dhfr	0.52	0.84	0.74	0.86	0.85	0.91	0.91

	egfr	0.53	0.74	0.70	0.70	0.74	0.70	0.76
	er_agonist	0.97	0.93	0.96	0.95	0.95	0.95	0.94
	er_antagonist	0.96	0.91	0.96	0.93	0.94	0.96	0.97
	fgfr1	0.47	0.66	0.64	0.53	0.63	0.54	0.65
	fxa	0.33	0.52	0.44	0.44	0.49	0.47	0.51
	gart	0.59	0.88	0.81	0.73	0.81	0.72	0.81
	gpb	0.87	0.64	0.89	0.90	0.89	0.91	0.90
	gr	0.89	0.67	0.85	0.82	0.78	0.87	0.84
	hivrt	0.70	0.74	0.81	0.68	0.71	0.69	0.71
	hivpr	0.28	0.69	0.75	0.22	0.53	0.44	0.56
	hmga	0.89	0.86	0.90	0.92	0.94	0.96	0.94
	hsp90	0.83	0.76	0.84	0.54	0.77	0.59	0.75
	inha	0.67	0.57	0.61	0.71	0.65	0.82	0.80
	mr	0.97	0.90	0.95	0.97	0.96	0.98	0.96
	na	0.57	0.91	0.85	0.94	0.95	0.95	0.96
	p38	0.51	0.50	0.51	0.65	0.64	0.57	0.61
	parp	0.47	0.95	0.91	0.93	0.96	0.97	0.98
	pde5	0.50	0.82	0.78	0.69	0.80	0.81	0.85
	pdgfrb	0.28	0.31	0.27	0.29	0.28	0.54	0.47
	pnp	0.77	0.87	0.84	0.81	0.87	0.88	0.90
	ppar_gamma	0.49	0.40	0.54	0.59	0.52	0.84	0.79
	pr	0.59	0.67	0.64	0.37	0.58	0.59	0.68
	rxr_alpha	0.87	0.96	0.98	0.94	0.97	0.91	0.96
	sahh	0.85	0.92	0.91	0.89	0.92	0.95	0.94
	src	0.35	0.81	0.79	0.72	0.80	0.78	0.82
	thrombin	0.62	0.74	0.76	0.53	0.75	0.65	0.77
	tk	0.77	0.87	0.88	0.92	0.94	0.94	0.94
	trypsin	0.85	0.89	0.97	0.85	0.92	0.84	0.94
	vegfr2	0.58	0.82	0.80	0.57	0.78	0.59	0.79
	average	0.66	0.74	0.77	0.71	0.76	0.76	0.80
	St_desv	0.21	0.17	0.16	0.19	0.16	0.16	0.14
	max	0.97	0.96	0.98	0.97	0.97	0.98	0.98
	min	0.28	0.31	0.27	0.22	0.28	0.44	0.47
BS2	**dhfr**	0.60	0.77	0.81	0.70	0.84	0.76	0.86
	vegfr2	0.62	0.74	0.77	0.68	0.71	0.68	0.71
	gr	0.86	0.64	0.80	0.71	0.70	0.65	0.64
	hiv1pr	0.81	0.76	0.86	0.83	0.87	0.84	0.87
	average	0.72	0.73	0.81	0.73	0.78	0.73	0.77
	St_desv	0.13	0.06	0.04	0.07	0.09	0.08	0.11
	max	0.86	0.77	0.86	0.83	0.87	0.84	0.87
	min	0.60	0.64	0.77	0.68	0.70	0.65	0.64
	average	0.66	0.7	0.8	0.7	0.8	0.8	0.8
	St_desv	0.20	0.2	0.2	0.2	0.2	0.2	0.1
	max	0.97	1.0	1.0	1.0	1.0	1.0	1.0
	min	0.28	0.3	0.3	0.2	0.3	0.4	0.5

Table S4. AUC and ROCe metrics for PharmScreen, Glide HTVS, and Combined methods derived from Glide HTVS. Parallel ranking (PR), rescoring ranking (RR) and consensus ranking (CR). Tv: Tversky coefficient.

		Pha	**Glide**	**PR**	**RR**	**CR**	**RR tv**	**CR tv**
BS1	ROCE 0.5	33.5	27.6	37.4	31.2	32.9	42.4	**42.9**
	ROCE 1	21.2	19.3	25.1	19.1	21.6	25.7	**25.8**
	ROCE 2	12.5	12.3	16.0	11.5	13.7	14.7	**15.9**
	ROCE 5	6.6	6.5	8.1	6.1	7.1	7.52	**8.2**
	AUC	0.7	0.7	0.8	0.7	0.7	0.72	0.7
BS2	ROCE 0.5	17.7	7.5	17.8	20.9	19.7	19.1	14.6
	ROCE 1	10.5	6.8	11.2	13.5	11.9	12.4	9.2

	ROCE 2	8.4	4.8	7.8	9.1	7.8	7.1	7.7
	ROCE 5	6.6	3.4	5.7	5.5	5.0	3.4	4.0
	AUC	0.7	0.6	**0.8**	0.6	0.7	0.6	0.6

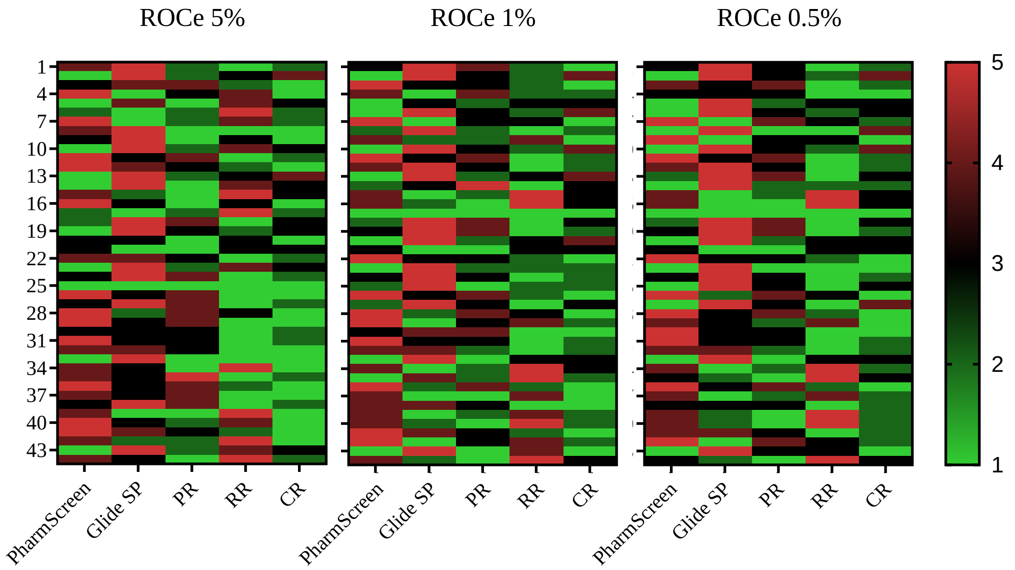


Figure S1. Heatmap of the hierarchical position of all methods among the others for ROCe 0.5%, 2% and 5%. The color scale is indicative of the position, being the first green and the fifth red. Parallel ranking (PR), rescoring ranking (RR) and consensus ranking (CR).

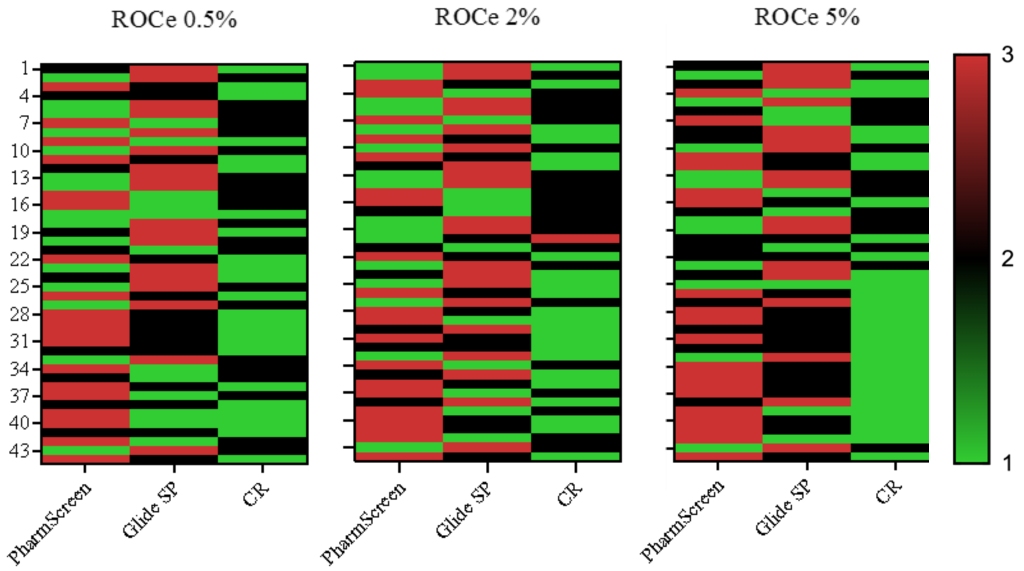


Figure S2. Heatmap of the hierarchical position of Consensus Ranking (CR) using Tversky coefficient for ROCe 0.5%, 2% and 5%. The method ranked first is shown in green, second in black and third in red.

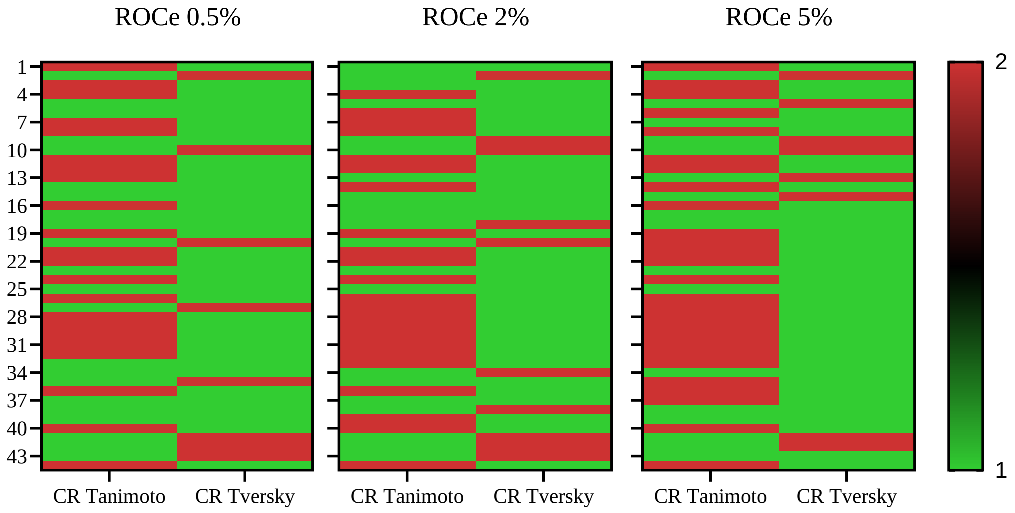


Figure S3. Heatmap of the hierarchical position of Consensus Ranking (CR) using Tversky coefficient for ROCe 0.5%, 2%, 5%. The method ranked first is shown in green and the second in red.

Table S6. awROCE values for all data sets included in BS1 and BS2 using PharmScreen (Pha), Glide in SP mode and Consensus Ranking using tv (CR tv).

		awROCEe 0.5%			**awROCe 1%**		
	Target	**Pha**	**Glide**	**CR tv**	**Pha**	**Glide**	**CR tv**
BS1	**ace**	33.7	10.5	16.9	28.4	5.3	14.7
	ache	46.4	0.0	8.4	40.5	0.3	12.6
	ada	0.0	3.1	8.3	0.0	1.6	7.8
	alr2	7.1	7.1	11.9	3.6	12.5	6.0
	ampc	19.0	0.0	33.3	9.5	1.2	16.7
	ar	11.3	3.1	7.3	6.6	3.1	9.2
	cdk2	0.0	31.4	28.1	1.6	16.8	16.4
	comt	100.0	0.0	25.0	50.0	0.0	18.8
	cox1	0.0	18.2	15.2	9.1	9.1	12.1
	cox2	85.9	12.6	26.1	54.3	10.1	38.4
	dhfr	0.0	14.6	34.6	0.1	8.3	22.3
	egfr	12.4	13.7	33.2	8.3	13.4	25.6
	er_agonist	60.6	43.0	95.2	52.7	40.1	60.8
	er_antagonist	22.5	0.0	12.5	43.8	40.0	40.0
	fgfr1	20.0	66.3	47.9	12.1	34.2	27.7
	fxa	21.1	6.3	5.9	10.5	8.4	3.1
	gart	0.0	0.0	0.0	0.0	0.0	0.0
	gpb	44.1	0.0	43.5	25.3	0.0	32.4
	gr	49.4	19.4	51.9	25.9	9.7	27.2
	hivrt	23.5	13.7	17.6	17.6	6.9	8.8
	hivpr	33.3	133.3	66.7	16.7	66.7	33.3
	hmga	27.8	100.0	140.4	15.8	50.0	72.1
	hsp90	25.0	0.0	25.0	15.6	0.0	12.5
	inha	50.7	23.2	67.1	25.4	11.6	37.9
	mr	58.3	33.3	50.0	29.2	20.8	29.2
	na	0.7	151.7	142.9	0.3	76.2	79.3
	p38	10.9	0.0	10.5	5.5	0.0	5.6
	parp	28.6	20.5	87.6	14.3	25.3	46.0
	pde5	15.9	20.5	25.0	8.0	17.0	12.5
	pdgfrb	26.0	27.3	38.5	18.2	13.6	19.2
	pnp	7.1	23.8	73.8	3.6	11.9	36.9
	ppar_gamma	33.3	33.3	66.7	16.7	16.7	33.3

	pr	50.0	50.0	62.5	50.0	33.3	32.8
	rxr_alpha	35.6	62.2	40.0	20.0	64.4	26.7
	sahh	12.9	16.1	12.9	12.9	9.7	11.3
	src	11.9	20.6	28.9	7.8	19.9	23.8
	thrombin	0.0	30.8	15.4	0.0	15.4	15.4
	tk	3.2	28.6	44.4	1.6	14.3	27.0
	trypsin	28.6	85.7	85.7	14.3	61.9	57.1
	vegfr2	32.3	48.4	48.4	19.4	27.4	27.4
	average	26.2	29.3	41.4	17.4	19.4	26.0
	St_desv	23.3	35.4	33.9	15.7	20.3	18.1
	max	100.0	151.7	142.9	54.3	76.2	79.3
	min	0.0	0.0	0.0	0.0	0.0	0.0
		awROCe 2%			**awROCe 5%**		
	Target	**Pha**	**Glide**	**CR tv**	**Pha**	**Glide**	**CR tv**
BS1	**ace**	18.0	2.9	7.6	9.5	1.4	4.6
	ache	21.8	0.1	10.5	9.6	0.4	5.3
	ada	0.0	2.3	3.9	2.8	0.9	5.9
	alr2	1.8	10.7	7.4	0.7	5.1	5.2
	ampc	5.4	0.6	8.9	2.1	0.2	3.6
	ar	3.9	2.5	12.7	3.2	3.7	10.0
	cdk2	2.4	17.8	11.1	1.4	10.4	7.8
	comt	25.0	0.0	9.4	10.0	4.6	5.0
	cox1	4.5	4.5	6.1	3.9	3.6	4.8
	cox2	28.5	10.4	27.6	14.2	6.7	12.6
	dhfr	0.1	10.0	14.0	0.0	8.7	11.6
	egfr	7.3	8.9	18.5	3.7	6.1	8.9
	er_agonist	28.4	33.0	37.7	13.7	14.4	15.9
	er_antagonist	43.8	28.8	37.1	17.5	12.0	15.7
	fgfr1	7.3	21.4	17.1	7.1	10.3	8.6
	fxa	5.3	4.3	4.2	2.6	1.7	3.8
	gart	0.0	12.5	0.0	5.0	6.0	5.0
	gpb	13.4	0.0	30.4	5.6	0.6	14.1
	gr	14.8	4.9	15.4	9.6	2.2	6.4
	hivrt	9.7	10.8	4.9	5.6	6.7	6.9
	hivpr	8.3	33.3	16.7	3.3	13.3	6.7
	hmga	22.8	31.3	36.1	10.6	12.9	16.9
	hsp90	9.1	0.0	6.3	5.3	3.8	4.2
	inha	14.9	6.5	21.1	8.0	3.9	9.0
	mr	16.7	12.5	16.7	8.3	8.3	8.3
	na	0.2	38.6	47.8	0.1	15.8	19.5
	p38	3.3	0.0	3.2	1.4	1.2	2.6
	parp	7.1	15.7	37.3	3.1	10.6	16.1
	pde5	4.0	10.8	9.1	2.5	7.0	8.2
	pdgfrb	9.1	6.8	9.6	3.6	2.8	4.1
	pnp	1.8	7.0	23.0	5.7	3.9	13.1
	ppar_gamma	8.3	8.3	16.7	3.3	3.3	10.0
	pr	25.0	16.7	17.2	10.0	7.0	6.9
	rxr_alpha	11.1	32.2	14.4	14.4	12.9	13.3
	sahh	8.1	4.8	9.7	4.8	5.5	7.1
	src	4.1	12.9	16.5	2.9	8.8	9.6
	thrombin	0.0	8.7	7.7	3.5	5.5	6.3
	tk	1.6	7.1	13.5	2.5	4.3	6.3
	trypsin	7.1	45.2	38.1	9.5	18.1	18.1
	vegfr2	9.7	16.9	18.5	4.2	7.8	9.8
	average	10.3	12.6	16.6	5.9	6.6	8.9
	St_desv	9.8	11.8	11.6	4.3	4.6	4.5
	max	43.8	45.2	47.8	17.5	18.1	19.5
	min	0.0	0.0	0.0	0.0	0.2	2.6

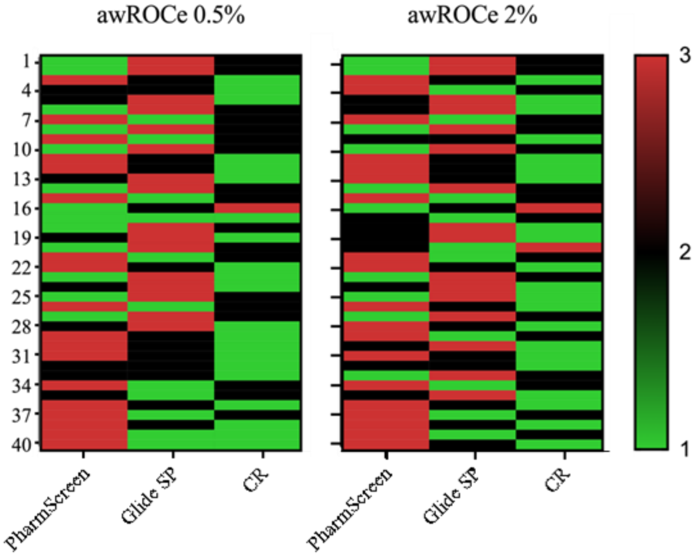


Figure S4. Heatmap of the hierarchical position of Consensus Ranking using Tanimoto and Tversky coefficient for awROCe 0.5% and 2%. The method ranked first is shown in green, second in black and third in red.

3.4 PATENT: *“Calculating molecular similarity”*

Inventors: VÁZQUEZ, Javier; HERRERO, Enric; GIBERT, Enric; LUQUE, Javier.

Applicant: PHARMACELERA, S.L. [ES/ES]; C. Esteve Pila 22, 1r1a, 08173 SANT CUGAT DEL VALLÈS (ES).

Abstract

Methods and tools for measuring degrees of similarity between molecules using information of molecules hydrophobicity is proposed. In example, sets of field values for each of the molecules are calculated, each field values representing hydrophobicity of the respective molecules. The calculated sets of field values are combined to generate similarity index.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
05 July 2018 (05.07.2018)

WIPO | PCT

(10) International Publication Number
WO 2018/121866 A1

(51) International Patent Classification:
G06F 19/00 (2018.01)

(21) International Application Number:
PCT/EP2016/082850

(22) International Filing Date:
29 December 2016 (29.12.2016)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **PHARMACELERA, S.L.** [ES/ES]; C. Esteve Pila 11, 1r1a, 08173 SANT CUGAT DEL VALLÈS (ES).

(72) Inventors: **VÁZQUEZ, Javier**; C. Poeta Joan Maragall 1, 3r 1a, 08160 MONTMELÓ (ES). **HERRERO, Enric**; C. Cucut, 6, 08440 CARDEDEU (ES). **GIBERT, Enric**; C. Esteve Pila 11, 1r 1a, 08173 SANT CUGAT DEL VALLÈS (ES). **LUQUE, Javier**; C. Xaloc 4, 9ºB, 08917 BADALONA (ES).

