UNIVERSITÀ DEGLI STUDI DI PARMA

PhD in Food Science and Technology

XXVIII Cycle

Multi-scale approaches to Food Sciences: Computational QM and MM explorations at the microscopic level

Coordinator:

Chiar.mo Prof. Furio Brighenti

Tutors:

Chiar.mo Prof. Pietro Cozzini

Chiar.mo Prof. F. Javier Luque

PhD Candidate: Tiziana Ginex

Table of Contents

effects of thioxanthone metabolites through structure-based molecular docking."

Part 2. New QM-derived molecular descriptors for 3D-QSAR

Background

Paper 1:

"Development and validation of hydrophobic molecular fields derived from the Quantum Mechanical IEF/PCM-MST Solvation Models in 3D-QSAR."

Paper 2:

"Application of the Quantum Mechanical IEF/PCM-MST hydrophobic descriptors to selectivity in ligand binding."

Conclusions

References

Author

Acknowledgments

The author wishes to thank all the people who have contributed and collaborated actively to the realization of all the works discussed in this doctoral thesis.

A special thank goes to Enric Herrero and Enric Gilbert of the Pharmacelera group, Barcelona for the PharmQSAR software.

The Author

Tiziana Ginex

Multi-scale modelling approaches in Food Sciences and Technology

Multi-scale modelling approaches in Food Science and Technology

"Ut quod ali cibus est aliis fuat acre venenum"

"What is food to one is bitter poison to another"

Lucretius, ca. 96 B.C.–55 B.C. De Rerum Natura, Book IV, line 637.

Food products can be described as complex matrices of exogenous and endogenous, biotic and abiotic components each of which play an important, specific role on human health from a nutritional or toxicological point of view. This so-called "soft matter" is constituted by different mesoscopic structures, as bubbles, colloidal particles, emulsions droplets, amphiphiles or polymers (see **Figure 1**).

The main objective of food safety policy is to ensure consumers health through specific rules and protocols of security. In this context, both EU and USA have defined and enforced control standards for food produced within the country and those imported from other countries.[1] To address the purposes of food safety and quality standardization, in 2002 the European Parliament and the Council of the EU (Regulation (EC)178/2002 (EC, 2002)), tried *to approximate these concepts, principles and procedures so as a common basis for measure governing food and feed taken in the Member States at Community level*.

Formalization of rules and protocols standardization should pass through a more detailed and accurate harmonization of the knowledge on the global (macroscopic), pseudo-local (mesoscopic), and possibly, local (microscopic) properties of food products.[2] This means one has to apply multi-scale approaches to gradually reduce the magnitude of the system in an attempt to find an explanation of the physico-chemical

features and processes that influence the global (technologic) and local (toxicological or nutraceutical) properties of food products.[3]

Within a multi-scale approach, while coarse graining Molecular Dynamics (MD) approaches could allow to better understand the thermodynamics and dynamical (macroscopic/mesoscopic) behaviour of soft matter, classical Quantum Mechanics (QM), Molecular Mechanics (MM) and Dynamics (MD) simulations could be useful to extend our knowledge of (toxicological or nutraceutical) structure-property/activity relationships in complex food materials from a molecular point of view.

In general, macroscopic food-related phenomena mainly depend on entropic rather than enthalpic (chemical bonds) effects: they are generally modelled by using the principle of thermodynamics[4,5] and can be described by applying classical mathematical models, neglecting the relative influence of each molecular component.

