- **Article title:** Lipophilicity in drug design: An overview of lipophilicity descriptors in 3D-QSAR studies

- **Short running title:** Novel lipophilicity descriptors in 3D-QSAR.

- **Author names:** Tiziana Ginex,[1,*] Javier Vazquez,[1,2] Enric Gilbert,[2] Enric Herrero,[2] & F. Javier Luque[1,*]

- **Author affiliations**

1 Department of Nutrition, Food Sciences and Gastronomy, Faculty of Pharmacy and Food Sciences, Campus Torribera, 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.
2 Pharmacelera, Plaça Pau Vila, 1, Sector 1, Edificio Palau de Mar, Barcelona 08039, Spain.

- **Corresponding author details:** Tiziana Ginex (tiziana.ginex@ub.edu), F. Javier Luque (fjluque@ub.edu)

- **Financial disclosure:** We 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). JV is fellowship from the Generalitat de Catalunya.

- **Acknowledgements:** Dr. J. Muñoz-Muriedas (GSK group, Stevenage, UK) is acknowledged for valuable comments and suggestions.

- **Information pertaining to writing assistance:** N/A.

- **Ethical disclosure:** N/A

- **Author Contributions:** TG and JV performed computations. All authors contributed to the analysis and discussion of results. TG and FJL wrote the manuscript.

Word count: 6262
Figure number: 3
Table number: 2

Abstract

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 modelling. 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.

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

1. The pharmacophore concept and 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 to 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 towards 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 approach, 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 virtual screening.

Another successful application of the pharmacophore concept is linked to 3D quantitative structure-activity relationships (3DQSAR) [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 favourable 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.

2. 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 non-aqueous 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 (Eq 1).

$$\Delta G_{tr}^{o/w} = -2.303\ RT\,logP \qquad \text{Eq 1}$$

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 physico-chemical properties of the compound that may be relevant for the preferential solvation in aqueous and non-aqueous solvents, as shown in Eq 2 [21, and references therein].

$$logP = \upsilon V - \Lambda + I + IE \qquad \text{Eq 2}$$

where v is a constant, V is the molar volume, which encompass 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 non-polar groups of a molecule to aggregate in order to minimize the unfavourable exposition to the surrounding polar (water) solvent, lipophilicity is a measure of the affinity of the molecule for the non-polar solvent in a biphasic system constituted by a polar and a non-polar 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 is associated to high clearance and low metabolic stability. Overall, a careful handling of lipophilicity is required to optimize compound availability at the biological target.

Figure 1 here

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 favourable 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 (Eq 3).

$$LipE = -log(\text{potency}) - logP \qquad \text{Eq 3}$$

LipE 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 programmes, thus giving support to the formulation of the “lipophilic pharmacophore” concept.

3. 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 Eq 4,

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

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

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 an 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 centres 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 three-dimensional structure of compounds, and thus the effect of molecular conformation. Among all the available techniques, the molecular lipophilicity potential (MLP) [51] 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 Eq 5,

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

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 [50], 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 Eq 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.

Eq. 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 molecular lipophilic potential (HMLP) [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, Güssregen 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 to CoMFA.

4. Lipophilicity from QM continuum solvation methods.

More elaborate methods for estimating the partition coefficients have been proposed in the framework of 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 aims 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.

4.1 COSMO and 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 one-dimensional 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 bioisosteric 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 four-dimensional 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].

4.2 Fragmental lipophilicity model from the MST 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 [85, 86] at both semiempirical, Hartree-Fock and B3LYP levels [87-90] (for a review see [91]). From the solvation free energies in water and *n*-octanol, one can derive the *n*-octanol/water partition coefficient (Eq 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 non-polar ($\Delta G^{o/w}_{cav,i}$, $\Delta \mathrm{G}^{o/w}_{vW,i}$) contributions to the solvation free energy (Eqs 7 and 8).

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

$$logP_M = \sum_{i=1}^{N} logP_i = \sum_{i=1}^{N} \left(logP_{ele,i} + logP_{cav,i} + logP_{vW,i} \right) \qquad \text{Eq 8}$$

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 [92], leading to Eq 9.

$$logP^{o/w}_{ele,i} = \frac{1}{2}\langle\Psi^0 \left| \sum_{\substack{k=1\\ k\in i}}^{K} \frac{q_k^w}{|r_k^w - r|} - \sum_{\substack{l=1\\ l\in i}}^{L} \frac{q_l^o}{|r_l^o - r|} \right| \Psi^0\rangle \qquad \text{Eq 9}$$

where Ψ^o 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_k^w) and *n*-octanol (q_l^o), located at positions r_k^w and r_l^o.

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 (Eqs 10 and 11).

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

$$logP^{o/w}_{vW,i} = \sum_{i=1}^{N} S_i \Delta\xi^{o/w} \qquad \text{Eq 11}$$

where $\Delta G^{\,o/w}_{P,i} = \Delta G^{\,w}_{P,i} - \Delta G^{\,o}_{P,i}$, $\Delta G_{P,i}$ 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 [84]. By using the MST-based hydrophobic descriptors $logP^{o/w}_{ele,i}$ and $logP^{o/w}_{cav,i}$, 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 pre-aligned 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 (Eq 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 virtual screening 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 5 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 virtual screening 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.