(74) Agent: **ZBM PATENTS - Zea, Barlocci & Markvardsen**; Pl. Catalunya, 1 2nd floor, 08002 BARCELONA (ES).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:
— *with international search report (Art. 21(3))*

WO 2018/121866 A1

(54) **Title:** CALCULATING MOLECULAR SIMILARITY

(57) **Abstract:** Methods and tools for measuring degrees of similarity between molecules using information of molecule hydrophobicity are proposed. In example methods, sets of field values for each of the molecules are calculated, each field value representing hydrophobicity of the respective molecules. The calculated sets of field values are combined to generate similarity indexes.

CALCULATING MOLECULAR SIMILARITY

The present disclosure relates to computational chemistry and more specifically to calculating similarity between molecules.

BACKGROUND

Correct molecular superposition and comparison is a well-known problem in computational chemistry. One of the main tasks of computational chemists is to search for new molecules in order to find alternative structures able to bind to a given receptor or decide which molecule modifications are more appropriate in order to improve its affinity, solubility, etc. To this end, they typically rely on known compounds with known properties and the aim is to disclose other molecules highly similar to the reference compounds. This is done by combining molecular alignment and similarity measures to capture the degree of similarity of these molecules. It is not trivial to know which is the correct alignment or similarity metric because different molecule properties are going to be more important depending on the problem. Therefore, multiple similarity metrics exist that help comparing molecules, some of them using steric or electrostatic fields or field extrema.

SUMMARY

As used in the specification and appended claims, the terms “a”, “an” and “the” include both singular and plural referents, unless the context clearly dictates otherwise. Thus, for example, “an apparatus” or “a device” includes one apparatus or device as well as plural apparatuses or devices.

A computer-implemented method of comparing molecules using information of molecule hydrophobicity is disclosed. This is performed by comparing two superposed molecules (m1, m2) and retrieving a Similarity Index (SI). The SI may be in the form of a numerical value that represents the similarity between m1 and m2.

Using information of molecular hydrophobicity permits to generate accurate similarity indexes, particularly for drug-like compounds. As drugs need to be soluble in a polar environment like blood, a drug needs to be hydrophilic. Furthermore, drugs need to pass through cellular membranes and other lipophilic barriers, thus they should not be too hydrophilic. Also, drugs need to interact with the target biological receptor, i.e. they need to desolvate both the ligand and the binding pocket.

In a first aspect, a computer-implemented method of measuring a degree of similarity between two molecules m1, m2 is provided. The method comprises calculating a set of field values for each of the two molecules, each field value representing hydrophobicity of the respective molecules, and combining the calculated sets of field values to generate a similarity index. The similarity index (SI(m1,m2)) may be a numerical value that represents the similarity between two molecules (m1 and m2). The field values may be calculated for a set of M points in space (C). This set of points may be uniformly distributed in space. For each point (c) we may have three coordinate values (c_x, c_y, c_z). In a cubic uniformly distributed grid, M may be equal to the multiplication of the number of points in each coordinate axis.

In some examples, calculating a set of field points for each of the two molecules may comprise creating a grid of points in space, i.e. a set of field point coordinates, and calculating the influence of different hydrophobicity values of each molecule at each field point to generate the value of the field at a given field point. A descriptor point (i) may be a point defined in space where a hydrophobicity value (hv) is present. Each descriptor point may have three coordinate values (i_x, i_y, i_z). A hydrophobicity value (hv(i)) may be a numerical value representing molecule hydrophobicity in a given descriptor point. The number of descriptor points (N) may be the total number of hydrophobicity values. If hydrophobicity values are defined at each atom, then N would be the number of atoms.

In some examples, a possible location of the descriptor points may be the atom centers. A possible way of calculating the hydrophobicity values (hv) may be by using the logarithm of the partition coefficient P (logP) or a partitional type of the logP using fractional components.

Calculating the set of points C may be performed by creating a 3D mesh (e.g. cube or sphere) of uniformly distributed points with the studied molecules in the center. This may be performed by defining a border length (b) and a grid spacing distance (s) and finding the molecule coordinate extrema in each coordinate axis. The 3D mesh origin (c_O) may be defined by subtracting the border length from the minimum molecule coordinates ($Coord_{Min}$):

$$c_{O\_x} = Coord_{Min\_x} - b \qquad \text{(Eq. 1)}$$

Then the size of the 3D mesh may be calculated by finding the first integer number of points (D) in each coordinate direction that multiplied by the grid spacing is bigger than the distance between the maximum molecule coordinates ($Coord_{Max}$) and the minimum molecule coordinates ($Coord_{Min}$) plus two times the border length:

$$D_x = round\left(\frac{|Coord_{Max\_x} - Coord_{Min\_x}| + 2b}{s}\right) \qquad \text{(Eq. 2)}$$

Finally, once the 3D mesh origin and the number of points in each coordinate direction is defined, iteration over all the field points (c) to calculate their coordinates may take place. Field point coordinates may be calculated by adding to the 3D mesh origin coordinates the number of the field point multiplied by the grid spacing in each coordinate direction. Eq.3 shows how the x-axis coordinates of field point Q are calculated (being a number from 0

$$f(hv_i, d_{ci}) = hv_i \cdot \exp(-\alpha \cdot {d_{ci}}^2) \quad \text{(Eq. 7)}$$

being α an adjusting factor.

In some examples, the method may further comprise calculating a local representation of hydrophobicity at different areas of the molecule.

In some examples, the hydrophobicity descriptor (hv_i) may be calculated using the contribution of each atom to the logP by using parameters related to the transfer of the molecule from apolar and polar phases. The logP is the ratio of concentrations of a compound in a mixture of two immiscible phases at equilibrium. These two phases are usually solvents, typically water and an organic phase like octanol. In that case, the logarithm of the partition coefficient P may be calculated as follows:

$$logP = -\frac{\Delta G_{tr}^{o/w}}{2.303\ R\ T} = \frac{\Delta G_{sol}^{water} - \Delta G_{sol}^{organic}}{2.303\ R\ T} \quad \text{(Eq. 8)}$$

where ΔG_{sol} is the solvation free energy or Gibbs free energy in solution (water, organic phase like octanol), R is gas constant and T is the temperature. The solvation free energy or Gibbs free energy (ΔG_{sol}) is the amount of free energy required from the transfer of the molecule from the gas phase to the interior of the solvent.

In other case hv could be calculated using the fractional logP (Pf), which may be defined as the logP calculated with any of the individual components of the solvation free energy like ΔG_{cav}, ΔG_{vW} or ΔG_{ele}:

$$Pf = \frac{\Delta G_X^{water} - \Delta G_X^{organic}}{2.303\ R\ T} \ (X{:}\ ele, cav\ or\ vW) \quad \text{(Eq. 9)}$$

In some examples, calculating the 3D distribution of polar and apolar regions in the molecule may comprise calculating the free energy of solvation (ΔGSol) by combining the cavitation (ΔG_Cav), the van der Waals (ΔG_VW) and the electrostatic components (ΔG_Ele).

In some examples, combining the cavitation (ΔG_Cav), the van der Waals (ΔG_VW) and the electrostatic components (ΔG_Ele) may comprise calculating the solvation free energy by using the accurate polarizable continuum model (PCM) developed by Miertus-Scrocco and Tomasi (MST) and is calculated by adding three energy contributions, the cavitation (ΔG_{Cav}), the van der Waals (ΔG_{VW}) and the electrostatic terms (ΔG_{Ele}):

$$\Delta G_{Sol} = \Delta G_{Cav} + \Delta G_{VW} + \Delta G_{Ele} \qquad \text{(Eq.10)}$$

where ΔG_{cav} is the free energy required for creating a cavity shaped to accommodate the solute in the solvent, ΔG_{VW} is the free energy accounting for dispersion-repulsion interactions between solute and solvent, and ΔG_{ele} is the free energy needed to build up the solute charge distribution in the solvent.

In some examples, the similarity index for each field point may be equal to the size of the intersection between the two molecules (SI(m1,m2)) for all the field points divided by the size of the union of all the field points of the two sets.

In some examples, the intersection of a given field point may be computed picking the smallest absolute field value at that point if both fields are positive or negative and is set to 0 otherwise.

In some examples, the method may further comprise combining multiple similarity indexes.

The proposed method provides key differences over existing solutions. First, the use of atomic contributions to the LogP or Pf is proposed to calculate field points, and second it is proposed to calculate a similarity metric using the hydrophobic field at the field points.

In another aspect, a computational chemistry tool for measuring a degree of similarity between two molecules is disclosed. The tool may comprise means for calculating a set of field points for each of the two molecules, each field point representing hydrophobicity of the respective molecules. The tool may further comprise means for combining the calculated sets of field points to generate a similarity index.

To reduce the computational cost of the proposed algorithm, the computational tool may employ multiple acceleration methods. These methods may include task parallelization or the usage of hardware accelerators to reduce the time required to perform similarity calculation.

Task parallelization may take advantage of the data independence of various tasks of the proposed algorithm. For example, instead of executing the algorithm in a sequential way, it may detect those tasks that may be executed in parallel and execute those tasks concurrently, thus reducing the overall execution time. Task parallelization may be achieved at different levels; in the proposed algorithm, it may be implemented at the molecule level if multiple similarity indexes need to be calculated by calculating all similarity indexes in parallel. It may also be implemented at the field point level in the field calculation or in the similarity index calculation. Additional techniques may be employed at the instruction level for a finer grain parallelization such as vectorization of mathematical operations.

One way of implementing the proposed algorithms would be through a software computer program to be executed in the computational tool. The computational tool may employ graphic processing units (GPUs) for the parallelization of tasks through the usage of specific programming languages.

Finally, critical tasks that are executed often may also be performed by ad-hoc computational devices designed specifically for this task and included in the computational tool. This may be implemented with an electronic circuit capable of performing such tasks and fabricating it or implementing it in reprogrammable hardware devices such as Field Programmable Gate-Arrays (FPGAs).

In another aspect, a computer program product is disclosed. The computer program product may comprise program instructions for causing a computing system to perform a method of measuring a degree of similarity between two molecules according to some examples disclosed herein.

The computer program product may be embodied on a storage medium (for example, a CD-ROM, a DVD, a USB drive, on a computer memory or on a read-only memory) or carried on a carrier signal (for example, on an electrical or optical carrier signal).

The computer program may be in the form of source code, object code, a code intermediate source and object code such as in partially compiled form, or in any other form suitable for use in the implementation of the processes. The carrier may be any entity or device capable of carrying the computer program.

For example, the carrier may comprise a storage medium, such as a ROM, for example a CD ROM or a semiconductor ROM, or a magnetic recording medium, for example a hard disk. Further, the carrier may be a transmissible carrier such as an electrical or optical signal, which may be conveyed via electrical or optical cable or by radio or other means.

When the computer program is embodied in a signal that may be conveyed directly by a cable or other device or means, the carrier may be constituted by such cable or other device or means.

Alternatively, the carrier may be an integrated circuit in which the computer program is embedded, the integrated circuit being adapted for performing, or for use in the performance of, the relevant methods.

BRIEF DESCRIPTION OF THE DRAWINGS

Non-limiting examples of the present disclosure will be described in the following, with reference to the appended drawings, in which:

Figure 1A schematically illustrates a method of measuring a degree of similarity between two molecules according to an example;

Figure 1B schematically illustrates two example molecules to be compared before and after superposition;

Figure 1C schematically illustrates an example superposition in a set of field points used for Similarity Index calculation;

Figure 2 schematically illustrates a method of measuring a degree of similarity between two molecules according to another example;

Figure 3 schematically illustrates a use case for molecular virtual screening according to an example.

Figure 4 schematically illustrates a computational tool for measuring a degree of similarity between two molecules according to an example.

DETAILED DESCRIPTION OF EXAMPLES

Figure 1A schematically illustrates a method of measuring a degree of similarity between two molecules according to an example. In block 105, a set of field point values for each of the two molecules may be calculated. Each field point value may represent hydrophobicity of the respective molecules. Then, in block 110, the calculated sets of field point values may be combined to generate a similarity index.

Figure 1B schematically illustrates two example superposed molecules m1 and m2 to be compared before and after superposition. Figure 1C schematically illustrates an example superposition of two molecules in a set of field points used for similarity index calculation;

In order to compute the Similarity Index, the proposed method uses a set of field points representing the hydrophobicity of each of the two compared molecules (m1 and m2). Field points are calculated creating a grid of points in space and calculating the influence of the different hydrophobicity values of each molecule in each field point. Once the two sets of field points are calculated (F_{m1} and F_{m2}), they are combined to obtain the Similarity Index.

In order to compute the values of the field in a given point (c) a possible implementation would use Eq. 4. A possible implementation of the field formula would be Eq. 6. and another possible implementation of the field formula would be Eq. 7.

A possible way of calculating the hydrophobicity values (hv_i) is using the fractional description of free energies of solvation or the LogP. These hydrophobic descriptors could be the atomic contribution to the solvation free energy or the LogP partition, or any of the terms in which it can be decomposed (Cavitation, Van der Waals or Electrostatic) or a combination of them.

The hydrophobic/hydrophilic character of a drug may be described by using parameters related to the transfer of the molecule from apolar and polar phases. These parameters are typically the free energy of solvation, which is calculated in polar phase usually with water, and in an apolar phase typically represented with apolar solvents such as chloroform, carbon tetrachloride, hexane or octanol. In this context, a well-known measure of hydrophobicity is the LogP, which is computed with Eq. 8.

Typically, a unique global measure of the hydrophobic/hydrophilic character of a drug is computed and used in the drug design process. However, the 3D distribution of polar/apolar regions in the molecule is also important and therefore it is desirable to have a local representation of hydrophobicity/hydrophilicity in different areas of the molecule.

Traditional approaches for atomic-level descriptors are based on empirical parameters obtained for fragments/chemical groups and calculating the influence in a given point with the distance to each fragment. Alternatively, quasi-empirical methods or theoretical methods can be used.

The proposed tool is based on the accurate polarizable continuum model (PCM) developed by Miertus-Scrocco and Tomasi (MST). In the MST method, the free energy of solvation (ΔG_{Sol}) is calculated summing up the cavitation, the van der Waals and the electrostatic components as shown in Eq. 10.

The cavitation and van der Waals components of ΔG_{Sol} can be easily decomposed into atomic contributions, since they depend on the exposure of atoms to the solvent.

The cavitation component (ΔG_{Cav}) is computed following Pierotti (Pierotti, R.A., "*A scaled particle theory of aqueous and nonaqueous solutions*", Chem. Rev. 76, 1976, 717) scaled particle theory adapted to cavities of molecular shape, so that the contribution of a given atom is weighted according to its exposure to the solvent.

$$\Delta G_{Cav}(i) = \frac{SAS_i}{SAS_T} \Delta G_P(i) \qquad \text{(Eq. 11)}$$

where *SAS*

where SAS_i is the solvent accessible surface of atom i, SAS_T is the surface area of such an atom, and $\Delta G_P(i)$ stands for the cavitation free energy of the isolated atom.

The van der Waals term (ΔG_{VW}) is computed using a linear relationship with the atomic surface as given by Equation (12), where the parameter α_i stands for the atomic van der Waals surface tension.

$$\Delta G_{VW}(i) = \alpha_i SAS_i \qquad \text{(Eq. 12)}$$

The electrostatic component of the free energy (ΔG_{Ele}) is computed considering the interaction of the solute charge distribution with the surface elements pertaining to the surface generated by such an atom.

$$\Delta G_{Ele}(i) = \sum_{j=1}^{M\in N} \sum_{k=1}^{N_q} \left\langle \psi^o \left| \frac{q_j^{sol}}{r-r_j} \right| \psi^o \right\rangle \text{ (Eq. 13)}$$

where ψ^o accounts for the wavefunction of the solute in vacuo, q_j^{sol} stands for the charge associated with the surface element j generated in response to the fully polarized charge distribution of the solute in solution, and M is the number of surface elements of the total number of atoms N.

These parameters are used in the proposed method either together or separately. For example, separate LogP values of each of the aforementioned components could be used and then generate multiple interaction fields that could be used for alignment and comparison.

A possible implementation of the Similarity Index (SI(m1,m2)) is to use the Tanimoto or Jaccard index, which is defined as the size of the intersection (I(m1,m2)) divided by the size of the union of the two sets. The result of this operation is a number that ranges between 0 and 1 being 0 completely different and 1 the same. It can be computed with the following formulas (Eq.

14) and (Eq. 15).

$$SI(m1,m2) = \frac{\sum_{i=0}^{P}|I(m1,m2,i)|}{\sum_{i=0}^{P}|F_{m1}(i)|+\sum_{i=0}^{P}|F_{m1}(i)|-\sum_{i=0}^{P}|I(m1,m2,i)|} \qquad \text{(Eq. 14)}$$

$$if\ \left((F_{m1}(i) > 0)\ and\ (F_{m2}(i) > 0)\right)\ or\ \left((F_{m1}(i) < 0)\ and\ (F_{m2}(i) < 0)\right)$$

$$I(m1,m2,i) = \min(|F_{m1}(i)|, |F_{m2}(i)|)$$

$$else \qquad \text{(Eq. 15)}$$

$$I(m1,m2,i) = 0$$

The intersection of a given field point (i) is computed picking the smallest absolute field value at that point if both fields are positive or negative and is set to 0 otherwise.

Additionally, multiple similarity indexes could be used, in which case their similarity metric could be calculated and combined with the hydrophobic similarity with the following formula:

$$Similarity = \sum_{i=0}^{M} W_i \cdot SI_i \quad \text{(Eq. 16)}$$

where M is the number of similarity indexes, W_N are the weights assigned to each similarity index and SI_i are the similarity indexes.

The previously presented combination of similarity indexes could include multiple hydrophobic similarity indexes computed using different hydrophobicity values (hv_i) like the terms in which it can be decomposed (Cavitation, Van der Waals or Electrostatic).

Furthermore, the presented similarity index could be used to guide the superposition process of two molecules with any iterative method either alone or combined with other similarity metrics.

Figure 2 schematically illustrates a method of measuring a degree of similarity between two molecules according to another example. In block 200, information regarding the first and second superposed molecules may be introduced. In block 205 grid values (set of field point coordinates) may be calculated based on the introduced information. In block 210 hydrophobic descriptor values for the first molecule may be determined (either calculated or retrieved from a memory of pre-calculated values) based on the information introduced for the first superposed molecule. In block 215 hydrophobic descriptor values for the second molecule may be determined (similarly, they may be either calculated or retrieved from a memory of pre-calculated values) based on the information introduced for the second superposed molecule. In block 220, the field values may be calculated for the first molecule based on the grid value calculation and the hydrophobic descriptor value determination for the first molecule. Accordingly, in block 225, the field values may be calculated for the second molecule based on the grid value calculation and the hydrophobic descriptor value calculation for the second molecule. In block 230, a similarity index may be calculated based on the field value calculations. Optionally, in block 235, the calculated similarity index may be combined with other similarity index or indexes to generate a final similarity index.

Figure 3 schematically illustrates a use case for molecular virtual screening according to an example. The computational tool 310 may be connected or may receive information about molecules from a molecule database 305. Then information about a reference molecule may be received or introduced. In a typical scenario, a reference molecule will be compared with some or all the molecules that may be included in the molecule database. The molecular database may comprise data related to the molecule such as the molecule name, the number of atoms in the molecule, the coordinates of each atom, the type of atom and the existing type of bond between the atoms. The computational tool 310 may comprise a molecule preprocessing module 315. The molecule preprocessing module 315 may preprocess the information received from the molecule database for any particular molecule. The computational tool 310 may further comprise a reference molecule preprocessing module 330. The reference molecule preprocessing module 320 may preprocess the information introduced or received for the reference molecule. The purpose of the two modules is to adjust the received information so that the two molecules (the one from the database and the reference one) to be comparable. The computational tool may further comprise a molecule superposition module 325 that performs superposition of the two molecules during any iteration. The computational tool may further comprise a similarity calculation module 330. The similarity calculation module 330 may perform calculations according to methods disclosed herein. The computational tool may iterate for all the molecules of the database and then provide the results to a similarity index sorting module 335. The sorting module 335 may rank the results to provide the best match.

Figure 4 schematically illustrates a computational tool for measuring a degree of similarity between two molecules according to an example, and its usage for performing a virtual screening of a molecule database. The computational tool 400 may be connected or may receive information about molecules from molecule database 405. This molecule database may be a storage device including multiple molecule data entries, each of them containing relevant information of the molecule like the molecule name, the number of atoms in the molecule, the coordinates of each atom, the type of atom and the existing type of bond between the atoms.