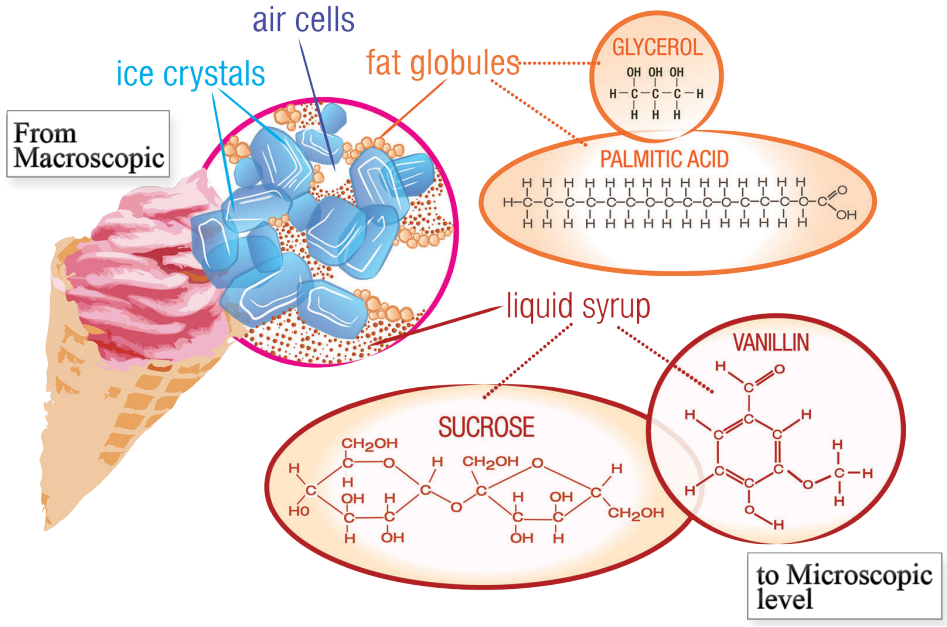


Figure 1. An example of food-related multi-scale approach.

Depending on the problem at hand, researchers look at motions occurring on time scales from femtoseconds (10^{-15} s) to hours, and covering distances from sub-atomic scale to cell size. A complete description in terms of all atomic coordinates would be desirable to explain the microscopic properties of food products but, if more chemical detail is included in a model, the computational requirements increase drastically, thereby affecting the length scales and timescales of the study. Another challenge is the

great chemical complexity and heterogeneity of biological systems. Even the smallest biological sample contains a wide variety of molecules.

We have to remember that a mathematical model is a simplified representation of a system where information input, an information processor and an output of expected results are contemplated. Its construction requires making some assumptions about the essential structure of the studied system and about the relationships between its constitutive elements and components. Accordingly, the objective is to reach a *necessary and sufficient* level of detail in a model to capture the physics that are responsible for specific and relevant food-related phenomena.[6]

Luckily, the increasing speed of molecular simulations due to advances in both hardware and software, and the development of force fields and methodologies, designed and adapted for specific purposes, have made already possible to describe increasingly complex biological systems and processes. It is interesting to note how they already succeeded in covering a significant (but not sufficient) range in a multi-scale biological context, from macro- to micro-level.

Progress in several directions can be foreseen in the near future to improve sensitivity, efficiency and efficacy of computational simulations, with the development of faster and more accurate methodologies for electronic structure calculations, more refined classical force fields, and improved coarse-graining techniques.

From meso- to microscale level

In computational studies, the equations that explain the behaviour of soft matter are solved using mesoscale simulations[7]: computational grids are used to treat the dispersed phase of the soft matter, generally represented as hard spheres into a continuum medium that can be modelled with free energy functionals.[8]

In his famous equation of state, van der Waals introduced some important concepts already used to study the soft matter as *free volume* and *mean fields*. These concepts derive from splitting the interaction between molecules into a strong short-range repulsive force, and a weak long-range attractive force. The repulsive forces determine the free volume, which is equal to the volume of a system available for insertion of new particles. The effect of attractive forces between particles can be captured with a (negative) internal pressure, which is added to the actual pressure field, to define an effective pressure or mean field. With these effective volumes and pressures van der

Waals reformulated the equation of state of an ideal gas as $p_{eff}V_{eff} = nRT$.[9] Another interesting concept applied for modelling of granulated and colloidal matter is the *effective temperature*,[10] which generally accounts for thermal fluctuations in a *non-equilibrium* thermodynamics context.

At specific conditions, mesoscopic sub-structures could self-assemble. An important requirement for self-assembly is the existence of predominant weak (non-covalent) attractive forces that lead to a final more ordered, self-assembled state. These systems are approached theoretically using numerical tools developed earlier for classical soft matter systems, such as Brownian Dynamics (BD),[11] Dissipative Particle Dynamics (DPD),[12] Self Consistent Field (SCF) theory or the related Dynamic Density Functional theory.[13] **Table 1** offers a rapid overview of some interesting applications of computational techniques to food science.

Among them, MD is probably the most commonly used approach for the study of biological macromolecules. In MD atomic motion are simulated by solving Newton's equation of motion simultaneously for all atoms in the system. MD simulations can be used to obtain both equilibrium and transport properties of a system. In this context, the development of accurate force-fields (FF) to treat complex biological systems[14-16] has contributed to the diffusion of these techniques also in fields not properly pertinent to theoretical chemistry. Examples of systems that can be simulated by classical MD simulations are:

- Ligand binding to enzymes.
- Self-assembly of lipids into relatively small aggregates.
- Small molecules adsorbing to a surface.
- The influence of molecules on the properties of a lipid bilayer.
- Coarse-grained simulations of polymeric nano-composites: lattice models,[1] bead-spring models.[17]

An illustrative example of the potential impact of computational tools is the risk assessment for pre-screening potential endocrine disruptors, improving experimental in vitro screening assay design and facilitating more thorough data analyses. They include, for instance, nuclear receptor (NR) crystal structures and homology models to examine potential modes of ligand binding by representative compounds through docking-based

approaches, multivariate principal component analyses (PCA) techniques to select best predicted cell lines for endocrine disrupting chemicals (EDC) in risk assessment purposes, and quantitative structure–activity relationships (QSARs) constructed from varied biological data sources using multivariate partial least squares (PLS) techniques and specific descriptors.[18]

In this context, a docking-based screening for preliminary (*in silico*) hazard evaluation of the potential xeno-androgenic effects of some food contaminants has been reported by the author and will be discussed in detail later in this thesis.