5. *A comparative analysis between Hyphar and 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 data set compiled by Sutherland and coworkers [95], which comprises 113 angiotensin converting enzyme (ACE) inhibitors, 111 acetylcholinesterase (AChE) inhibitors, 147 ligands for benzodiazepine receptors (BZR), 282 cyclooxygenase-2 (COX-2) inhibitors, 361 dihydrofolatereductase (DHFR) inhibitors, 66 glycogen phosphorylase b (GPB) inhibitors, 74 thermolysin (THER) inhibitors, and 87 thrombine (THR) inhibitors.

Accordingly, the CoMFA/CoMSiA results reported in ref. 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 non-electrostatic 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 Material (Tables S1-S3).

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 [90]. Comparison of the statistical parameters obtained for the subset of training and test compounds defined for each molecular system is shown in Table 1.

Table 1. Statistical parameters of the 3D-QSAR HyPhar models obtained from MST/B3LYP and MST/RM1 calculations for the four sets of compounds.[a]

	Training set				Test set			Field (%)	
System	***r^2***	***q^2***	***S***	***Spress***	***r^2***	***S***	**Nc**	**Elec**	**Non-elec**
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

[a] See [92] 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 Non-elec stand for the fraction (in percentage) of electrostatic ($logP_{ele,i}$) and non-electrostatic ($logP_{cav,i}$) hydrophobic contributions to the final model.

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 data sets. 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 data set 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 (RM1) and Hyphar models in GPB and THERM systems compared to 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 Material, Table S4). On the other hand, both BZR and COX2 were confirmed to be challenging systems for QSAR modelling, as already noted by Sutherland and coworkers [95]. For instance, in case of COX2, part of the reason for the poor predictive behaviour may probably be ascribed to the fact that training and test set cover different ranges of in the property space.

Table 2. Statistical parameters obtained for CoMFA and CoMSiA models reported in [95] 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).[a]

	Training set				**Test set**			**Field (%)**		
System	***r^2***	***q^2***	***S***	***Spress***	***r^2***	***S***	**Nc**[c]	**Ele**	**N-Ele**	**HB**
ACE[b]										
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 [c]										
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 [d]										
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

[a] 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 Material (Table S4).

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

[c] Partition between training and test sets made as indicated in [15].

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

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 Material (Table S5). 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 Eqs. 13-15.

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

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

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

Figure 2 here

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 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 in 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 to the biological activity low active compounds.

Figure 3 here

Overall, the results obtained 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.

6. *Final consideration and perspectives*.

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, virtual screening, 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 [50] 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 [82-84,97-99], 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 to 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.

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 towards 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 sub-optimal 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 MST-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 MST-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

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.* 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, New York, (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, New York, 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, 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, Glase S, Pugsley T, Cruciani G, Pastor M,Clementi S. GRID/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, 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, 49–71 (1996).

20 *Drug bioavailability: Estimation of solubility, permeability, absorption and bioavilability*. van de Waterbeemd H, Lennernäs H, Artursson P (Eds.), Wiley-VCH, Weinheim, (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, 5, 425-452 (2007).

22 Van de Waterbeemd H, Carter RE, Grassy G, Kubinyi H, Martin YC, Tute MS, Willett P. 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, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol. 25*, 71–75 (2007).

** This study 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. 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 Caron G, Ermondi G, Scherrer RA. Lipophilicity, polarity and hydrophobicity. In: *Comprehensive Medicinal Chemistry II. (Volume 5.18).* Elsevier (Ed.), 425-452 (2006).

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 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).

51 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).

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, Cozzini P, Alberti S, Kellogg GE, Mozzarelli A. 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, Tripathi PK, Mozzarelli A, Scarsdale JN, Safo MA, Kellogg GE. 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).

64 Bodor N, Gabanyi Z, Wong C. A new method for the estimation of partition coefficient. *J. Am. Chem. Soc.* 111, 3783–3786 (1989).

65 Bodor N, Huang MJ. An extended version of a novel method for the estimation of partition

coefficients. *J. Pharm. Sci.* 81, 272–281 (1992).

66 Breindl A, Beck B, Clark T, Glen RC. Prediction of the n-octanol/water partition coefficient, logP, using a combination of semiempirical MO-calculations and a neural network. J. Mol. Model. 3, 142–155 (1997).

67 Beck B, Breindl A, Clark T. QM/NN QSPR models with error estimation: Vapor pressure and Log P. *J. Chem. Inf. Comput. Sci.* 40, 1046–1051 (2000).

68 Du Q, Liu PJ, Mezey PG. Theoretical derivation of heuristic molecular lipophilicity potential: A quantum chemical description for molecular solvation. *J. Chem. Inf. Model.* 45, 347–353 (2005).

69 Du Q-S, Li D-P, He W-Z, Chou K-C. Heuristic molecular lipophilicity (HMLP): Lipophilicity and hydrophilicity of amino acid side chains. *J. Comput. Chem.* 27, 685–692 (2006).