Information about molecules in the database may be read by a database read processing element 410 in the computational tool 400 and stored in the computational tool system memory 415. The system memory 415 may be in the form of a unified or distributed storage medium. In a typical scenario, one of the stored molecules may be selected as the reference molecule and compared to the rest of the molecules. In other scenarios, a molecule that is not part of the molecule database may be compared with the molecules in the molecule database. A system memory read function is indicated with a straight line whereas a system memory write function is indicated with a dashed line in the example of Fig. 4.

The computational tool 400 may comprise one or multiple hydrophobic descriptors calculation processing elements 420. The hydrophobic descriptors calculation processing element 420 may calculate the hydrophobicity descriptor values for any particular molecule from an entry of the system memory and store these descriptors in another entry in the system memory 415. In the example of Fig. 4, there is data for three molecules stored in system memory entries 415a, 415b and 415c. Accordingly, the hydrophobic descriptors calculation processing elements 420 may write the result of the hydrophobicity descriptor value calculations to memory entries 415e, 415f, 415g. The hydrophobicity descriptor values may be calculated every time or may be calculated once and retrieved from the memory whenever there is a similarity index calculation.

The computational tool may further comprise one or multiple molecule superposition processing elements 425. The molecule superposition processing element may generate a new set of coordinates for any particular molecule from system memory so this molecule is overlapped in space with the reference molecule and then store the new coordinates in system memory. In the example of Fig. 4 for the 3 molecules, three sets of coordinates (indicated as Super Coord 1, 2 and 3) in memory entries 415h, 415i and 415j, respectively.

The computational tool may further comprise one or multiple grid calculation processing elements. The grid calculation processing element 430 may read the coordinates of one or multiple molecules and generate a set of points in space and store them in system memory. System memory entry 415d indicates the memory entry where the field points may have been stored. The purpose of this module is to generate the set of points in space that will be used later to calculate the field values of each molecule.

The computational tool may further comprise one or multiple field calculation processing elements. The field calculation processing element 435 may read the coordinates of a molecule, its hydrophobicity descriptors and a set of field points from system memory and calculate the field value in each of the field points. The calculated field values may then be stored in system memory. In the example, the system memory entries used are entries 415k, 415l and 415m, each entry for the field values of the corresponding molecule 1, 2 and 3, respectively. The coordinates used in the field calculation processing element 435 may be the coordinates previously read from the molecule database or the coordinates generated by a molecule superposition processing element 425.

The computational tool may further comprise one or multiple similarity index processing elements 440. The similarity index processing element 440 may read two sets of field values from system memory, each from a different molecule, and calculate a similarity index that may be stored in system memory. System memory entries 415n and 415o may be used to store the SI for two comparisons. The similarity index module may perform calculations according to methods disclosed herein. In a typical scenario, one set of field values would be always the set of field values of the reference molecule.

The computational tool may further comprise one or multiple sorting processing elements 450. The sorting processing elements may read multiple similarity indexes and order them by the similarity index value, obtaining a molecule order. In a typical scenario, the sorting processing element may generate a molecule ranking that may be used to select those molecules of the molecule database that are more similar or less similar to the reference molecule.

Those of skill would further appreciate that the various illustrative logical blocks, modules, circuits, and algorithm steps described in connection with the embodiments disclosed herein may be implemented as electronic hardware, computer software, or combinations of both. To clearly illustrate this interchangeability of hardware and software, various illustrative components, blocks, modules, circuits, and steps have been described above generally in terms of their functionality. Whether such functionality is implemented as hardware or software depends upon the particular application and design constraints imposed on the overall system. Skilled artisans may implement the described functionality in varying ways for each particular application, but such implementation decisions should not be interpreted as causing a departure from the scope of the exemplary embodiments of the invention.

The various illustrative logical blocks, modules, and circuits described in connection with the embodiments disclosed herein may be implemented or performed with a general purpose processor, a Graphics Processing Unit (GPU), a Digital Signal Processor (DSP), an Application Specific Integrated Circuit (ASIC), a Field Programmable Gate Array (FPGA) or other programmable logic device, discrete gate or transistor logic, discrete hardware components, or any combination thereof designed to perform the functions described herein. A general purpose processor may be a microprocessor, but in the alternative, the processor may be any conventional processor, controller, microcontroller, or state machine. A processor may also be implemented as a combination of computing devices, e.g., a combination of a DSP and a microprocessor, a plurality of microprocessors, one or more microprocessors in conjunction with a DSP core, or any other such configuration.

The steps of a method or algorithm described in connection with the embodiments disclosed herein may be embodied directly in hardware, in a software module executed by a processor, or in a combination of the two. A software module may reside in Random Access Memory (RAM), flash memory, Read Only Memory (ROM), Electrically Programmable ROM (EPROM), Electrically Erasable Programmable ROM (EEPROM), registers, hard disk, a removable disk, a CD-ROM, or any other form of storage medium known in the art. An exemplary storage medium is coupled to the processor such that the processor can read information from, and write information to, the storage medium. In the alternative, the storage medium may be integral to the processor. The processor and the storage medium may reside in an ASIC. The ASIC may reside in a user terminal. In the alternative, the processor and the storage medium may reside as discrete components in a user terminal.

In one or more exemplary embodiments, the functions described may be implemented in hardware, software, firmware, or any combination thereof. If implemented in software, the functions may be stored on or transmitted over as one or more instructions or code on a computer-readable medium. Computer-readable media includes both computer storage media and communication media including any medium that facilitates transfer of a computer program from one place to another. A storage media may be any available media that can be accessed by a computer. By way of example, and not limitation, such computer-readable media can comprise RAM, ROM, EEPROM, CD-ROM or other optical disk storage, magnetic disk storage or other magnetic storage devices, or any other medium that can be used to carry or store desired program code in the form of instructions or data structures and that can be accessed by a computer. Also, any connection is properly termed a computer-readable medium.

Although only a number of examples have been disclosed herein, other alternatives, modifications, uses and/or equivalents thereof are possible. All possible combinations of the described examples are also covered. Thus, the scope of the present disclosure should not be limited by particular examples, but should be determined only by a fair reading of the claims that follow. If reference signs related to drawings are placed in parentheses in a claim, they are solely for attempting to increase the intelligibility of the claim, and shall not be construed as limiting the scope of the claim.

Further, although the examples described with reference to the drawings comprise computing apparatus/systems and processes performed in computing apparatus/systems, the invention also extends to computer programs, particularly computer programs on or in a carrier, adapted for putting the system into practice.

CLAIMS

1. A computer-implemented method of measuring a degree of similarity between two molecules, comprising:

calculating a set of field values for each of the two molecules, each field value representing hydrophobicity of the respective molecules,

combining the calculated sets of field values to generate a similarity index.

2. The method according to claim 1, wherein calculating a set of field points for each of the two molecules comprises

defining a set (C) of points (c) in space;

identifying a set (I) of descriptor points (i) in space, each descriptor point (i) having an associated hydrophobicity value (hv)

calculating the influence of different hydrophobicity values (hv) at each point c to generate the value of the hydrophobicity field.

3. The method according to claim 2, wherein calculating the influence of different hydrophobicity values of each molecule at each point c is performed with the following formula

$$F(c) = \sum_{i=0}^{N} f(hv_i, d_{ci})$$

where the field value $F(c)$ is the sum of the contributions of the different descriptor points to that field point c, being N the number of hydrophobicity values, c the field point from the set C and $f(hv_i, d_{ci})$ a field formula using hydrophobicity values (hv_i) and the distance (d_{ci}) between each descriptor point i and the field point c.

4. The method according to claim 3, wherein the field formula $f(hv_i, d_{ci})$ is:

$$f(hv_i, d_{ci}) = \frac{hv_i}{{d_{ci}}^2}$$

5. The method according to claim 3, wherein the field formula $f(hv_i, d_{ci})$ is:

$$f(hv_i, d_{ci}) = hv_i \cdot \exp(-\alpha \cdot {d_{ci}}^2)$$

6. The method according to any of claims 1 to 5, further comprising calculating a local representation of hydrophobicity at different areas of the molecule.

7. The method according to claim 6, wherein the hydrophobicity value (hv_i) is calculated using the logarithm of the partition coefficient P, LogP or fractional components of LogP, by using parameters related to the transfer of the molecule from apolar and polar phases of the molecule wherein

$$LogP = \frac{\Delta G_{polar} - \Delta G_{apolar}}{2.303RT}$$

where ΔG is the solvation energy in solution (water, organic phase like octanol), R is gas constant and T is the temperature

8. The method according to claim 7, wherein the hydrophobicity value is calculated for the center of each atom of the molecule.

9. The method according to claim 7 or 8, wherein calculating the 3D distribution of polar and apolar regions in the molecule comprises calculating the free energy of solvation (ΔG) by using one or more of the cavitation (ΔG_Cav), the van der Waals (ΔG_VW) and the electrostatic components (ΔG_Ele).

10. The method according to claim 9, wherein using comprises summing up the cavitation (ΔG_{Cav}), the van der Waals (ΔG_{VW}) and the electrostatic components (ΔG_{Ele}), wherein:

$$\Delta G_{Sol} = \Delta G_{Cav} + \Delta G_{VW} + \Delta G_{Ele}$$

11. The method according to any of claims 1 to 10, wherein the similarity index for each field point is equal to the size of the intersection (I(m1,m2))

between the two molecules for all the field points divided by the size of the union of all the field points of the two sets.

12. The method according to claim 1, wherein the intersection of a given field point is computed picking the smallest absolute field value at that point if both fields are positive or negative and is set to 0 otherwise.

13. The method according to any of claims 1 to 12, further comprising combining multiple similarity indexes.

14. A computational chemistry tool for measuring a degree of similarity between two molecules, comprising:

means for calculating a set of field points for each of the two molecules, each field point representing hydrophobicity of the respective molecules; and

means for combining the calculated sets of field points to generate a similarity index.

15. The computational chemistry tool according to claim 14, wherein the means for calculating a set of field points for each of the two molecules comprises:

a memory;

one or more hydrophobicity descriptor processing elements to determine hydrophobicity descriptor values; and

one or more field calculation processing elements to read the determined hydrophobicity descriptor values and a set of field points from the memory and calculate the field value in each of the field points.

16. The computational chemistry tool according to claim 15, wherein the one or more hydrophobicity descriptor processing elements are configured to calculate the hydrophobicity descriptor values.

17. The computational chemistry tool according to claim 15, wherein the hydrophobicity descriptor values are pre-calculated and the one or more

hydrophobicity descriptor processing elements are configured to retrieve the hydrophobicity descriptor values from a memory.

18. The computational chemistry tool according to any of claims 14 to 17, further comprising a grid calculation processing element to calculate the field points.

19. The computational tool according to any of claims 15 to 18, further comprising one or more molecule superposition processing elements to generate new coordinates for the molecules and store them in the system memory.

20. The computational chemistry tool according to any of claims 14 to 19, wherein the means for combining the calculated sets of field points to generate a similarity index comprises a similarity index processing element.

21. The computational chemistry tool according to claim 20, further comprising a sorting processing element.

22. The computational chemistry tool according to any of claims 14 to 21, comprising multiple hydrophobic descriptor calculation processing elements, molecule superposition processing elements and field calculation processing elements to compare molecules in a parallel configuration.

23. A computer program product comprising program instructions for causing a computing system to perform a method according to any of claims 1 to 13.

24. A computer program product according to claim 23, embodied on a storage medium.

25. A computer program product according to claim 23, carried on a carrier signal.

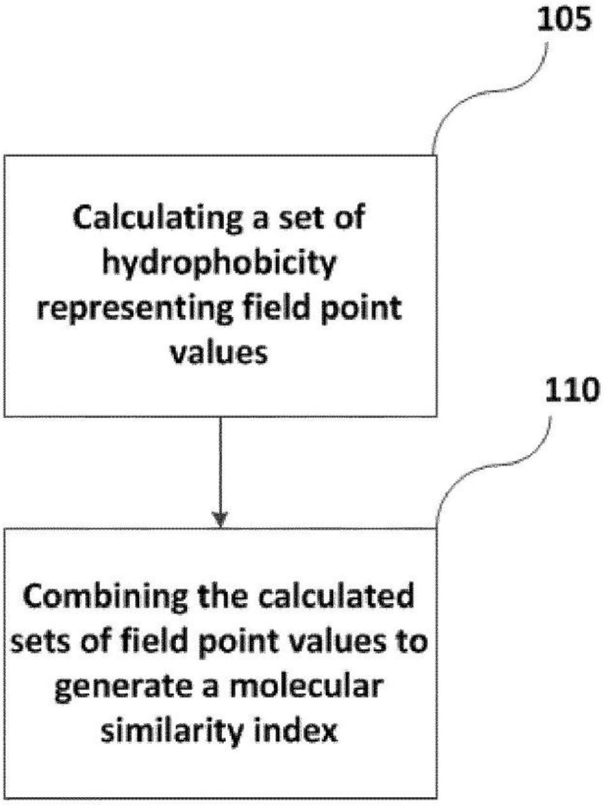

105
Calculating a set of hydrophobicity representing field point values
110
Combining the calculated sets of field point values to generate a molecular similarity index

Fig. 1A

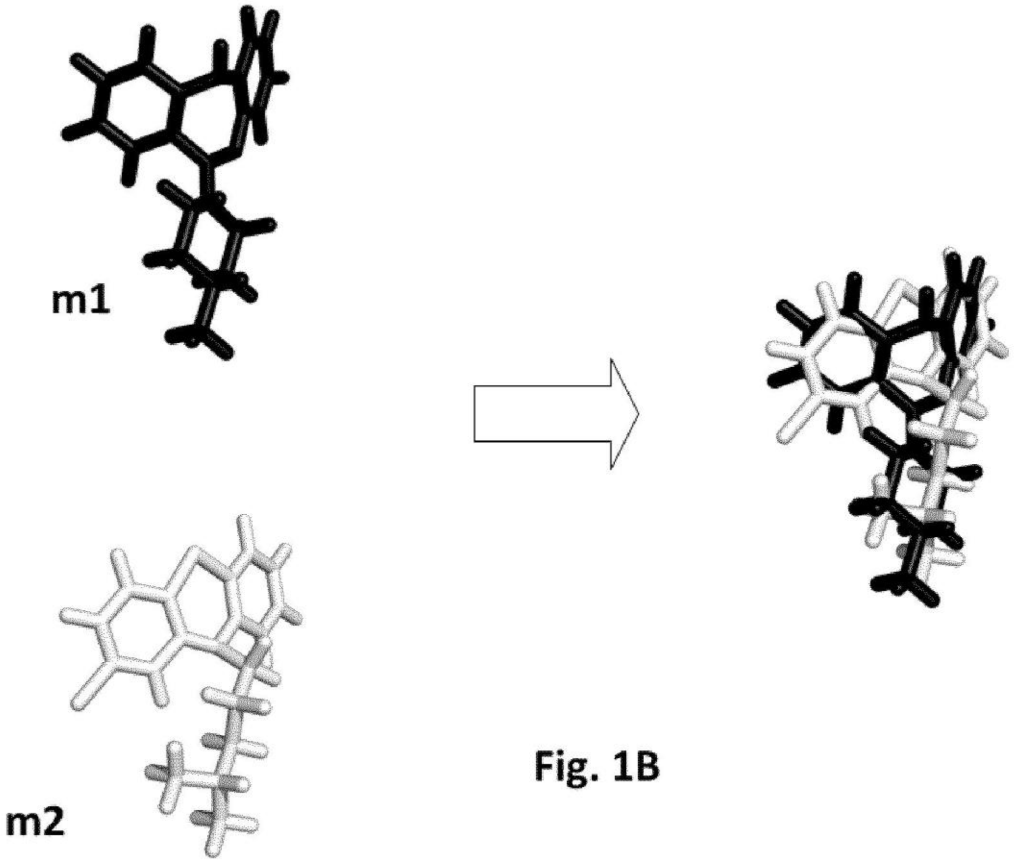


Fig. 1B

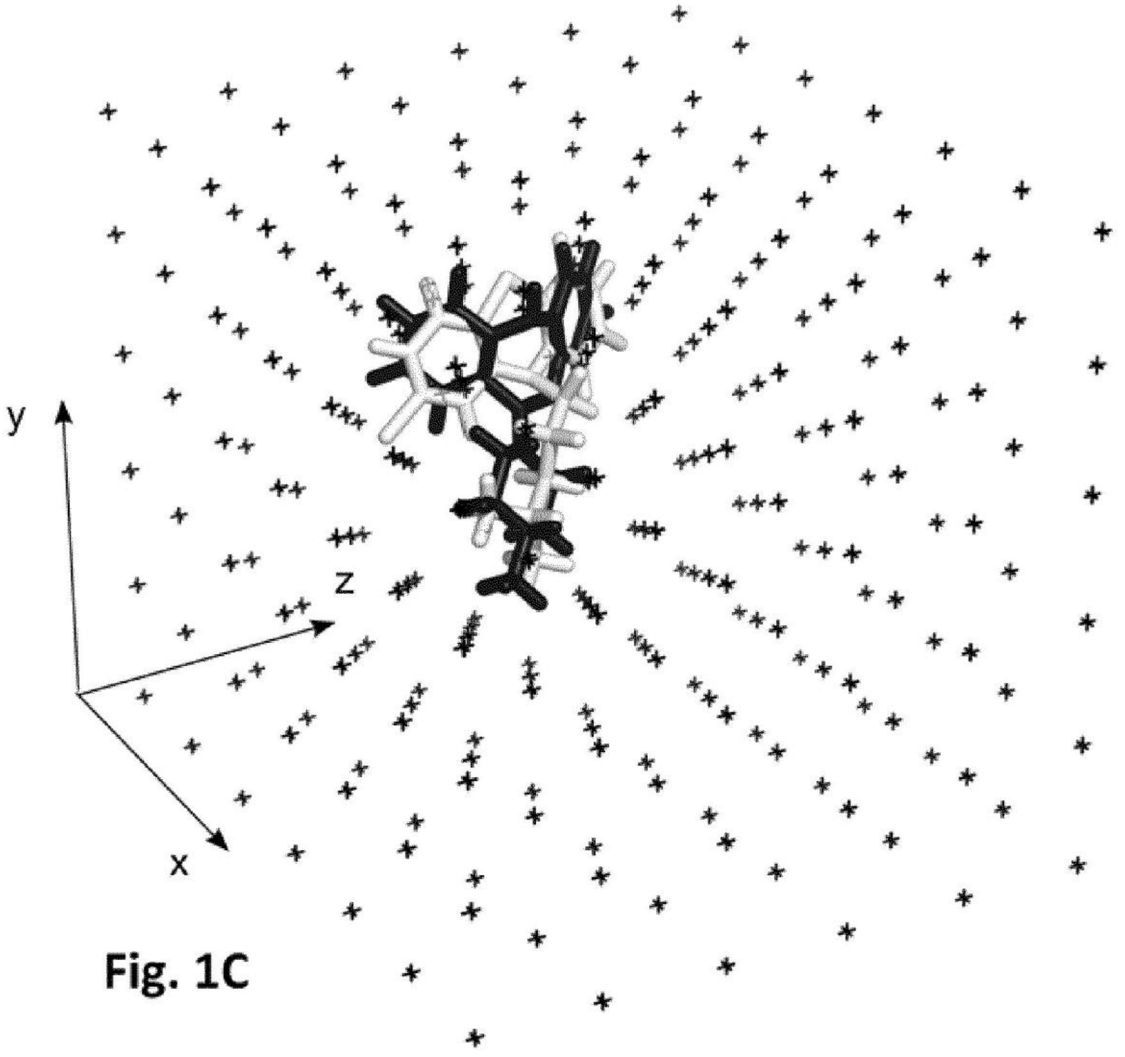


Fig. 1C

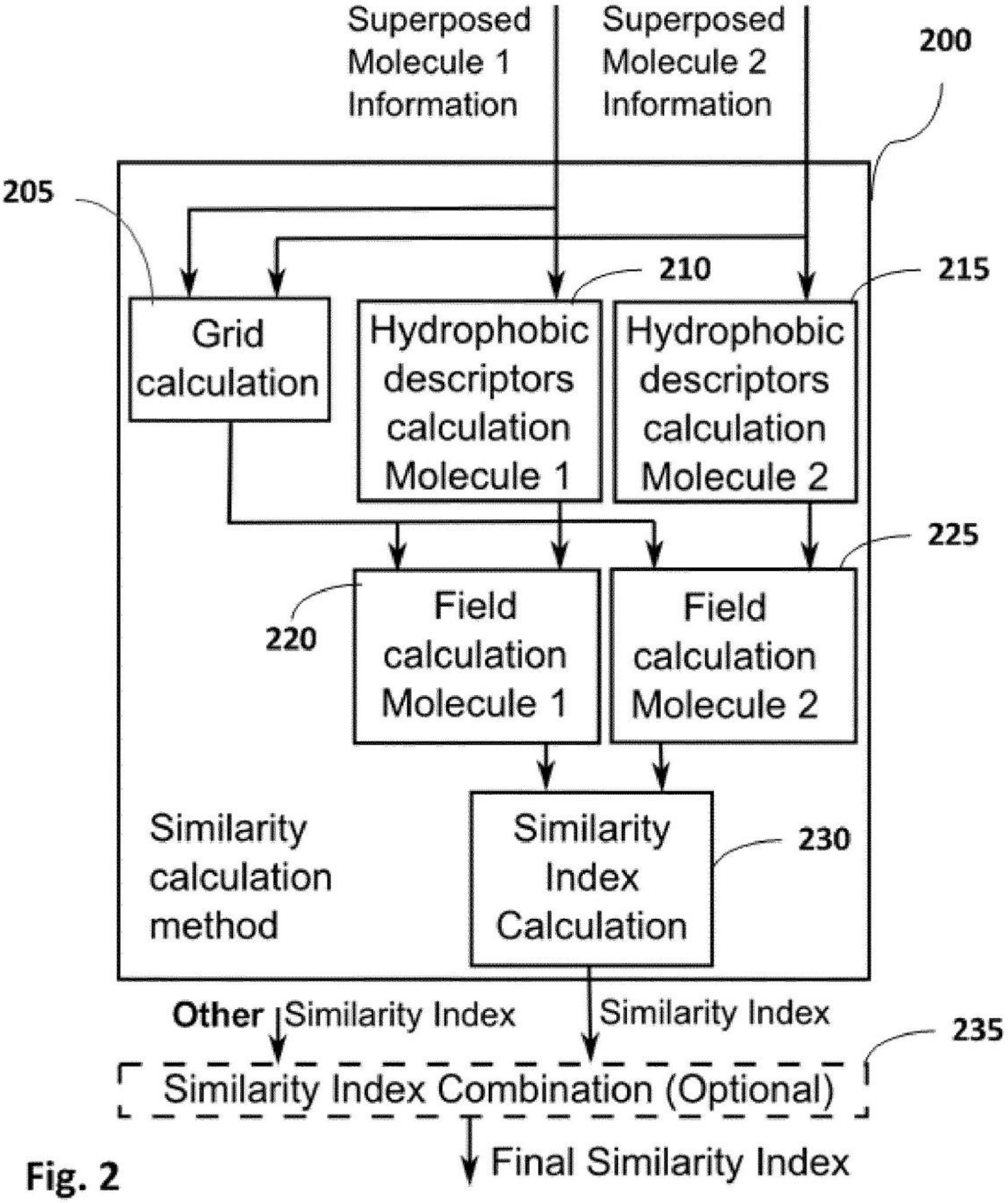

Superposed
Molecule 1
Information
Superposed
Molecule 2
Information
200
205
210
215
Grid calculation
Hydrophobic descriptors calculation Molecule 1
Hydrophobic descriptors calculation Molecule 2
225
220
Field calculation Molecule 1
Field calculation Molecule 2
Similarity calculation method
Similarity Index Calculation
230
Other Similarity Index
Similarity Index
235
Similarity Index Combination (Optional)
Final Similarity Index