Scale	Technique	System	Ref.
Micro	MD	Carbohydrates	19-21
		Protein/polysaccharides mix	22
	SCF	Casein at interfaces	23
		Liquid crystals	24
		Protein/polysaccharide mix at interface	25
	Monte Carlo	Proteins adsorption at interfaces	26,27
		Surfactants at interfaces	28
		Bulk protein/polysaccharides mix	28
Mesoscale	Brownian Dynamics	Colloids at interfaces	29
		Food gels	30
		Colloids displacement at interface	31
	Lattice Boltzmann	Emulsions	32
		Suspensions	33
Macro	Porous media approach	Intensively heated foods	34
Multiscale	Cell model + Macro	Expanded snacks	35

Table 1. Advanced simulations techniques in food soft matter approaches. Reproduced from R. G. M. van der Smam, 2012.[2]

MD simulations in food sciences

To give a direct and rapid explanation of the potentialities of MD simulations in food sciences, some relevant cases (caffeine self-association in water, spherification, catechins-lipid bilayers interactions, the role of cholesterol in lipid bilayers, and

olfactory receptors) of recent applications of MD chosen from literature have been selected and discussed.[36]

Caffeine: Solubility and self-association in water

Coffee is among the most popular beverages in the world. The most important active substance is caffeine (1,3,7-trimethylxanthine; **Figure 2**), which is also found in chocolate, tea and other kinds of soft drinks. Its huge consumption is related to the role as stimulant of the central nervous system. In such beverages, caffeine is in solution with many other constituents, such as polysaccharides, lignans, polyphenolics, lipids, melanoidins, and chlorogenic acid.[37] The physicochemical properties of caffeine justify its mild solubility in water, which is estimated to be around 0.11 M at 298 K.[38] Thus, even though it has a significant dipole moment (~3.7 Debyes), the presence of the methyl groups attached to the heterocyclic ring makes the surface of this flat molecule to be only moderately hydrated. Hence, caffeine is sparingly soluble in water at room temperature and the solubility increases with temperature. In fact, caffeine self-associates to form dimers, and to some extent trimers and higher aggregates, and the association is enthalpically driven.[38-40] The apparent self-association constant of the vertical stacking process has been estimated to be 10.5-16.5 M^{-1} at 298 K.[41-43]

MD simulations have been recently used to gain insight into the structural details of caffeine hydration and association in aqueous solution.[44] The results show the complex organization of the hydrating water molecules, which appear in three distinct areas around the caffeine solute (**Figure 2**). Preferred hydration sites are found above and below the molecular plane, close to the carbonylic oxygens. An additional hydration site is localized in the molecular plane facing the lone pair of the pyridine-like nitrogen in the five-membered ring. Finally, a set of weakly bound water molecules is found above and below the molecular plane.

The trajectories also revealed the dynamical exchange of caffeine molecules in self-association. Thus, even though association is guided by stacking of caffeine molecules, steric clashes between methyl groups limits the number of relative arrangements between stacked molecules.