70 Palmer DS, Mišin M, Fedorov MV, Llinas A. Fast and general method to predict the physicochemical properties of druglike molecules using the integral equation theory of molecular liquids. *Mol. Pharm.* 12(9), 3420–3432 (2015).

71 Güssregen S, Matter H, Hessler G, Lionta E, Heil J, Kast SM. Thermodynamic characterization of hydration sites from integral equation-derived free energy densities: Application to protein binding sites and ligand series. *J. Chem. Inf. Model.* 57, 1652–1666 (2017).

72 Ansari SM, Palmer DS. Comparative molecular field analysis using molecular integral equation theory. *J. Chem. Inf. Model.* 58 (6), 1253–1265 (2018).

73 Orozco M, Luque FJ. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.* 100, 4187–4226 (2000).

74 Tomasi J, Mennucci B, Cammi R. Quantum mechanical continuum solvation models. *Chem. Rev.* 105, 2999–3094 (2005).

75 Cramer CJ, Truhlar DG. A universal approach to solvation modeling. *Acc. Chem. Res.* 41, 760–768 (2008).

76 Klamt A, Mennucci B, Tomasi J, Barone V, Curutchet C, Orozco M, Luque FJ. On the performance of continuum solvation methods. A comment on "Universal approaches to solvation modeling". *Acc. Chem. Res.* 42, 489–492 (2009).

77 Klamt A. The COSMO and COSMO-RS solvation models. *WIRES Comput. Mol. Sci.* 8, e1338 (2018).

78 Thormann M, Klamt A, Hornig M, Almstetter M. COSMO*sim*: Bioisosteric similarity based on COSMO-RS σ-profiles. *J. Chem. Inf. Model.* 64, 1040–1053 (2006).

** This study reports the application of the σ-profiles derived from Conductor-like Screening Model for Realistic Solvation (COSMO-RS) calculations in drug similarity measurements.

79 Hornig M, Klamt A. COSMOfrag: A novel tool for high-throughput ADME property prediction and similarity screening based on quantum chemistry. *J. Chem. Inf. Model.* 45, 1169–1177 (2005).

80 Thormann M, Klamt A, Wichmann K. COSMO*sim3D*: 3D-Similarity and alignment based on COSMO polarization charge densities. *J. Chem. Inf. Model.* 52, 2149–2156 (2012).

81 Klamt A, Thormann M, Wichmann K, Tosco P. COSMO*sar3D*: Molecular field analysis based on local COSMO σ-profiles. *J. Chem. Inf. Model.* 52, 2157–2164 (2012).

** The contribution examines the usage of local σ-profiles in molecular field analysis, providing an interpretation about the features of the virtual free energy field generated from the target binding pocket.

82 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).

** This study examines the performance of QM-based MST lipophilic (Hyphar) descriptors for calculation of molecular fields in the derivation of structure-activity relationships models.

83 Ginex T, Muñoz-Muriedas J, Herrero E, Gibert E, Cozzini P, Luque FJ. Application of the quantum mechanical IEF/PCM-MST hydrophobic descriptors to selectivity in ligand binding. *J. Mol. Model.* 22, 136 (2016).

84 Vázquez J, Deplano A, Herrero A, Ginex T, Gilbert E, Rabal O, Oyarzabal J, Herrero E, Luque FJ. 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 lipophilici (Hyphar) descriptors for predicting molecular overlays from three-dimensional similarity measurements.

85 Miertus S, Scrocco E, Tomasi J. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* 55, 117–129 (1981).

86 Cancès E, Mennucci B, Tomasi J. A new integral equation formalism for the polarizable continuum

model: Theoretical background and applications to isotropic and anisotrpic dielectrics. *J. Chem. Phys.* 107, 3032 (1997).

87 Bachs M, Luque FJ, Orozco M. Optimization of solute cavities and van der Waals parameters in *ab initio* MST-SCRF calculations of neutral molecules. *J. Comput. Chem.* 15, 446–454 (1994).

88 Luque FJ, Bachs M, Orozco M. An optimized AM1/MST method for the MST-SCRF representation of solvated systems. *J. Comput. Chem.* 15, 847–857 (1994).

89 Curutchet C, Orozco M, Luque FJ. Solvation in octanol: parametrization of the continuum MST model. *J. Comput. Chem.* 22, 1180–1193 (2001).

90 Soteras I, Curutchet C, Bidon-Chanal A, Orozco M, Luque FJ. Extension of the MST model to the IEF formalism: HF and B3LYP parametrizations. *THEOCHEM-J Mol. Struct.* 727, 29–40 (2005).

91 Luque FJ, Curutchet C, Muñoz-Muriedas J, Bidon-Chanal A, Soteras I, Morreale A, Gelpí JL, Orozco M. Continuum solvation models: Dissecting the free energy of solvation. *Phys. Chem. Chem. Phys.* 5, 3827–3836 (2003).

92 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).

93 PharmScreen - PharmQSAR, Pharmacelera, Barcelona, Spain.

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).