Fig. 2

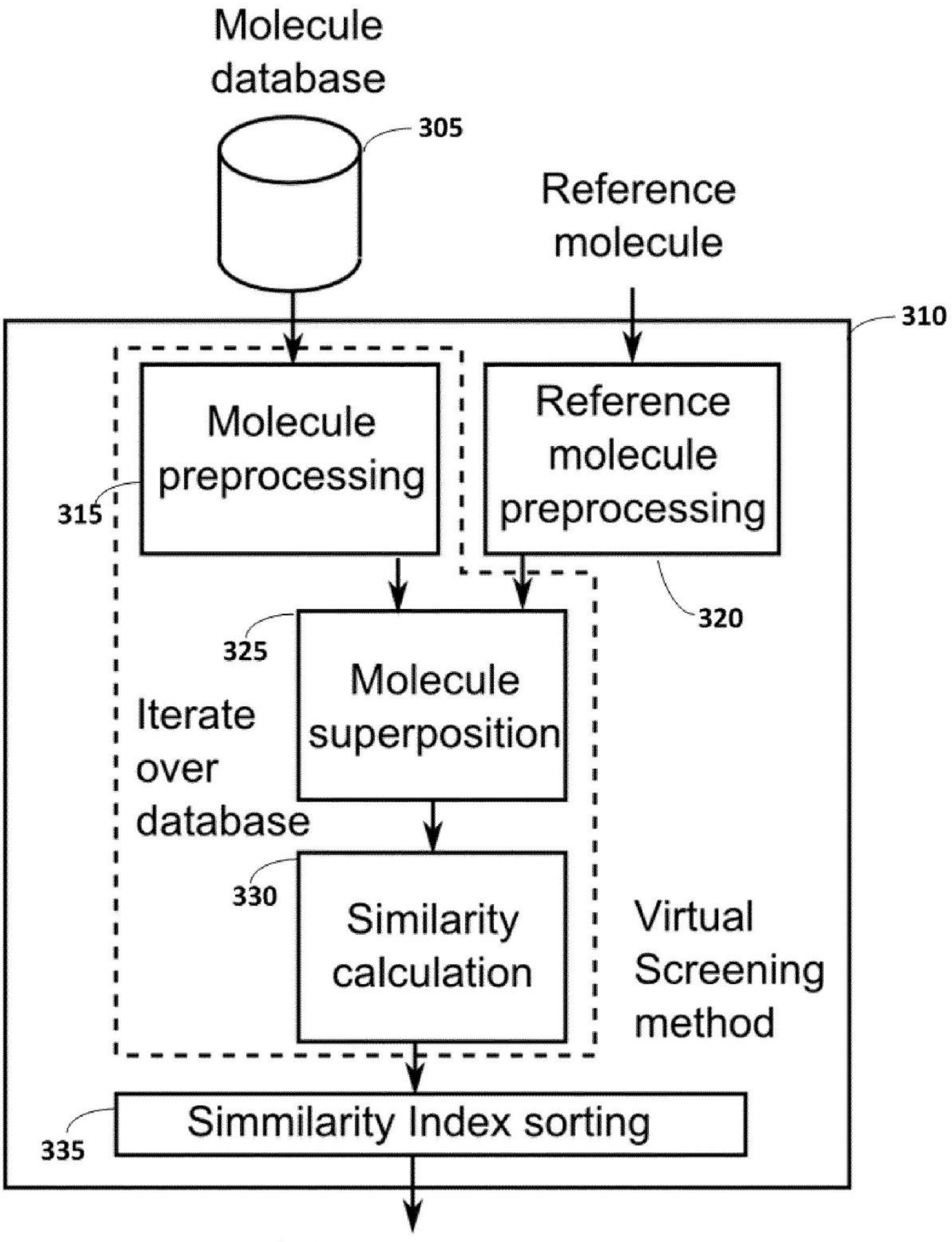

Molecule database
305
Reference molecule
310
Molecule preprocessing
315
Reference molecule preprocessing
320
325
Molecule superposition
Iterate over database
330
Similarity calculation
Virtual Screening method
Simmilarity Index sorting
335
Molecule ranking

Fig. 3

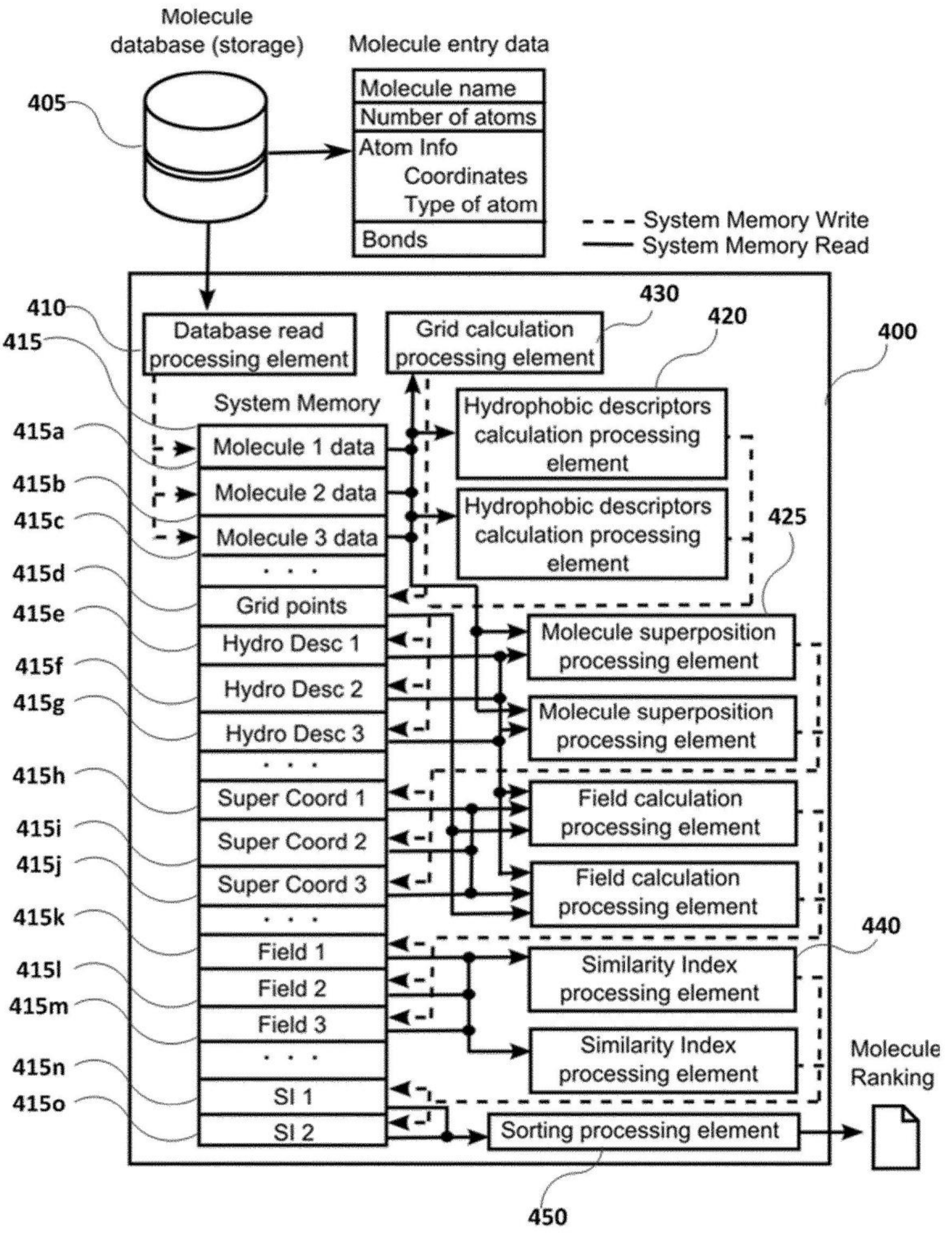

Molecule database (storage)
Molecule entry data
Molecule name
Number of atoms
Atom Info
Coordinates
Type of atom
Bonds
System Memory Write
System Memory Read
405
410
415
Database read processing element
Grid calculation processing element
430
420
400
System Memory
415a
Molecule 1 data
415b
Molecule 2 data
415c
Molecule 3 data
. . .
415d
Grid points
415e
Hydro Desc 1
415f
Hydro Desc 2
415g
Hydro Desc 3
. . .
415h
Super Coord 1
415i
Super Coord 2
415j
Super Coord 3
. . .
415k
Field 1
415l
Field 2
415m
Field 3
. . .
415n
SI 1
415o
SI 2
Hydrophobic descriptors calculation processing element
Hydrophobic descriptors calculation processing element
425
Molecule superposition processing element
Molecule superposition processing element
Field calculation processing element
Field calculation processing element
440
Similarity Index processing element
Similarity Index processing element
Molecule Ranking
Sorting processing element
450

Fig. 4

CHAPTER FOUR
Results summary

4 RESULTS SUMMARY

The study of molecular similarity and the aim to exploit 3D-distribution patterns of lipophilicity have provided the necessary impetus to carry out this research project. Under this framework, we have developed a LBVS tool that exploits a novel set of hydrophobic descriptors derived from QM self-consistent reaction field calculations. The outcome of this work is presented in two main sections that (1) define a computational approach that enables to perform a reliable molecular alignment, and (2) establish an ordered similarity relationship between aligned pairs of compounds with respect to a molecular reference template. The latter includes a preliminary study in 3D-QSAR, where our alignment descriptors are applied to structure-activity relationships.

4.1 Lipophilic descriptors in molecular alignment

Molecular alignment is a core procedure in 3D CADD tools, which enables pharmacophore elucidation, QSAR analysis, or to perform VS campaigns. Obtaining a correct alignment is not trivial and is influenced by several factors, including the quality of the physicochemical descriptors used. Traditionally, steric and electrostatic descriptors have dominated the choice of molecular descriptors,[178,179] whereas other molecular determinants of drug activity have been mostly ignored, or given a secondary role, such as the hydrophobic/hydrophilic balance.

Relying on the hypothesis that the maximal achievable binding affinity variation for an optimized drug-like molecule is largely due to desolvation,[109] PharmScreen is presented as a novel strategy for 3D alignment of small molecules. PharmScreen exploits the usage of molecular lipophilicity and hydrogen-bond donors/acceptors to obtain accurate 3D molecular alignments.

4.1.1 Molecular descriptors and implementation

The alignment method is based on the hydrophobic descriptors obtained by using the Miertus-Scrocco-Tomasi self-consistent continuum solvation method (MST-SCRF),[166] particularly the version parametrized for the semiempirical Hamiltonian RM1. Under this framework, the molecular hydrophobicity can be partitioned into atomic contributions, each decomposable into electrostatic, and non-electrostatic (cavitation and van der Waals) components. However, due to the large redundancy between non-electrostatic components, molecular overlays have been determined only from the $\log P_{ele}$ and $\log P_{cav}$, which have demonstrated to have a good performance in comparison to other standard 3D-QSAR techniques.[162,163]

Starting from the 3D topological distribution of hydrophobicity constituted by the atomic contributions, a set of molecular moments inspired on the multipolar expansion of the electrostatic potential are defined.[128] Each moment is defined by using an expansion center and the principal axes of a specific tensor matrix. The use of LogP_{total} and LogP_{ele} allows the definition of a “hydrophobic monopole” where the “quadrupole tensor” is positioned. The former exploits the information about the net polar/apolar character of atoms, and the latter relies on the differential electrostatic interaction of the individual atoms arising upon hydration relative to solvation in *n*-octanol. In the case of $\log P_{cav}$, which encodes information about shape, the alignment is accomplished throughout calculations of the moments of inertia. For each molecular pair, a pool of alignments is obtained based on these moments.

In the second stage, a score function selects the best overlay among the multiple alignments derived from the hydrophobic descriptors (Figure 6). The $\log P_{total}$ and $\log P_{ele}$ of polar groups are negative, reflecting the preference for solvation in water, but they do not include information about the hydrogen-bond donor/acceptor character. For this reason, this information is included in the score function as a third component to preserve the information about the acceptor/donor recognition properties of the compound. To this end, an arbitrary parameter of +1 is assigned to all hydrogen considered as donors, and − 1 for N and O atoms that may act as acceptors. For each overlay, the atomic contributions are projected into a 3D grid using the exponential function implemented in CoMSiA.[180]

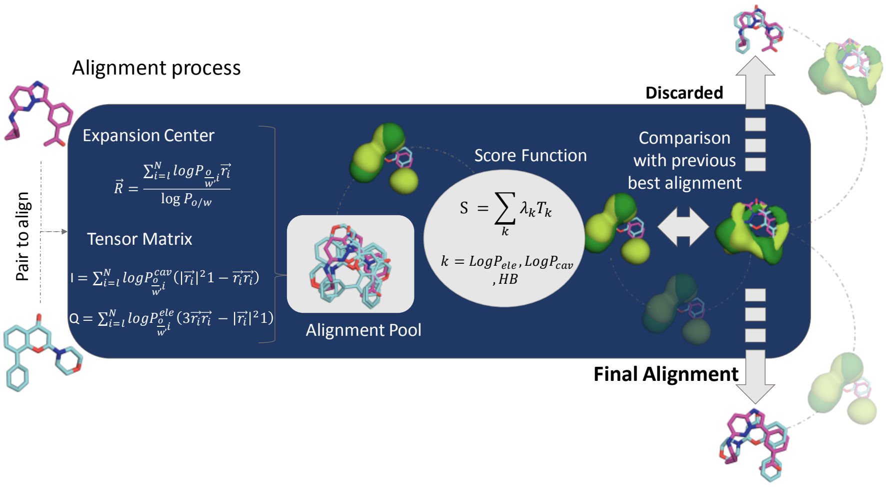

Alignment process
Pair to align
Expansion Center
$\vec{R} = \frac{\sum_{i=l}^{N} logP_{\frac{o}{w},i}\vec{r_i}}{\log P_{o/w}}$
Tensor Matrix
$I = \sum_{i=l}^{N} logP_{\frac{o}{w},i}^{cav}(|\vec{r_i}|^2 1 - \vec{r_i}\vec{r_i})$
$Q = \sum_{i=l}^{N} logP_{\frac{o}{w},i}^{ele}(3\vec{r_i}\vec{r_i} - |\vec{r_i}|^2 1)$
Alignment Pool
Score Function
$S = \sum_k \lambda_k T_k$
$k = LogP_{ele}, LogP_{cav}, HB$
Comparison with previous best alignment
Discarded
Final Alignment

Figure 6 | Alignment scheme. The quadrupole tensor (Q) and inertial tensors (I) positioned in the center of expansion ($\vec{R}$) define the moments for each pair of compounds to build the alignment pool. The global score (S) is determined by the Tanimoto coefficient (T_k) obtained for the different molecular fields using a normalized weighting factor (λ_k). All proposed alignments in the pool are evaluated, and only the one with the highest score is reported as the final alignment for each molecule pair.

4.1.2 Weights calibration and QM methods

The scoring function is determined combining the Tanimoto coefficients obtained for the different molecular fields using a normalized weighting factor. The training set used to calibrate the score function weights, as well as to evaluate the QM method used to derive the hydrophobic contributions, consisted of 14 series of compounds used as a benchmarking ensemble (with 410 ligands in total). This set combines X-ray crystal structures[181,182] and pharmacophoric models.[162] Molecular overlays were determined using either two or three molecular fields. The former combines the atomic contributions of the $\log P_{total}$ supplemented with a hydrogen bond field (HB). The third molecular field combines the electrostatic ($\log P_{ele}$) and cavitation ($\log P_{cav}$) contributions in addition to the HB field. The analysis of the weighting factors was performed in the presence and absence of the HB contribution.

The optimal weights are found to be close to 30/70 for the combinations of descriptors $\log P_{total}$/HB and $\log P_{ele}$/$\log P_{cav}$. Upon inclusion of the HB field in the latter, the weighting factors were refined to 15 ($\log P_{ele}$), 55 ($\log P_{cav}$), and 30 (HB).

On average, the weighted factors chosen for the combination of $logP_{ele}$ and $logP_{cav}$ give rise to obtain better results compared to the exclusive use of the total hydrophobicity. The addition of HB field also enhances the performance in the two field combinations, as can be noted in Table 2.

Table 2 | Weighting factors chosen for molecular overlays upon combination of electrostatic and cavitation components of hydrophobicity and upon combination with the HB descriptors. The average value (%) of successful overlaps for all ligand-template pairs are reported. The molecular overlay is considered to be correct when the RMSD of the heavy atoms is ≤ 2.0 Å from the X-ray arrangement or pharmacophoric model.

	weights (LogP contributions/HB)	
$logP_{total}$	**100/0**	**70/30**
Subset 1 (155)	20.8	24.0
Subset 2 (255)	46.0	52.9
total (410)	36.7	41.9
$logP_{ele}$/log P_{cav}	**30/70/0**	**15/55/30**
Subset 1 (155)	34.1	38.0
Subset 2 (255)	64.0	65.6
total (410)	52.7	54.0

Due to the low computational cost compared to *ab initio* methods, the calibration of the weighting factors in the similarity function was performed using the semiempirical Hamiltonian RM1. Nevertheless, the performances of descriptors derived from MST/RM1[4] and MST/B3LYP/6-31G(d)[8] calculations (calculated using locally modified versions of MOPAC[183] and Gaussian 09[184], respectively) were compared to evaluate the influence of the method in the alignment accuracy for the molecules included in the training set. As shown in Figure 7, similar results were obtained at the two levels of theory. In light of these results, the subsequent generation of hydrophobic descriptors was carried out by using the MST solvation model parametrized for the semiempirical Hamiltonian RM1, which was integrated in PharmScreen.

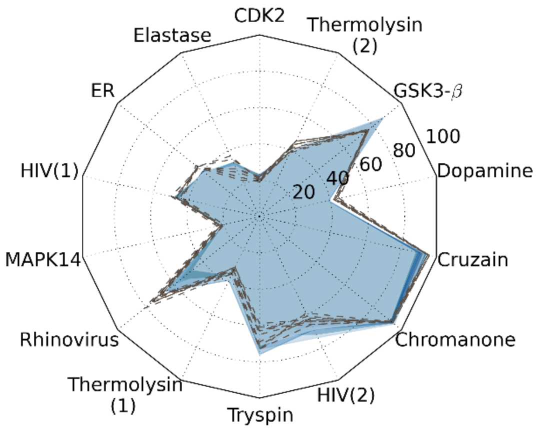


Figure 7 | Overlay accuracy (%) determined for logP$_{ele}$/logP$_{cav}$/HB fields. The average values of correctly predicted alignments from semiempirical RM1 and B3LYP hydrophobic contributions are shown as dashed lines and solid area, respectively. The molecular overlay is considered to be correct when the RMSD of the heavy atoms is ≤ 2.0 Å from the X-ray arrangement or pharmacophoric model.

4.1.3 Overlay accuracy validation

To validate the alignment approach presented in this study, a retrospective validation study was carried out with the CCDC AstraZeneca Overlays Validation Test Set[185] (AZ), which comprises 1456 ligands in 121 sets. In Ref.[186] the ligands in the distinct X-ray crystallographic structures were examined and classified in four categories based on how easy or difficult it would be to reproduce the experimental overlay. This systematic analysis yielded a classification of the ligands as easy, moderate, hard, and unfeasible, comprising 210, 1447, 187 and 103 compounds, respectively.

The accuracy of the molecular overlays predicted from logP$_{ele}$/logP$_{cav}$/HB fields (weights of 15/55/30, respectively) is compared with the results obtained by using electrostatic/steric fields. To further validate the findings of this comparison, the weights of these latter fields were also optimized using the compounds in the training set. The most efficient weighting fields were found to be close to 50/50.

The results for the AZ set confirm the suitability of the MST based-hydrophobic parameters for generating molecular overlays with correct predictions. A similar success rate was obtained using the two sets of descriptors for compounds of easy and moderate categories (96.5% and 79.4%). However, a success rate of 54.4% and 31.3% were obtained with MST-based descriptors for the molecules classified into hard and unfeasible sets. Using electrostatic/steric fields the performance for the hard and unfeasible set category was slightly reduced to 48% and 27%, respectively.

It is worth noting that the apparently similar overall performance of logP$_{ele}$/logP$_{cav}$/HB and electrostatic/steric fields does not necessarily imply that these descriptors lead to identical overlays for a given compound. Figure 8 shows the number (%) of identical superpositions between the best pose predicted from the hydrophobic/HB and electrostatic/steric fields. This comparison shows that the number of similar orientations decreases as the difficulty of the category increases. Thus, for the 22 sets in the easy category, logP$_{ele}$/logP$_{cav}$/HB and electrostatic/steric fields lead to the same molecular alignments in 18 cases, and the agreement is larger than 80% in the remaining 4 cases. This level of identity is attained in 50 out of the 73 sets included in the moderate category, and it is found only in 4 cases out of the 18 sets pertaining to the hard category. Finally, only a single set reaches a number of identical overlays larger than 50% for the unfeasible targets. Thus, although the two sets of descriptors yielded similar overall performances, they do not lead necessarily to similar overlays for the same compounds, especially for those included in the most difficult categories.

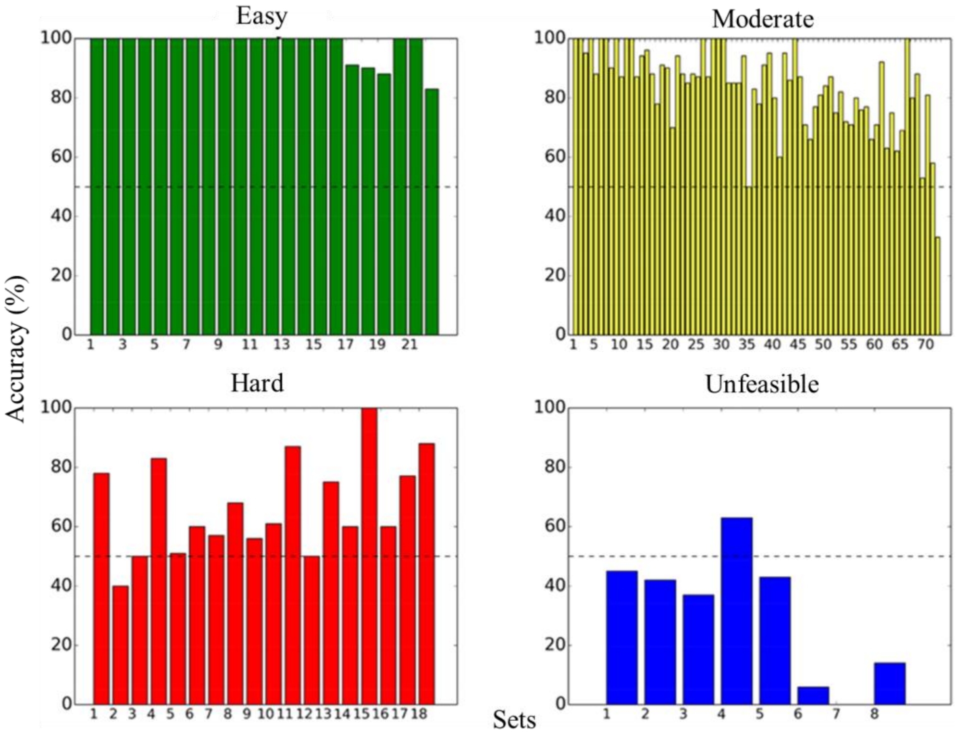


Figure 8 | Degree of identity (%) between the molecular overlays obtained predicted from logP$_{ele}$/logP$_{cav}$/HB and electrostatic/steric fields. Two orientations are considered to be equal when the RMSD of the heavy atoms is ≤ 2.0 Å from the X-ray arrangement

To illustrate the preceding comments, Figure 9 shows the alignment of the ligand taken from X-ray structure 1YI3 over the compound extracted from 2C3I when hydrophobic descriptors are used in comparison to the alignment produced by electrostatic/steric descriptors. In this particular case, hydrophobic descriptors yield the closest orientation to the X-ray structure (RMSD of 1.64 Å), whereas electrostatic/steric fields lead to a wrong orientation (RMSD of 4.34 Å). On the other side, Figure 10 shows how the steric/electrostatic fields bring the ligand extracted from 3ORX to the closest alignment to the X-ray structure (RMSD of 1.54 Å), while hydrophobic descriptors induce an incorrect orientation (RMSD of 2.14 Å).

These results highlight the complementarity that may exist between hydrophobic/HB and electrostatic/steric fields. The chemical space is broad and diverse, and two molecular moieties that appear seemingly equivalent according to a given set of descriptors may not be equivalent when another set is used. Accordingly, each subset of actives can be addressed from different perspectives.[187,188] Thus, the possibility to select different properties in searching for a drug-like compound go beyond the range of molecular similarity relationships under investigation and may facilitate the study of specific cases.

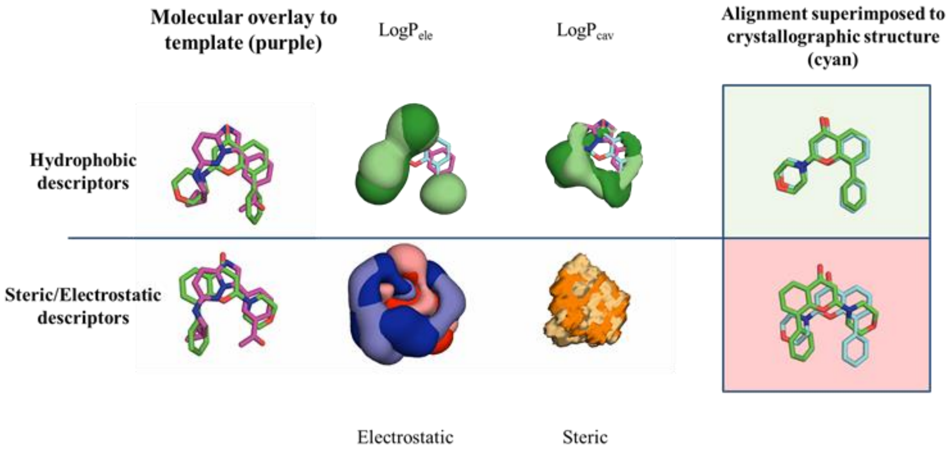

Molecular overlay to template (purple)
LogP$_{ele}$
LogP$_{cav}$
Alignment superimposed to crystallographic structure (cyan)
Hydrophobic descriptors
Steric/Electrostatic descriptors
Electrostatic
Steric

Figure 9 | Molecular overlays obtained for compounds linked to the serine protein kinase/threonine PIM-1(PDB code:2c3i, purple). Representation of the overlapping 1yi3 (green) superimposed using logPele / logPcav / HB (upper). Representation of the 2c3i (green) binder superimposed using electrostatic / steric descriptors (lower).

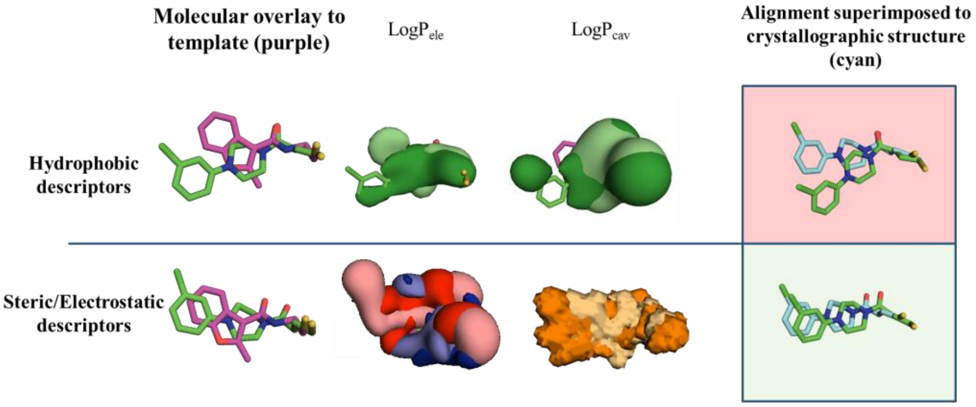

Molecular overlay to template (purple)
LogP$_{ele}$
LogP$_{cav}$
Alignment superimposed to crystallographic structure (cyan)
Hydrophobic descriptors
Steric/Electrostatic descriptors

Figure 10 | Molecular overlays obtained for compounds linked to the serine protein kinase/threonine pdpk1(PDB code:3orz, purple). Representation of the overlapping 3orx (green) superimposed using logPele / logPcav / HB (upper). Representation of the 3orx (green) binder superimposed using electrostatic / steric descriptors (lower).

In the interest of standardizing the assessment of alignment tools, Jones et al. proposed the AlignScore metric: only sets with a percentage of success ≥ 50% are considered as correct. The analysis of PharmScreen’s overlays making use of this metric returns values of 100%, 93%, 55%, and 13% for easy, moderate, hard, and unfeasible sets. Previous studies used AlignScore to analyze the performance of two alignment tools: the Cambridge Structural Database-driven overlay program (CSD)[189] and MolAlign[190]. For the former, the percentage of correct overlays was 95%, 76%, 39%, and 0%, although it must be noticed that in this case the starting point was not the experimental conformation, which increases the complexity of the evaluation and might justify the lower performance. With regard to MolAlign, the number (%) of correctly predicted overlays was 95%, 68%, 44%, and 13% (results derived by using conformers generated with Balloon and Confect, considering a geometrically successful arrangement in any of the top five solutions). Although caution is required for a quantitative comparison due to the differences in the computational protocol and performance metrics, present results suggest that PharmScreen produces competitive overlays.

Overall, based on logPele/logPcav/HB scoring function, the template-ligand pairs can be compared and ranked, enabling the use of PharmScreen as a LBVS tool through a pre-generated ensemble of conformers for flexible compounds.

4.2 Lipophilic descriptors in 3D-QSAR

After developing and applying the proposed descriptors successfully to molecular alignment in Section 3.1, this approach was tested in 3D-QSAR studies to validate its ability to predict SAR models. Due to the partitioning of lipophilicity in atomic contributions, the graphical representation of the distribution pattern of polar and apolar regions can be adapted to 3D-QSAR methods in a straightforward way.

4.2.1 On the QM method applied to derive the 3D-QSAR model

As in the case of the alignment study, the effect of the QM accuracy level in 3D-QSAR was evaluated. The hydrophobic descriptors derived from MST method were obtained using both the semiempirical RM1 Hamiltonian and the version parametrized at the B3LYP/6-31G(d) level. The models were assessed for a subset of four systems (D2 inhibitors, antifungal chromanones, GSK3-β, and cruzain inhibitors)[162]. Figure 11 reveals that there is a large resemblance in the overall performance of MST/RM1 and MST/B3LYP for all datasets. Accordingly, the computationally less demanding RM1 method seems to be a promising choice for 3D-QSAR studies using hydrophobic parameters. Thus, the benchmark dataset was examined using the MST/RM1 descriptors.

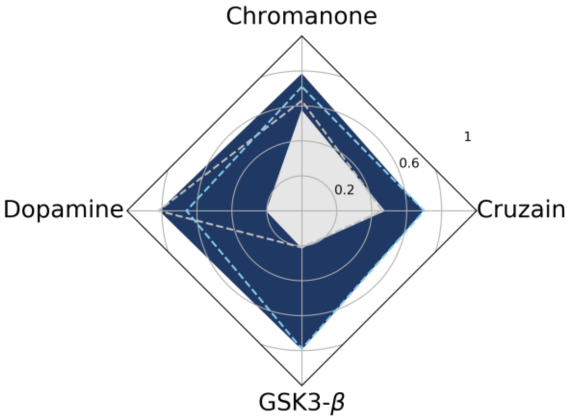


Figure 11 | Statistical parameters (r^2, dashed line, and S, solid area) for the test set of the 3D-QSAR HyPhar models obtained from MST/B3LYP (grey) and MST/RM1(blue) calculations for the four sets of compounds.

4.2.2 3D-QSAR validation

The standard CoMFA/CoMSiA systems were compared with the 3D-QSAR model presented in this work. This analysis was carried out using the comprehensive benchmark dataset compiled by Sutherland and coworkers,[191] which comprises 1265 structures grouped in 8 categories (AChE inhibitors, ligands for BZR, ACE inhibitors, COX-2 inhibitors, THER inhibitors, DHFR inhibitors, GPB inhibitors, and THR inhibitors).

In general, the different 3D-QSAR models return similar statistical parameters for the test set (R-squared and the standard error of the regression), though there is a slight improvement for GPB and THERM systems according to hydrophobic and CoMFA/RM1 models.

To measure the value of these models to rank the compounds, a comparison was made between predicted and experimental potencies using the Spearman correlation coefficient, whereas the ability to discern between active and inactive compounds were estimated from the specificity and sensitivity properties. The Spearman (*Rs*) coefficient for the first (Q1) and second (Q2) quartile performed by CoMFA/RM1 and hydrophobic models are reported in Figure 12. Except for COX2 and THR, higher *Rs* values are observed for the presented model in both Q1 and Q2. In this latter case, CoMFA/RM1 returns a higher *Rs* coefficient only for BZR.

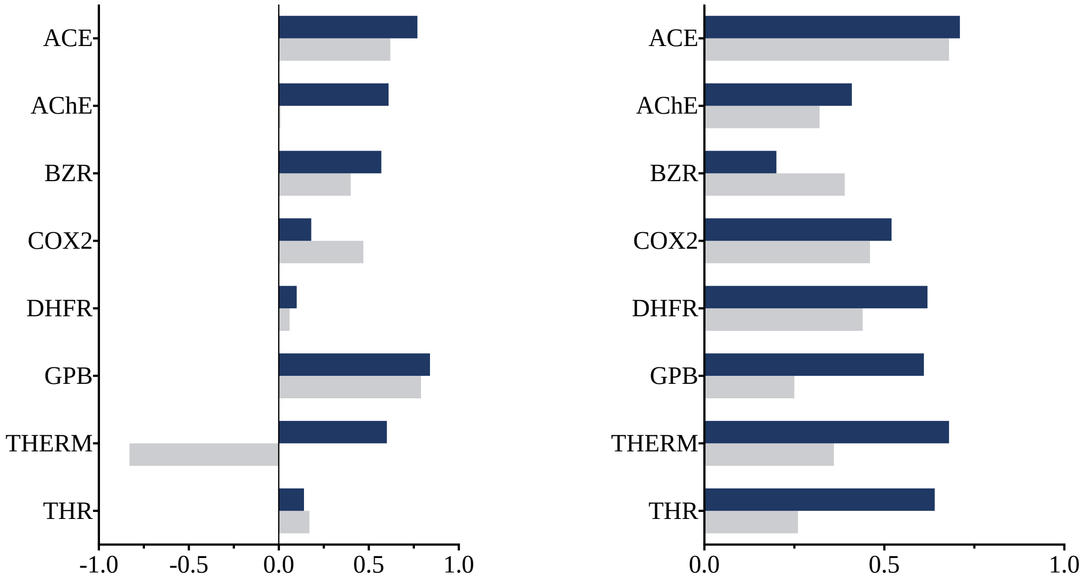


Figure 12| Spearman coefficient (Rs) for the first (Q1), left, and the second (Q2), right, quartiles for CoMFA/RM1 (grey) and hydrophobic (dark blue) models.

Figure 13 shows the values for specificity (inactive molecules are correctly detected as such) and sensitivity (active molecules correctly recognized as such) obtained from the tested models. The hydrophobic approach has a slightly better performance in sensitivity/specificity values for AChE, THERM and THR systems, whereas the opposite trend is found for CoMFA/RM1 in ACE and COX2.

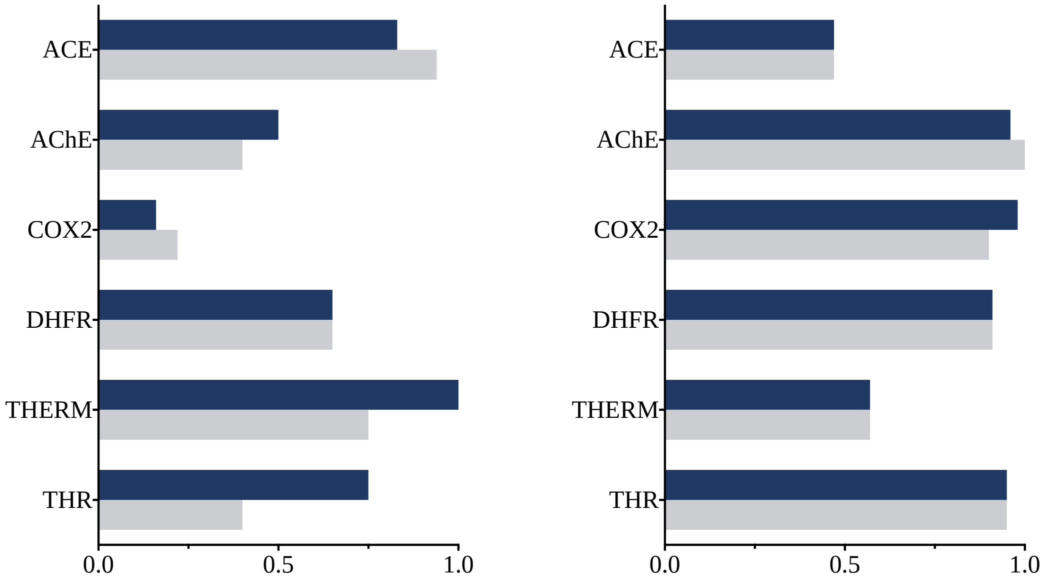


Figure 13 | The specificity, left, and sensitivity, right, for CoMFA/RM1 (grey) and hydrophobic (dark blue) models.

Overall, the results obtained reveal that the hydrophobic descriptors yield 3D-QSAR models with an overall performance that compares with standard CoMFA/CoMSiA results. Moreover, these models may be a valuable measure to rank molecules based on molecular similarity (high sensitivity and Rs).

4.3 Three-dimensional similarity in combination with molecular docking

In the final part of the thesis, the complementarity between the two main groups of techniques that have traditionally divided VS has been evaluated: structure-based and ligand-based methods. It is well known that both methods have inherent limitations that could be overcome by the usage of a hybrid approach.

In structure-based virtual screening, the most used tool has been molecular docking. The balance between predicted accuracy and computational cost is one of its major drawbacks. The consideration of different structural conformations and the use of simplified score functions may limit the accuracy of the ranking score, due to deficiencies in the definitions of enthalpic and entropic contributions to the binding affinity. Consequently, the inclusion of 3D similarity information can be valuable in the identification of new active compounds.

In ligand-based virtual screening, on the other hand, no information of the target is used, giving equal importance to all the regions of the molecules while only a small part could be relevant for the binding mode.

Validation is a sensitive stage dependent on the selected test set. In order to compare the proposed protocol, the Directory of useful decoys (DUD)[192] was used, specifically the set known as DUD_LIB_VS_1.[193] This set addresses the limitations of the original DUD data set to not overestimate the performance of ligand-based methods. In addition, four sets especially compiled to evaluate combined virtual screening methods[194] were added, downloaded directly from the DEKOIS V.2 database.[195] As molecular docking software, Glide[196,197] was used, and as molecular 3D similarity approach PharmScreen was applied, and by extension its hydrophobic descriptors derived from the MST calculations. PharmScreen and Glide results running as a standalone tool are also reported.

4.3.1 Combined score function exploration

Three different protocols were tested. Namely, parallel ranking, rescoring ranking, and consensus ranking.

- Parallel Ranking (PR): ligand-based and structure-based approaches are run independently, and subsequently, the final rankings from the two approaches are combined. The new classification is based on the original ranking position obtained from LB and docking approaches, treating both methods with equal parity. Accordingly, the first molecule of docking ranking and the first molecule of LB ranking will occupy the first and second position or vice versa in the final ranking. The first will be the one with the lowest mark of both ranking positions. Accordingly, the molecules ranked second for each method would be re-ranked third and fourth, and so on until all molecules are reordered. This framework is guided by the assumption that there is no single approach that will provide optimal screening in all circumstances.[198]
- Rescoring Ranking (RR): the final docking ranking is rescored based on the 3D similarity of the best pose obtained for each compound with known co-crystallographic ligands. This protocol aims to propose an alternative to solve the limitations of the scoring function approximations performed in molecular docking[199,200].
- Consensus Ranking (CR): the docking ranking and the rescoring ranking are merged following the same protocol proposed in PR, but in this case, the overlays used by PharmScreen are taken from the docking tool. CR was intended to solve the limitations that RR presents due to the bias introduced toward poses with high overlap between the docking ligand and the template. In large binding sites, a single structure cannot cover all possible binding modes (Figure 14). Accordingly, a consensus between both docking and RR ranking is a convenient proposal to cover the mutual limitations. Ideally, the final ranking will include on the top the poses detected as best by both methods. Although the limitations of the docking score function have been reported,[121,199] due to the existence of multiple bindings modes, there are still scenarios where 3D similarity cannot provide an alternative to evaluate docking poses. A consensus between docking result and the rescored outcome using 3D similarity aims to offer a hybrid scenario to assess the outcome of VS.

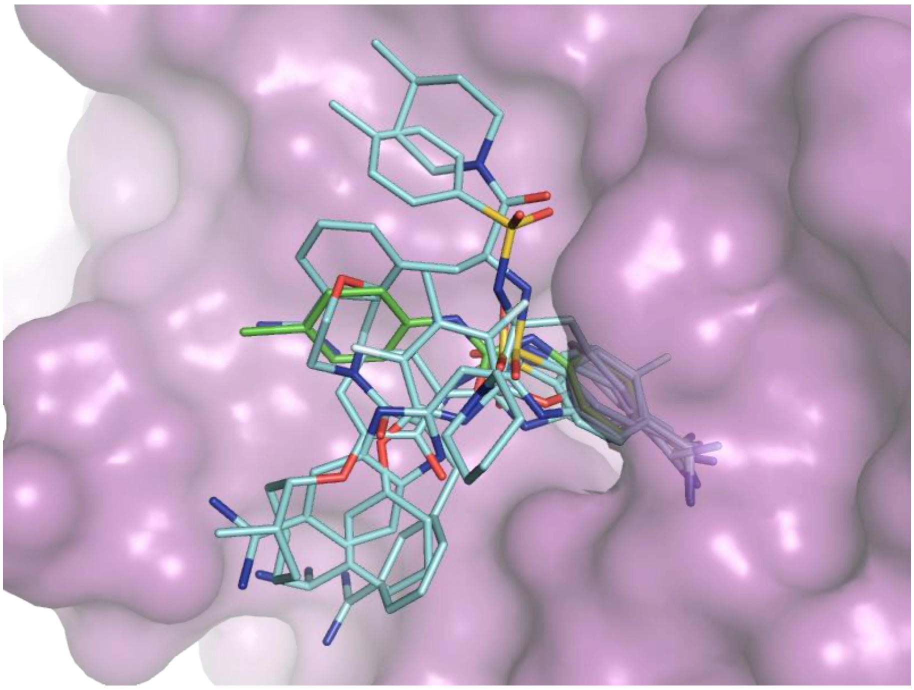


Figure 14 | Binding mode of beta-trypsin (PDB code: 1bju, purple), co-crystallized reference molecule (green), docked molecules using Glide (cyan)

Two similarity coefficients have been examined to score the docked poses of RR and CR: the Tanimoto coefficient (Tn), which is conceived to measure the global similarity (default option in PharmScreen, Tn), and the Tversky coefficient (Tv), which evaluates the partial similarity between the poses.

Although the molecular similarity of pairs of compounds with different size is computed, when only a specific region of the ligand interacts with the receptor, considering the similarity as a whole could introduce noise. As an example, Figure 15 shows a selection of docked ligands by Glide in the binding site of COMT. The overlap of the molecules with the co-crystallized ligand is reduced to the 3-nitrocyclohexane-1,2-diol moiety, making them more convenient to take into account a partial similarity measure.

As an advantage, Tversky coefficient allows focusing the evaluation in the small molecule, highlighting the overlapped section. The improvement in recovering hits using Tv in comparison with Tn for COMT is illustrated in Figure 16.

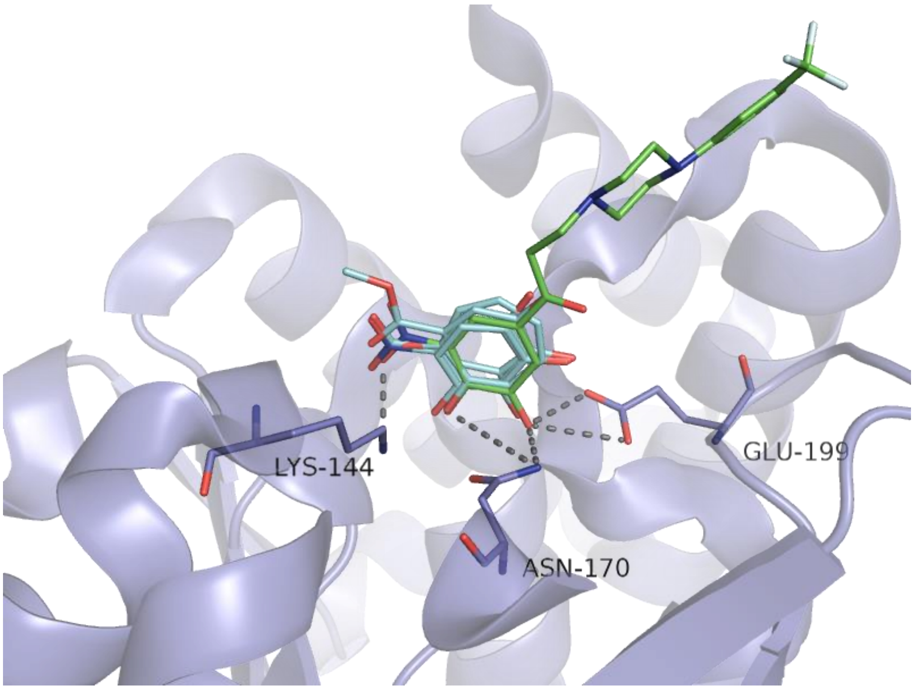


Figure 15 | Reference molecule (green, co-crystallized structure BIA) and 5 docked ligands of COMT set (ZINC03814485, ZINC00392003, ZINC03814484, ZINC00021789, and ZINC00330141) by Glide (cyan) in the binding site of Catecol O-metiltransferasa (blue). 5 possible hydrogen bonds are reported between the reference and lysine 144, asparagine 170 and glutamine 199.

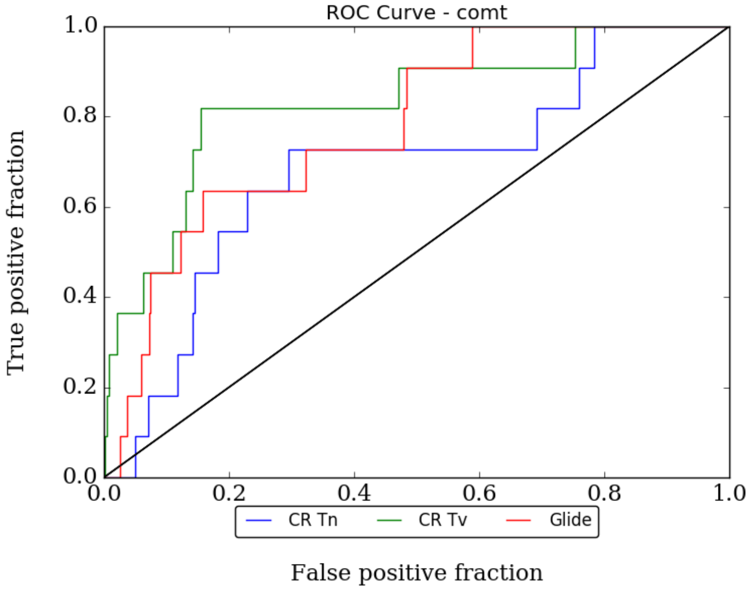


Figure 16 | Comparison of ROC Curves for COMT system performed by Glide and Consensus ranking using Tanimoto (Consensus Tn) and Consensus ranking using Tversky (Consensus Tv).

4.3.2 Performance evaluation

Regarding the three protocols, CR in combination with Tv is the one that recovers a higher number of hits, especially in the first part of the ranking. As shown in Table 3, for all the analyzed metrics (ROCe and AUC), the best combined approach (CR) leads to an increase in the results relative to both PharmScreen and Glide.

Table 3 | AUC and ROCe metrics. The higher values are highlighted in blond. CR and RR are performed using Tv.

	PharmScreen	**Glide**	**PR**	**RR**	**CR**
ROCe 0.5	32.0	28.9	33.5	42.2	**42.7**
ROCe 1	20.2	18.9	22.1	26.5	**27.7**
ROCe 2	12.1	12.2	13.0	16.5	**17.4**
ROCe 5	6.5	6.8	6.7	8.3	**9.4**
AUC	0.66	0.70	0.77	0.71	**0.76**

The ROC curves of pnp (purine nucleoside phosphorylase) and inha (inhibin alpha protein) systems performed by the ligand-based and structure-based methods, and the best combined method (consensus ranking using partial similarity) are illustrated as example in Figure 17.

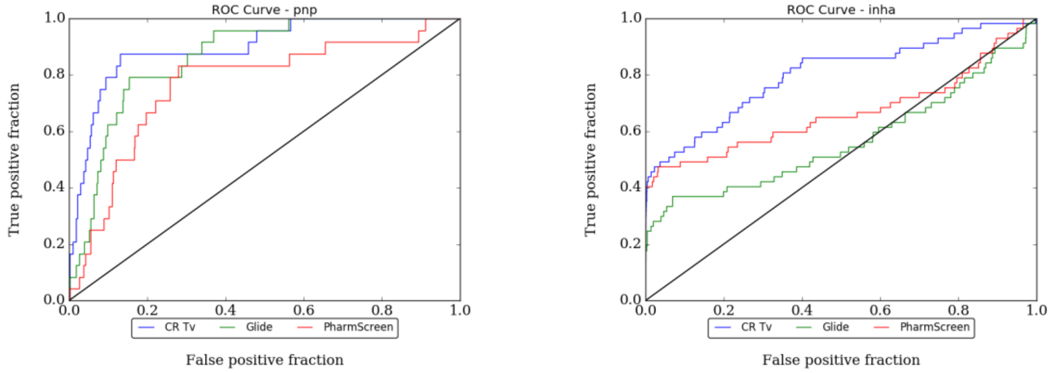


Figure 17 | Comparison of ROC Curves for pnp (purine nucleoside phosphorylase), right, and inha (inhibin alpha chain), left, performed by Glide, PharmScreen and Consensus ranking using Tversky (Consensus Tv).

To further analyze the consistency of the methods, considering the use of Tv for RR and CR, the average of the hierarchical position of each approach among the others is presented in Table 4. As an illustrative case, Figure 18 reports the individual relative position of each method for all sets in ROCe 1%. Consistent with the previous results, CR shows higher robustness among the combined methods. It can be seen that CR is positioned on average as the 2nd best tool, while PharmScreen and Glide alone are positioned on average as the 3rd best approach.

Table 4 | The average of the hierarchical position of each approach among the others. A perfect method that always gets the highest performance gets a value of 1.0. The higher values are highlighted in blond. RR and CR use Tv.

	PharmScreen	**Glide**	**PR**	**RR**	**CR**
ROCe 0.5	3.1	3.3	2.6	**2.3**	**2.3**
ROCe 1	3.2	3.3	2.8	2.4	**2.2**
ROCe 2	3.2	3.5	2.7	**2.2**	2.3
ROCe 5	3.3	3.4	2.6	2.4	**1.9**
AUC	3.8	3.3	2.8	3.0	**2.2**

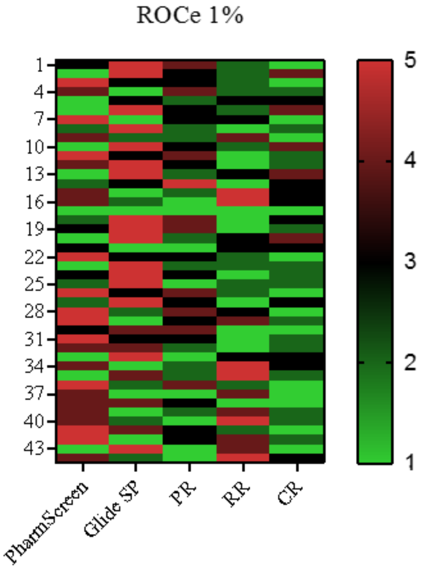


Figure 18 | Heatmap of the hierarchical position of all methods among the others for ROCe 1 %. The color scale is indicative of the position, being the first green and the fifth red. RR and CR use Tv.

5

CHAPTER FIVE
Discussion

5 DISCUSSION

3D LBVS methods have been used for many years in drug discovery, with a variable success depending on different factors, such as the complexity of the target system or the suitability of the molecular descriptors. New approaches are still necessary to cover the broad spectrum of relationships that a drug-like molecule may establish with the organism. In spite of the complexity of processes that modulate the activity of a drug, most tools are primarily focused on the use of shape or electrostatic descriptors. In contrast, since the maximum ligand-receptor binding affinity can be explained mainly by the curvature and apolar surface of the protein-binding site,[109] an exact representation of the 3D pattern of hydrophobic/hydrophilic regions can be a valuable guideline in the construction of a pharmacodynamic profile.

This influence is not only related to the ligand-receptor complex but also to a considerable variety of pharmacokinetic processes. Lipophilicity, therefore, is a crucial parameter to consider in the rational drug design pipeline. To enhance the molecular similarity studies with the hydrophobic/hydrophilic balance, PharmScreen was conceived as a tool to exploit lipophilic 3D similarity. The papers reported in this thesis exemplify the efforts performed to examine the reliability of MST-based hydrophobic descriptors, and the main findings and challenges that arise will be briefly discussed.

The overlays based on the MST contributions to octanol/water partition coefficients and the ability of MST-derived descriptors to predict molecular activity using 3D-QSAR models are the main issues discussed in the first and second publications. Results support the assumption that lipophilicity, supplemented by HB acceptor/donor descriptors, provides a useful signature to enrich the information that can be retrieved from (i) molecular alignment and (ii) QSAR models, complementing the results obtained traditionally from electrostatic and steric properties.

Correct superpositions of 94%, 79%, 54% and 13% of the molecules classified in easy, moderate, difficult and unfeasible sets (AZ), respectively, have been predicted. In addition, there is a low percentage of overlap between the alignments reported by the hydrophobic and traditional descriptors, highlighting the complementarity of both methods and, thus, facilitating the analysis of the growing number of complex systems under investigation. Therefore, plausible molecular overlays are provided by these sets of descriptors throughout alternative information to traditional descriptors.

Along the same lines, as reported in the second paper, the strong resemblance of the statistical values of the regression (r^2 and S) and cross-validation (q^2) disclosed by the standard models (CoMFA and CoMSiA) and our descriptors corroborates the competitive results provided by the atomic decomposition of lipophilicity as 3D-QSAR descriptor. Moreover, $\text{LogP}_{\text{ele}}/\text{LogP}_{\text{cav}}$/HB models also seem to be more effective ranking (high Rs) and locating (high sensibility) true positives, especially in the first quartiles (molecules ordered by activity/affinity).

As a common issue, the performance of both alignment process and 3D-QSAR models from MST/RM1 and MST/B3LYP levels were compared. The similar overall performance of the two QM methods and the lower computational requirement of the RM1 approach make the latter a promising choice for PharmScreen's alignment protocol and 3D-QSAR studies in the exploitation of atomic solvation free energy parameters.

Taken together, the results obtained for the benchmarking sets confirm the usefulness of lipophilicity as a valuable alternative for molecular alignment and structure-activity relationship prediction.

Finally, the applicability of our descriptors in VS has been explored in order to re-evaluate the complexes constituted by docking techniques (in our case, Glide). Since (de)solvation is fundamental for the establishment of the ligand-receptor complex, it can be expected that the docked ligands in the same pocket share lipophilic characteristics, even if there are several binding modes. However, approximations that affect solvation contribution[113] are applied in the docking score functions, and by extension, some docking programs show problems performing VS, especially in hydrophobic binding pockets.[201]

In view of the work presented in the third publication, the $LogP_{ele}$/$LogP_{cav}$/HB similarity is introduced as a valid scoring function for discerning between active and inactive compounds. Specific binding typically requires the formation of key interactions between targets and ligands. Thus, 3D similarity relative to experimental binding modes could be sufficient to distinguish active compounds from decoys. However, multiple binding modes usually exist and, hence, a re-evaluation that computes the similarity from a single query is not always sufficient. Therefore, a consensus between docking ranking and the re-evaluation ranking based on 3D similarity is proposed (CR), which returns a considerable improvement compared to each VS method alone. CR provides an increase in AUC and ROCe metrics, both in performance and in robustness. Although PR and RR (Tv) return lower results than CR, their performance overcomes PharmScreen and Glide as standalone tools. RR and CR improve the success rate, with only a slight increase in time and resources. Nevertheless, the obvious higher computational resources demand for PR should also be considered.

These findings support the usefulness of $LogP_{ele}$/$LogP_{cav}$/HB as relevant descriptors in molecular similarity studies, promoting their use in virtual screening campaigns considering LB approaches or in combination with SB. Indeed, PharmScreen is being tested by national and international companies: Merck, Eli Lilly, Almirall and Esteve amongst others, and two national scientific institutions have acquired a license: CNIO and CIMA.

6

CHAPTER SIX
Conclusions

6 CONCLUSIONS

This thesis comprises both the development and validation of a LBVS tool that exploits 3D atomic contributions to lipophilicity, PharmScreen. This section summarizes the main conclusions found in the course of this work:

- Atomic descriptors of hydrophobicity obtained from Miertus–Scrocco–Tomasi-based continuum solvation models are a valuable alternative to explore novel frameworks in CADD. In fact, the partition of molecular lipophilicity into atomic contributions using the MST/RM1 model provides a three-dimensional lipophilicity pattern of drug-like compounds valuable for molecular similarity studies.

- The similar overall performance of MST/RM1 and MST/B3LYP/6-31G(d) and the lower computational requirement for the former makes MST/RM1 to be a balanced choice for PharmScreen's alignment protocol and 3D-QSAR studies to represent atomic solvation free energy contributions.

- MST-derived Hydrophobic descriptors have demonstrated to be competitive for molecular alignments in comparison to traditional properties, especially for targets that may be challenging for predictive molecular similarity techniques.

- In 3D-QSAR studies, the proposed descriptors provide models for structure-activity relationships with predictive accuracy comparable to CoMFA/CoMSiA models based on electrostatic/steric parameters.

- PharmScreen exhibits a competitive performance as a VS tool compared to the tested docking software. On average, the performance of molecular docking is improved thanks to the similarity from the topological distribution of lipophilicity characteristics between docked ligands and a known active with only a slight increase in time and resources.

- The results obtained from the analysis of hydrophobic/hydrophilic descriptors presented in this thesis opens a new window to explore the vast chemical space, complementing the information derived from traditional descriptors in ligand- and structure-based approaches.

BIBLIOGRAPHY

Bibliography

(1) Nielsen, A. K. The Periodic Table: Its Story and Its Significance - by Eric R. Scerri. *Centaurus* **2008**, *50* (4), 339–341.

(2) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B. T. Molecular Similarity and Diversity in Chemoinformatics: From Theory to Applications. *Mol. Divers.* **2006**, *10* (1), 39–79.

(3) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-Art in Ligand-Based Virtual Screening. *Drug Discov. Today* **2011**, *16* (9–10), 372–376.

(4) Forti, F.; Barril, X.; Luque, F. J.; Orozco, M. Extension of the MST Continuum Solvation Model to the RM1 Semiempirical Hamiltonian. *J. Comput. Chem.* **2008**, *29* (4), 578–587.

(5) Luque, F. J.; Curutchet, C.; Muñoz-Muriedas, J.; Bidon-Chanal, A.; Soteras, I.; Morreale, A.; Gelpí, J. L.; Orozco, M. Continuum Solvation Models: Dissecting the Free Energy of Solvation. *Phys. Chem. Chem. Phys.* **2003**, *5* (18), 3827–3836.

(6) Barril, X.; Muñoz, J.; Luque, F. J.; Orozco, M. Simplified Descriptions of the Topological Distribution of Hydrophilic/Hydrophobic Characteristics of Molecules. *Phys. Chem. Chem. Phys.* **2000**, *2* (21), 4897–4905.

(7) Curutchet, C.; Orozco, M.; Luque, F. J. Solvation in Octanol: Parametrization of the Continuum MST Model. *J. Comput. Chem.* **2001**, *22* (11), 1180–1193.

(8) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Luque, F. J. Extension of the MST Model to the IEF Formalism: HF and B3LYP Parametrizations. *J. Mol. Struct. THEOCHEM* **2005**, *727* (1–3), 29–40.

(9) Vázquez, J.; Deplano, A.; Herrero, A.; Ginex, T.; Gibert, E.; Rabal, O.; Oyarzabal, J.; Herrero, E.; Luque, F. J. Development and Validation of Molecular Overlays Derived from Three-Dimensional Hydrophobic Similarity with PharmScreen. *J. Chem. Inf. Model.* **2018**, *58* (8), 1596–1609.

(10) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2* (22), 3204–3218.

(11) Bajorath, J. Molecular Similarity Concepts for Informatics Applications. In *Methods in Molecular Biology*; Humana Press, New York, NY, **2017**; Vol. 1526, pp 231–245.

(12) Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discov. Des.* **1998**, *9*, 225–252.

(13) Maggiora, G. M.; Vogt, M.; Stumpfe, D.; Bajorath, J. J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2013**, *57* (8), 3186–3204.

(14) Klopmand, G. Concepts and Applications of Molecular Similarity. *J. Comput. Chem.* **1992**, *13* (4), 539–540.

(15) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B. T. Molecular Similarity and Diversity in Chemoinformatics: From Theory to Applications. *Mol. Divers.* **2006**, *10* (1), 39–79.

(16) Kopp, H. Annalen Der Chemie Und Pharm. *Ann. der Chemie und pharm* **1842**, *41*, 79.

(17) Hansch, C.; Fujita, T. *P* -σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616–1626.

(18) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Model.* **1985**, *25* (2), 64–73.

(19) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nearest-Neighbor Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Model.* **1986**, *26* (1), 36–41.

(20) Macalino, S. J. Y.; Gosu, V.; Hong, S.; Choi, S. Role of Computer-Aided Drug Design in Modern Drug Discovery. *Arch. Pharm. Res.* **2015**, *38* (9), 1686–1701.

(21) Aouidate, A.; Ghaleb, A.; Ghamali, M.; Chtita, S.; Ousaa, A.; Choukrad, M.; Sbai, A.; Bouachrine, M.; Lakhlifi, T. QSAR Study and Rustic Ligand-Based Virtual Screening in a

Search for Aminooxadiazole Derivatives as PIM1 Inhibitors. *Chem. Cent. J.* **2018**, *12* (1), 32.

(22) Erlina, L.; Yanuar, A. Pharmacophore-Based Virtual Screening from Indonesian Herbal Database to Finding New Inhibitor of HDAC4 and HDAC7. *J. Young Pharm.* **2018**, *10* (1), 7–11.

(23) Jin, X.; Kwon, W.; Kim, T. S.; Heo, J.-N.; Chung, H. C.; Choi, J.; No, K. T. Identification of Natural Products as Novel PI3Kβ Inhibitors Through Pharmacophore-Based Virtual Screening. *Bull. Korean Chem. Soc.* **2018**, *39* (3), 294–299.

(24) Al-Sha'er, M. A.; Taha, M. O. Ligand-Based Modeling of Akt3 Lead to Potent Dual Akt1/Akt3 Inhibitor. *J. Mol. Graph. Model.* **2018**, *83*, 153–166.

(25) Che, J.; Wang, Z.; Sheng, H.; Huang, F.; Dong, X.; Hu, Y.; Xie, X.; Hu, Y. Ligand-Based Pharmacophore Model for the Discovery of Novel CXCR2 Antagonists as Anti-Cancer Metastatic Agents. *R. Soc. Open Sci.* **2018**, *5* (7), 180176.

(26) Saroj Devi, N.; Shanmugam, R.; Ghorai, J.; Ramanan, M.; Anbarasan, P.; Doble, M. Ligand-Based Modeling for the Prediction of Pharmacophore Features for Multi-Targeted Inhibition of the Arachidonic Acid Cascade. *Mol. Inform.* **2018**, *37* (3), 1700073.

(27) Grisoni, F.; Merk, D.; Consonni, V.; Hiss, J. A.; Tagliabue, S. G.; Todeschini, R.; Schneider, G. Scaffold Hopping from Natural Products to Synthetic Mimetics by Holistic Molecular Similarity. *Commun. Chem.* **2018**, *1* (1), 44.

(28) Beccari, A. R.; Gemei, M.; Lo Monte, M.; Menegatti, N.; Fanton, M.; Pedretti, A.; Bovolenta, S.; Nucci, C.; Molteni, A.; Rossignoli, A.; et al. Publisher Correction: Novel Selective, Potent Naphthyl TRPM8 Antagonists Identified through a Combined Ligand- and Structure-Based Virtual Screening Approach. *Sci. Rep.* **2018**, *8* (1), 4250.

(29) Dietrich, R. C.; Alberca, L. N.; Ruiz, M. D.; Palestro, P. H.; Carrillo, C.; Talevi, A.; Gavernet, L. Identification of Cisapride as New Inhibitor of Putrescine Uptake in Trypanosoma Cruzi by Combined Ligand- and Structure-Based Virtual Screening. *Eur. J. Med. Chem.* **2018**, *149*, 22–29.

(30) Yu, M.; Gu, Q.; Xu, J. Discovering New PI3Kα Inhibitors with a Strategy of Combining Ligand-Based and Structure-Based Virtual Screening. *J. Comput. Aided. Mol. Des.* **2018**, *32* (2), 347–361.

(31) Martínez-Muñoz, A.; Prestegui-Martel, B.; Méndez-Luna, D.; Fragoso-Vázquez, M. J.; García-Sánchez, J. R.; Bello, M.; Martínez-Archundia, M.; Chávez-Blanco, A.; Dueñas-González, A.; Mendoza-Lujambio, I.; et al. Selection of a GPER1 Ligand via Ligand-Based Virtual Screening Coupled to Molecular Dynamics Simulations and Its Anti-Proliferative Effects on Breast Cancer Cells. *Anticancer. Agents Med. Chem.* **2018**, *18* (11), 1629–1638.

(32) Honegr, J.; Dolezal, R.; Malinak, D.; Benkova, M.; Soukup, O.; Almeida, J.; Franca, T.; Kuca, K.; Prymula, R.; Honegr, J.; et al. Rational Design of a New Class of Toll-Like Receptor 4 (TLR4) Tryptamine Related Agonists by Means of the Structure- and Ligand-Based Virtual Screening for Vaccine Adjuvant Discovery. *Molecules* **2018**, *23* (1), 102.

(33) Patel, S.; Modi, P.; Chhabria, M. Rational Approach to Identify Newer Caspase-1 Inhibitors Using Pharmacophore Based Virtual Screening, Docking and Molecular Dynamic Simulation Studies. *J. Mol. Graph. Model.* **2018**, *81*, 106–115.

(34) K, R.; V, S. Discovery of Potent Neuraminidase Inhibitors Using a Combination of Pharmacophore-Based Virtual Screening and Molecular Simulation Approach. *Appl. Biochem. Biotechnol.* **2018**, *184* (4), 1421–1440.

(35) Sepehri, B.; Ghavami, R. The Identification of New CD38 Inhibitors by Combined Structure and Ligand Based Virtual Screening Approaches of ZINC Database. *Lett. Drug Des. Discov.* **2018**, *15* (6), 654–660.

(36) Ramezani, M.; Shamsara, J. An Integrated Structure- and Pharmacophore-Based MMP-12 Virtual Screening. *Mol. Divers.* **2018**, *22* (2), 383–395.

(37) James, N.; Ramanathan, K. Ligand-Based Pharmacophore Screening Strategy: A Pragmatic Approach for Targeting HER Proteins. *Appl. Biochem. Biotechnol.* **2018**, *186* (1), 85–108.

(38) Hawkins, P. C. D.; Stahl, G. Ligand-Based Methods in GPCR Computer-Aided Drug Design. In *Methods in Molecular Biology*; **2018**; Vol. 1705, pp 365–374.

(39) Floris, M.; Olla, S. Molecular Similarity in Computational Toxicology. In *Methods in Molecular Biology*; Humana Press, New York, NY, **2018**; Vol. 1800, pp 171–179.

(40) Fischer, J.; Rotella, D. P. *Successful Drug Discovery*; Fischer, J., Klein, C., Childers, W. E., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, **2015**.

(41) Vansal, S. S.; Feller, D. R. Direct Effects of Ephedrine Isomers on Human Beta-Adrenergic Receptor Subtypes. *Biochem. Pharmacol.* **1999**, *58* (5), 807–810.

(42) Abourashed, E. A.; El-Alfy, A. T.; Khan, I. A.; Walker, L. Ephedra in Perspective - a Current Review. *Phyther. Res.* **2003**, *17* (7), 703–712.

(43) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57* (8), 3186–3204.

(44) Bajorath, J. Virtual Screening: Methods, Expectations, and Reality. *Curr. Drug Discov.* **2002**, *2*, 24–28.

(45) Wintner, E. A.; Moallemi, C. C. Quantized Surface Complementarity Diversity (QSCD): A Model Based Oil Small Molecule-Target Complementarity. *J. Med. Chem.* **2000**, *43* (10), 1993–2006.

(46) Bath, P. A.; Morris, C. A.; Willett, P. Effect of Standardization on Fragment-Based Measures of Structural Similarity. *J. Chemom.* **1993**, *7* (6), 543–550.

(47) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41* (18), 3325–3329.

(48) Cuissart, B.; Touffet, F.; Crémilleux, B.; Bureau, R.; Rault, S. The Maximum Common Substructure as a Molecular Depiction in a Supervised Classification Context: Experiments in Quantitative Structure/Biodegradability Relationships. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (5), 1043–1052.

(49) Wiswesser, W. J. 107 Years of Line-Formula Notations (1861-1968). *J. Chem. Doc.* **1968**, *8* (3), 146–150.

(50) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31–36.

(51) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (1), 71–79.

(52) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method to Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45* (2), 386–393.

(53) Jørgensen, A. M. M.; Pedersen, J. T. Structural Diversity of Small Molecule Libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 338–345.

(54) Ivanciuc, O.; Taraviras, S. L.; Cabrol-Bass, D. Quasi-Orthogonal Basis Sets of Molecular Graph Descriptors as a Chemical Diversity Measure. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (1), 126–134.

(55) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods. *J. Mol. Graph. Model.* **2010**, *29* (2), 157–170.

(56) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 338–345.

(57) Shin, W. H.; Zhu, X.; Bures, M. G.; Kihara, D. Three-Dimensional Compound Comparison Methods and Their Application in Drug Discovery. *Molecules* **2015**, *20* (7), 12841–12862.

(58) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. Toward High Throughput 3D Virtual Screening Using Spherical Harmonic Surface Representations. *J. Chem. Inf. Model.* **2007**, *47* (5), 1787–1796.

(59) Sheridan, R. P.; Kearsley, S. K. Why Do We Need so Many Chemical Similarity Search Methods? *Drug Discov. Today* **2002**, *7* (17), 903–911.

(60) Matter, H.; Pötter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 1211–1225.

(61) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *40* (2), 295–307.

(62) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal Assignment Methods for Ligand-Based Virtual Screening. *J. Cheminform.* **2009**, *1* (1), 14.

(63) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996.

(64) Muegge, I.; Mukherjee, P. An Overview of Molecular Fingerprint Similarity Search in Virtual Screening. *Expert Opin. Drug Discov.* **2016**, *11* (2), 137–148.

(65) Anighoro, A.; Pinzi, L.; Rastelli, G.; Bajorath, J. Virtual Screening for Dual Hsp90/B-Raf Inhibitors; Humana Press, New York, NY, **2017**; pp 355–365.

(66) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model* **2012**, *52*, 881.

(67) Chen, H.; Kogej, T.; Engkvist, O. Cheminformatics in Drug Discovery, an Industrial Perspective. *Mol. Inform.* **2018**, *37* (9–10), 1800041.

(68) Passeri, G. I.; Trisciuzzi, D.; Alberga, D.; Siragusa, L.; Leonetti, F.; Mangiatordi, G. F.; Nicolotti, O. Strategies of Virtual Screening in Medicinal Chemistry. *Int. J. Quant. Struct. Relationships* **2018**, *3* (1), 134–160.

(69) Hopfinger, A. J. A QSAR Investigation of Dihydrofolate Reductase Inhibition by Baker Triazines Based upon Molecular Shape Analysis. *J. Am. Chem. Soc.* **1980**, *102* (24), 7196–7206.

(70) Cross, S.; Cruciani, G. Molecular Fields in Drug Discovery: Getting Old or Reaching Maturity? *Drug Discov. Today* **2010**, *15* (2).

(71) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110* (18), 5959–5967.

(72) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37* (24), 4130–4146.

(73) James, T. Cheminformatics in the Service of GPCR Drug Discovery. In *Methods in Molecular Biology*; Humana Press, New York, NY, **2018**; Vol. 1705, pp 395–411.

(74) Manoharan, P.; Ghoshal, N. Computational Modeling of Gamma-Secretase Inhibitors as Anti-Alzheimer Agents. In *Neuromethods*; Humana Press, New York, NY, **2018**; Vol. 132, pp 283–303.

(75) Correa, C. M. Structure Activity of CB1 Cannabinoid Receptor Antagonists. *Curr. Top. Med. Chem.* **2018**, *8* (3), 53–62.

(76) Cozza, G.; Cozza; Giorgio. The Development of CK2 Inhibitors: From Traditional Pharmacology to in Silico Rational Drug Design. *Pharmaceuticals* **2017**, *10* (4), 26.

(77) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 1912–1928.

(78) Ballester, P. J.; Richards, W. G. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* **2007**, *28* (10), 1711–1723.

(79) Schreyer, A. M.; Blundell, T. USRCAT: Real-Time Ultrafast Shape Recognition with Pharmacophoric Constraints. *J. Cheminform.* **2012**, *4* (11), 1–12.

(80) Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Moretti, L.; Cooper, R. I.; Richards, W. G. ElectroShape: Fast Molecular Similarity Calculations Incorporating Shape, Chirality and Electrostatics. *J. Comput. Aided. Mol. Des.* **2010**, *24* (9), 789–801.

(81) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50* (1), 74–82.

(82) ROCS: OpenEye Scientific Software, Santa Fe, NM. Hhtp://Www.Eyesopen.Com.

(83) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential. *J. Chem. Inf. Model.* **2009**, *49* (2), 492–502.

(84) Liu, X.; Jiang, H.; Li, H. SHAFTS: A Hybrid Approach for 3D Molecular Similarity Calculation. 1. Method and Assessment of Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51* (9), 2372–2385.

(85) Roy, A.; Skolnick, J. LIGSIFT: An Open-Source Tool for Ligand Structural Alignment and Virtual Screening. *Bioinformatics* **2015**, *31* (4), 539–544.

(86) Sun, H. Pharmacophore-Based Virtual Screening. *Curr. Med. Chem.* **2008**, *15* (10), 1018–1024.

(87) Moser, D.; Wittmann, S. K.; Kramer, J.; Blöcher, R.; Achenbach, J.; Pogoryelov, D.; Proschak, E. PENG: A Neural Gas-Based Approach for Pharmacophore Elucidation. Method Design, Validation, and Virtual Screening for Novel Ligands of LTA4H. *J. Chem. Inf. Model.* **2015**, *55* (2), 284–293.

(88) Tosco, P.; Balle, T.; Shiri, F. Open3DALIGN: An Open-Source Software Aimed at Unsupervised Ligand Alignment. *J. Comput. Aided. Mol. Des.* **2011**, *25* (8), 777–783.

(89) Sastry, G. M.; Dixon, S. L.; Sherman, W. Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlap Scoring. *J. Chem. Inf. Model.* **2011**, *51* (10), 2455–2466.

(90) Phase, Schrödinger, LLC., New York, NY, 2009.

(91) Jain, A. N. Ligand-Based Structural Hypotheses for Virtual Screening. *J. Med. Chem.* **2004**, *47* (4), 947–961.

(92) Hofbauer, C.; Lohninger, H.; Aszódi, A. SURFCOMP: A Novel Graph-Based Approach to Molecular Surface Comparison. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 837–847.

(93) Artese, A.; Cross, S.; Costa, G.; Distinto, S.; Parrotta, L.; Alcaro, S.; Ortuso, F.; Cruciani, G. Molecular Interaction Fields in Drug Discovery: Recent Advances and Future Perspectives. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3* (6), 594–613.

(94) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen : Virtual Screening Using Molecular Fields . Application to the DUD Data Set FieldScreen : Virtual Screening Using Molecular Fields . Application to the DUD Data Set. *J. Chem. Inf. Model* **2008**, *48* (11), 2108–2117.

(95) Blaze, Version , Cresset®, Litlington, Cambridgeshire, UK, ; Http://Www.Cresset-Group.Com/Blaze/; Cheeseright, T.J.; Mackey,.

(96) Tervo, A. J.; Rönkkö, T.; Nyrönen, T. H.; Poso, A. BRUTUS: Optimization of a Grid-Based Similarity Function for Rigid-Body Molecular Superposition. 1. Alignment and Virtual Screening Applications. *J. Med. Chem.* **2005**, *48* (12), 4076–4086.

(97) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. MIMIC: A Molecular-Field Matching Program. Exploiting Applicability of Molecular Similarity Approaches. *J. Comput. Chem.* **1997**, *18* (7), 934–954.

(98) Yang, S. Y. Pharmacophore Modeling and Applications in Drug Discovery: Challenges and Recent Advances. *Drug Discov. Today* **2010**, *15* (11–12), 444–450.

(99) Cross, S.; Baroni, M.; Carosati, E.; Benedetti, P.; Clementi, S. FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation Using the DUD Data Set. *J. Chem. Inf. Model.* **2010**, *50* (8), 1442–1450.

(100) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849–857.

(101) Fox, P. C.; Wolohan, P. R. N.; Abrahamian, E.; Clark, R. D. Parameterization and Conformational Sampling Effects in Pharmacophore Multiplet Searching. *J. Chem. Inf. Model.* **2008**, *48* (12), 2326–2334.

(102) Abrahamian, E.; Fox, P. C.; Naerum, L.; Christensen, I. T.; Thøgersen, H.; Clark, R. D. Efficient Generation, Storage, and Manipulation of Fully Flexible Pharmacophore Multiplets and Their Use in 3-D Similarity Searching. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 458–468.

(103) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. MIMIC: A Molecular-Field Matching Program. Exploiting Applicability of Molecular Similarity Approaches. *J. Comput. Chem.* **1997**, *18* (7), 934–954.

(104) Abrahamian, E.; Fox, P. C.; Nærum, L.; Thøger Christensen, I.; Thøgersen, H.; Clark, R. D. Efficient Generation, Storage, and Manipulation of Fully Flexible Pharmacophore Multiplets and Their Use in 3-D Similarity Searching. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 458–468.

(105) Kumar, A.; Zhang, K. Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front. Chem.* **2018**, *6* (6), 315.

(106) Pliška, V.; Testa, B.; van de Waterbeemd, H. *Lipophilicity in Drug Action and Toxicology*; VCH, **2008**; Vol. 4.

(107) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of LogP Methods on More than 96,000 Compounds. *J. Pharm. Sci.* **2009**, *98* (3), 861–893.

(108) Tsopelas, F.; Giaginis, C.; Tsantili-Kakoulidou, A. Lipophilicity and Biomimetic Properties to Support Drug Discovery. *Expert Opin. Drug Discov.* **2017**, *12* (9), 885–896.

(109) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-Based Maximal Affinity Model Predicts Small-Molecule Druggability. *Nat. Biotechnol.* **2007**, *25* (1), 71–75.

(110) Volkamer, A.; von Behren, M. M.; Bietz, S.; Rarey, M. Prediction, Analysis, and Comparison of Active Sites. In *Applied Chemoinformatics*; **2018**; pp 283–311.

(111) Lavecchia, A.; Di Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.* **2013**, *20* (23), 2839–2860.

(112) Macalino, S. J. Y.; Gosu, V.; Hong, S.; Choi, S. Role of Computer-Aided Drug Design in Modern Drug Discovery. *Arch. Pharm. Res.* **2015**, *38* (9), 1686–1701.

(113) Hein, M.; Zilian, D.; Sotriffer, C. A. Docking Compared to 3D-Pharmacophores: The Scoring Function Challenge. *Drug Discov. Today Technol.* **2010**, *7* (4), e229–e236.

(114) Wilson, G. L.; Lill, M. A. Integrating Structure-Based and Ligand-Based Approaches for Computational Drug Design. *Future Med. Chem.* **2011**, *3* (6), 735–750.

(115) Anighoro, A.; Bajorath, J. Three-Dimensional Similarity in Molecular Docking: Prioritizing Ligand Poses on the Basis of Experimental Binding Modes. *J. Chem. Inf. Model.* **2016**, *56* (3), 580–587.

(116) Banoglu, E.; Çalışkan, B.; Luderer, S.; Eren, G.; Özkan, Y.; Altenhofen, W.; Weinigel, C.; Barz, D.; Gerstmeier, J.; Pergola, C.; et al. Identification of Novel Benzimidazole Derivatives as Inhibitors of Leukotriene Biosynthesis by Virtual Screening Targeting 5-Lipoxygenase-Activating Protein (FLAP). *Bioorg. Med. Chem.* **2012**, *20* (12), 3728–3741.

(117) Smith, J. R.; Evans, K. J.; Wright, A.; Willows, R. D.; Jamie, J. F.; Griffith, R. Novel Indoleamine 2,3-Dioxygenase-1 Inhibitors from a Multistep in Silico Screen. *Bioorg. Med. Chem.* **2012**, *20* (3), 1354–1363.

(118) Drwal, M. N.; Agama, K.; Wakelin, L. P. G.; Pommier, Y.; Griffith, R. Exploring DNA Topoisomerase I Ligand Space in Search of Novel Anticancer Agents. *PLoS One* **2011**, *6* (9), e25150.

(119) Swann, S. L.; Brown, S. P.; Muchmore, S. W.; Patel, H.; Merta, P.; Locklear, J.; Hajduk, P. J. A Unified, Probabilistic Framework for Structure- and Ligand-Based Virtual Screening. *J. Med. Chem.* **2011**, *54* (5), 1223–1232.

(120) Larsson, M.; Fraccalvieri, D.; Andersson, C. D.; Bonati, L.; Linusson, A.; Andersson, P. L. Identification of Potential Aryl Hydrocarbon Receptor Ligands by Virtual Screening of Industrial Chemicals. *Environ. Sci. Pollut. Res.* **2018**, *25* (3), 2436–2449.

(121) 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* (4), 943–955.

(122) Anighoro, A.; Bajorath, J. A Hybrid Virtual Screening Protocol Based on Binding Mode Similarity. In *Rational Drug Design*; Humana Press, New York, NY, 2018; pp 165–175.

(123) Sperandio, O.; Miteva, M.; Villoutreix, B. Combining Ligand- and Structure-Based Methods in Drug Design Projects. *Curr. Comput. Aided-Drug Des.* **2008**, *4* (3), 250–258.

(124) Sutter, J.; Li, J.; Maynard, A. J.; Goupil, A.; Luu, T.; Nadassy, K. New Features That Improve the Pharmacophore Tools from Accelrys. *Curr. Comput. Aided. Drug Des.* **2011**, *7* (3), 173–180.

(125) Anighoro, A.; Bajorath, J. A Hybrid Virtual Screening Protocol Based on Binding Mode Similarity. In *Methods in Molecular Biology*; Humana Press, New York, NY, **2018**; Vol. 1824, pp 165–176.

(126) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39* (11), 2129–2140.

(127) Naray-Szabo, G.; Ferenczy, G. G. Molecular Electrostatics. *Chem. Rev.* **1995**, *95* (4), 829–847.

(128) Platt, D. E.; Silverman, B. D. Registration, Orientation, and Similarity of Molecular Electrostatic Potentials through Multipole Matching. *J. Comput. Chem.* **1996**, *17* (3), 358–366.

(129) Moretti, L.; Graham Richards, W. Molecular Alignment Using Multipole Moments. *Bioorg. Med. Chem. Lett.* **2010**, *20* (19), 5887–5890.

(130) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99* (11), 3503–3510.

(131) Grant, J. a.; Gallardo, M. a.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *17* (14), 1653–1666.

(132) Nicholls, A.; MacCuish, N. E.; MacCuish, J. D. Variable Selection and Model Validation of 2D and 3D Molecular Descriptors&gt; *J. Comput. Aided. Mol. Des.* **2004**, *18* (7–9), 451–474.

(133) Carbó, R.; Leyda, L.; Arnau, M. How Similar Is a Molecule to Another? An Electron Density Measure of Similarity between Two Molecular Structures. *Int. J. Quantum Chem.* **1980**, *17* (6), 1185–1189.

(134) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110* (18), 5959–5967.

(135) Arkin, M. R.; Wells, J. A. Small-Molecule Inhibitors of Protein–Protein Interactions: Progressing towards the Dream. *Nat. Rev. Drug Discov.* **2004**, *3* (4), 301–317.

(136) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *J. Med. Chem.* **2005**, *48* (7), 2518–2525.

(137) Nayal, M.; Honig, B. On the Nature of Cavities on Protein Surfaces: Application to the Identification of Drug-Binding Sites. *Proteins Struct. Funct. Bioinforma.* **2006**, *63* (4), 892–906.

(138) Egner, U.; Hillig, R. C. A Structural Biology View of Target Drugability. *Expert Opin. Drug Discov.* **2008**, *3* (4), 391–401.

(139) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-Based Maximal Affinity Model Predicts Small-Molecule Druggability. *Nat. Biotechnol.* **2007**, *25* (1), 71–75.

(140) Schmidtke, P.; Barril, X. Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *J. Med. Chem.* **2010**, *53* (15), 5858–5867.

(141) Schmidtke, P.; Luque, F. J.; Murray, J. B.; Barril, X. Shielded Hydrogen Bonds as Structural Determinants of Binding Kinetics: Application in Drug Design. *J. Am. Chem. Soc.* **2011**,

133 (46), 18903–18910.

(142) Alvarez-Garcia, D.; Barril, X. Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. *J. Med. Chem.* **2014**, *57* (20), 8530–8539.

(143) Davis, A. M.; Teague, S. J. Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angew. Chemie Int. Ed.* **1999**, *38* (6), 736–749.

(144) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Model.* **1989**, *29* (3), 163–172.

(145) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. An Estimation of the Atomic Contribution to Octanol-Water Partition Coefficient and Molar Refractivity from Fundamental Atomic and Structural Properties: Its Uses in Computer Aided Drug Design. *Math. Comput. Model.* **1990**, *14* (C), 505–510.

(146) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 868–873.

(147) Gaillard, P.; Carrupt, P. A.; Testa, B.; Boudon, A. Molecular Lipophilicity Potential, a Tool in 3D QSAR: Method and Applications. *J. Comput. Aided. Mol. Des.* **1994**, *8* (2), 83–96.

(148) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA. *J. Comput. Aided. Mol. Des.* **1991**, *5* (6), 545–552.

(149) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43* (20), 3714–3717.

(150) Nayyar, A.; Malde, A.; Jain, R.; Coutinho, E. 3D-QSAR Study of Ring-Substituted Quinoline Class of Anti-Tuberculosis Agents. *Bioorganic Med. Chem.* **2006**, *14* (3), 847–856.

(151) Ahmed, M. H.; Spyrakis, F.; Cozzini, P.; Tripathi, P. K.; Mozzarelli, A.; Scarsdale, J. N.; Safo, M. A.; Kellogg, G. E. Bound Water at Protein-Protein Interfaces: Partners, Roles and Hydrophobic Bubbles as a Conserved Motif. *PLoS One* **2011**, *6* (9), e24712.

(152) Rogers, K. S.; Cammarata, A. Superdelocalizability and Charge Density. A Correlation with Partition Coefficients. *J. Med. Chem.* **1969**, *12* (4), 692–693.

(153) Rogers, K. S.; Cammarata, A. A Molecular Orbital Description of the Partitioning of Aromatic Compounds between Polar and Nonpolar Phases. *Biochim. Biophys. Acta - Biomembr.* **1969**, *193* (1), 22–29.

(154) Bodor, N.; Gabanyi, Z.; Wong, C. K. A New Method for the Estimation of Partition Coefficient. *J. Am. Chem. Soc.* **1989**, *111* (11), 3783–3786.

(155) Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. Prediction of the N-Octanol/Water Partition Coefficient, LogP, Using a Combination of Semiempirical MO-Calculations and a Neural Network. *J. Mol. Model.* **1997**, *3* (3), 142–155.

(156) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR Models with Error Estimation: Vapor Pressure and LogP. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (4), 1046–1051.

(157) Du, Q.; Arteca, G. A.; Mezey, P. G. Heuristic Lipophilicity Potential for Computer-Aided Rational Drug Design. *J. Comput. Aided. Mol. Des.* **1997**, *11* (5), 503–515.

(158) Du, Q.; Liu, P. J.; Mezey, P. G. Theoretical Derivation of Heuristic Molecular Lipophilicity Potential: A Quantum Chemical Description for Molecular Solvation. *J. Chem. Inf. Model.* **2005**, *45* (2), 347–353.

(159) Cramer, C. J.; Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999**, *99* (8), 2161–2200.

(160) Orozco, M.; Luque, F. J. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chem. Rev.* **2000**, *100* (11), 4187–4225.

(161) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105* (8), 2999–3093.

(162) Ginex, T.; Muñoz-Muriedas, J.; Herrero, E.; Gibert, E.; Cozzini, P.; Luque, F. J. Development and Validation of Hydrophobic Molecular Fields Derived from the Quantum Mechanical IEF/PCM-MST Solvation Models in 3D-QSAR. *J. Comput. Chem.* **2016**, *37* (13), 1147–1162.

(163) Ginex, T.; Muñoz-Muriedas, J.; Herrero, E.; Gibert, E.; Cozzini, P.; Luque, F. J. Application of the Quantum Mechanical IEF/PCM-MST Hydrophobic Descriptors to Selectivity in Ligand Binding. *J. Mol. Model.* **2016**, *22* (6), 1–15.

(164) Vázquez, J.; Deplano, A.; Herrero, A.; Ginex, T.; Gibert, E.; Rabal, O.; Oyarzabal, J.; Herrero, E.; Luque, F. J. Development and Validation of Molecular Overlays Derived from Three-Dimensional Hydrophobic Similarity with PharmScreen. *J. Chem. Inf. Model.* **2018**, *58* (8), 1596–1609.

(165) Tomasi, J.; Persico, M. Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chem. Rev.* **1994**, *94* (7), 2027–2094.

(166) Miertuš, S.; Scrocco, E.; Tomasi, J. Electrostatic Interaction of a Solute with a Continuum. A Direct Utilizaion of Ab Initio Molecular Potentials for the Prevision of Solvent Effects. *Chem. Phys.* **1981**, *55* (1), 117–129.

(167) Miertuš̃, S.; Tomasi, J. Approximate Evaluations of the Electrostatic Free Energy and Internal Energy Changes in Solution Processes. *Chem. Phys.* **1982**, *65* (2), 239–245.

(168) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105* (8), 2999–3093.

(169) Luque, F. J.; Bachs, M.; Orozco, M. An Optimized AM1/MST Method for the MST-SCRF Representation of Solvated Systems. *J. Comput. Chem.* **1994**, *15* (8), 847–857.

(170) Bachs, M.; Luque, F. J.; Orozco, M. Optimization of Solute Cavities and van Der Waals Parameters Inab Initio MST-SCRF Calculations of Neutral Molecules. *J. Comput. Chem.* **1994**, *15* (4), 446–454.

(171) Pierotti, R. A. A Scaled Particle Theory of Aqueous and Nonaqueous Solutions. *Chem. Rev.* **1976**, *76* (6), 717–726.

(172) Claverie, P.; Daudey, J. P.; Langlet, J.; Pullman, B.; Piazzola, D.; Huron, M. J. Studies of Solvent Effects. 1. Discrete, Continuum, and Discrete-Continuum Models and Their Comparison for Some Simple Cases: Ammonium(1+) Ion, Methanol, and Substituted Ammonium(1+) Ion. *J. Phys. Chem.* **1978**, *82* (4), 405–418.

(173) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Javier Luque, F. Extension of the MST Model to the IEF Formalism: HF and B3LYP Parametrizations. *J. Mol. Struct. THEOCHEM* **2005**, *727* (1–3), 29–40.

(174) Luque, F. J.; Barril, X.; Orozco, M. Fractional Description of Free Energies of Solvation. *J. Comput. Aided. Mol. Des.* **1999**, *13* (2), 139–152.

(175) Muñoz-Muriedas, J.; Perspicace, S.; Bech, N.; Guccione, S.; Orozco, M.; Luque, F. J. Hydrophobic Molecular Similarity from MST Fractional Contributions to the Octanol/Water Partition Coefficient. *J. Comput. Aided. Mol. Des.* **2005**, *19* (6), 401–419.

(176) Muñoz-Muriedas, J.; Barril, X.; López, J. M.; Orozco, M.; Luque, F. J. A Hydrophobic Similarity Analysis of Solvation Effects on Nucleic Acid Bases. *J. Mol. Model.* **2007**, *13* (2), 357–365.

(177) Scannell, J. W.; Blanckley, A.; Boldon, H.; Warrington, B. Diagnosing the Decline in Pharmaceutical R&amp;D Efficiency. *Nat. Publ. Gr.* **2012**, *11*, 191–200.

(178) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; et al. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010**, *53* (10), 3862–3886.

(179) Tresadern, G.; Bemporad, D. Modeling Approaches for Ligand-Based 3D Similarity. *Future Med. Chem.* **2010**, *2* (10), 1547–1561.

(180) Klebe, G.; Abraham, U. Comparative Molecular Similarity Index Analysis (CoMSIA) to

Study Hydrogen-Bonding Properties and to Score Combinatorial Libraries. *J. Comput. Aided. Mol. Des.* **1999**, *13* (1), 1–10.

(181) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41* (23), 4502–4520.

(182) Chen, Q.; Higgs, R. E.; Vieth, M. Geometric Accuracy of Three-Dimensional Molecular Overlays. *J. Chem. Inf. Model.* **2006**, *46* (5), 1996–2002.

(183) MOPAC 6.0.; Version Locally Modified by Luque, F. J.; Orozco M. University of Barcelona **2008**.

(184) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E. .; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V. . P.; G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P. . I.; A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M. .; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T. .; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., J. .; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N. .; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J. .; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S. . T.; J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J.; et al. Gaussian09. Wallingford CT **2009**.

(185) Giangreco, I.; Cosgrove, D. A.; Packer, M. J. An Extensive and Diverse Set of Molecular Overlays for the Validation of Pharmacophore Programs. *J. Chem. Inf. Model.* **2013**, *53* (4), 852–866.

(186) Giangreco, I.; Olsson, T. S. G.; Cole, J. C.; Packer, M. J. Assessment of a Cambridge Structural Database-Driven Overlay Program. *J. Chem. Inf. Model.* **2014**, *54* (11), 3091–3098.

(187) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49* (1), 108–119.

(188) Sheridan, R. P. Chemical Similarity Searches: When Is Complexity Justified? *Expert Opin. Drug Discov.* **2007**, *2* (4), 423–430.

(189) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr. B.* **2002**, *58* (Pt 3 Pt 1), 380–388.

(190) Chan, S. L. MolAlign: An Algorithm for Aligning Multiple Small Molecules. *J. Comput. Aided. Mol. Des.* **2017**, *31* (6), 523–546.

(191) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47* (22), 5541–5554.

(192) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.

(193) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal Assignment Methods for Ligand-Based Virtual Screening. *J. Cheminform.* **2009**, *1* (1), 1–23.

(194) Anighoro, A.; Bajorath, J. Three-Dimensional Similarity in Molecular Docking: Prioritizing Ligand Poses on the Basis of Experimental Binding Modes. *J. Chem. Inf. Model.* **2016**, *56* (3), 580–587.

(195) Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 – A Public Library of Challenging Docking Benchmark Sets. *J. Chem. Inf. Model.* **2013**, *53* (6), 1447–1462.

(196) 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. *J. Med. Chem.* **2004**, *47* (7), 1750–1759.

(197) Schrodinger. Glide. LLC: New York.

(198) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47* (4), 1504–1519.

(199) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty

in Predicting Free Eneregies for Protein- Ligand Bindning. *J. Med. Chem.* **2006**, *49*, 5880−5884.

(200) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein - Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.

(201) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017**, *9* (2), 91–102.

APPENDIX

Other outputs related to the work of this thesis:

Papers

"Design, synthesis and biological evaluation of N-methyl-N-[(1, 2, 3-triazol-4-yl) alkyl] propargylamines as novel monoamine oxidase B inhibitors". Di Pietro O; Alencar, N, Esteban, G; Viayna, E; Szałaj, S; **Vázquez, J**.;Juárez-Jiménez, J; Sola, I;Perez,B; Solé,M; Unzeta, M; Muñoz-Torrero, D; Luque F J. *Bioorganic & Medicinal Chemistry* 2016, 24 (20), 4835-4854.

Oral Communications

"Hydrophobic similarity between molecules: Application to three-dimensional molecular overlays with PharmScreen". **Vázquez, J**. 255th ACS National Meeting/2018. American Chemistry Society, 20 March 2018, New Orleans Convention Center, EEUU.

"Hydrophobic similarity based on MST: Application to VS". **Vázquez, Javier**. Biomed PhD Day, 12 December 2017, Campus Torribera Santa Coloma de Gramanet, Spain

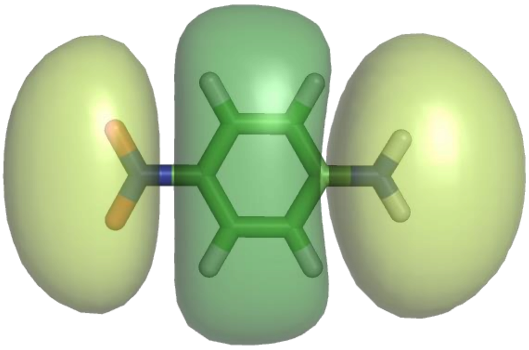
"Development and application of algorithm for virtual screening of chemical databases". **Vázquez, Javier**. Biomed PhD Day, 7 December 2016, Campus Torribera Santa Coloma de Gramanet, Spain

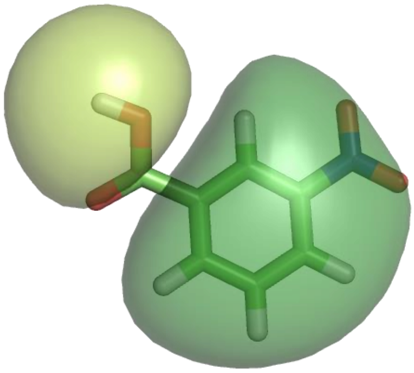
Posters

"Similarity assessment using lipophilic profile: A boost for structure-based methods with PharmScreen". **Vázquez, J**; Deplano, A; Herrero, A; Campos, L; Gibert, E; Herrero, E; Luque, F. J. GRC Computer Aided Drug Design. July 2019, Vermont, EEUU.

"Application to virtual screening of chemical databases with PharmScreen". **Vázquez, J**; Herrero, E; Luque, F. J. 5th Bioinformatics and Genomics Symposium, 20th December 2017, 20 December 2017, Hospital Universitari Vall d'Hebron (HUVH) Barcelona, Spain

"From continuum solvation models to hydrophobic descriptors: Application to virtual screening of chemical databases with PharmScreen". **Vázquez, J**; Deplano, A;Herrero, E; Luque, F. J.. The Royal Society of Chemistry, Medicinal Chemistry Residential School, June 27, 2017, Loughborough, England.





Facultat de Farmàcia i Ciències de l’Alimentació
Universitat de Barcelona