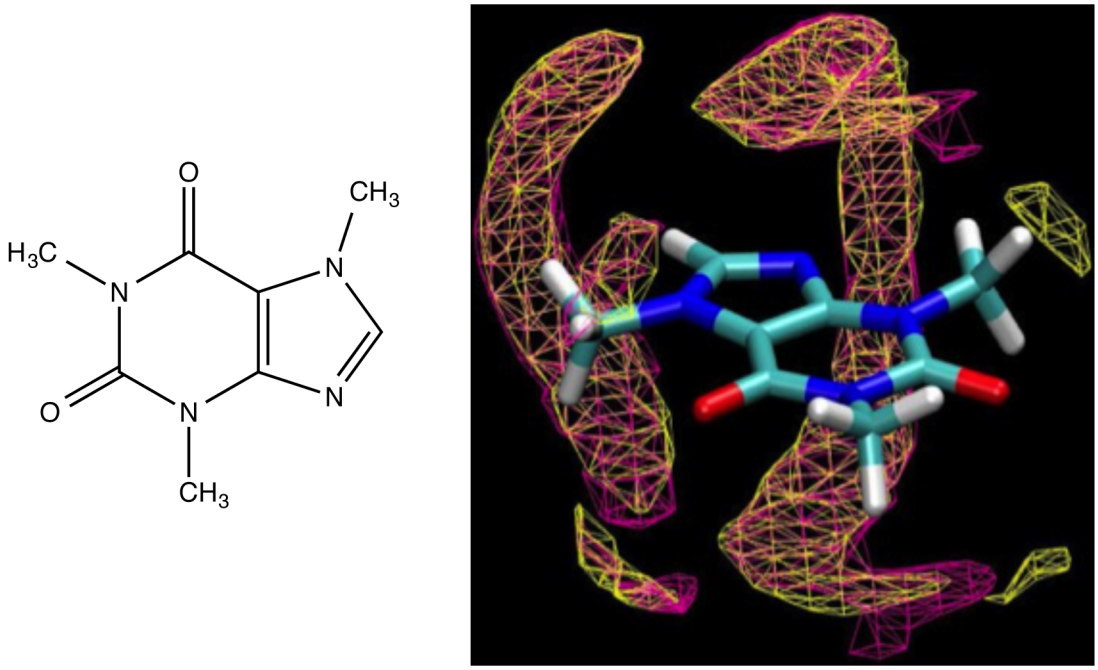


Figure 2. (Left) Chemical structure of caffeine. (Right) Isocontour representation of the water distribution around a single caffeine molecule. Isocontour surfaces encloses regions with a water oxygen atom density 1.3 times (yellow) and 1.4 times (red) the bulk density (Reprinted with permission from J. Phys. Chem. B. 2011, 115, 10957-10966. Copyright 2011 American Chemical Society).

The effect of cosolutes on the association of caffeine has also been studied. The results reveal distinctive trends in the influence exerted by some inorganic salts on the self-association of caffeine.[43] In particular, sodium chloride tends to destabilize monomeric caffeine, leading to slightly larger self-association. This trend has been reproduced by MD studies performed for a system comprising 8 caffeine molecules and 4500 water molecules, with a salt (NaCl) concentration ranging from pure water to 0.83 M.[45] The analysis of the caffeine-caffeine centre of mass distribution function showed that, on addition of salt, there is an increase in the first peak (located at around 4 Å), but also pointed out the appearance of a second peak (at around 7.2 Å). Furthermore, it was found that higher order clusters of caffeine are formed as the concentration of NaCl increases.

The effect of co-solutes on the caffeine aggregation has also been extended to other substances, such as sucrose, which was found to lower the solubility in water as the sugar concentration increases.[38] The interaction between caffeine and sugars has been recently examined.[46] Simulations showed that sugars weakly interact with caffeine by forming face-to-face stacking, with the nonpolar regions of sugars being primarily oriented toward the hydrophobic face of caffeine. Hence, it can be speculated that

liberation of water molecules structured over the hydrophobic face of caffeine might facilitate self-aggregation.

While the preceding findings are in agreement with available NMR data,[46] they also raise a number of questions, such as the impact of sugars on the interaction of caffeine with other components extracted from coffee, and how the intertwined set of interactions may affect the taste of the beverage. As an illustrative example, it is worth noting the influence of ions on the dissolution and extraction of flavoursome compounds, so that different rates and compositions of coffee extraction can be achieved by varying the nature of the ionic species present in water.[47]

Spherification: Role of calcium cations in alginate aggregation

Alginates are naturally occurring unbranched polysaccharides isolated from brown algal species and from the extracellular matrix of certain bacteria. They are composed of (1→4)-linked residues of β-D-mannuronic acid (M) and its C5-epimer, α-L-guluronic acid (G), and the precise chemical composition depends on the biological source.[48] Alginates are biodegradable, biocompatible and nontoxic. They are commonly used as thickeners in food industry, as well as stabilizing agents and emulsifiers.[49] Hence, there is technological interest in understanding their biochemical and biophysical properties, the nutritional properties of dietary alginates and the potential use as drug carriers.[50-54]

The sol-gel transition of sodium alginate is induced by the presence of calcium ions, which exhibit a preferential affinity for G units compared to the M ones.[55-57] Thus, the gel forming properties of alginates are derived from their capacity to bind a large number of divalent ions and the gel strength is correlated with the proportion and length of the G residues in the alginate chains. This process is the basis of 'spherification', a technique amply utilized in modernist cuisine pioneered at El Bulli.[58] This technique, which relies on the formation of alginate spheres that contain a physical outer gel membrane with a liquid core, is central to the formation of faux caviar, eggs, gnocchi and ravioli. In the process of spherification a liquid containing sodium alginate is submersed in a bath of calcium or, alternatively, a calcium source is added to the edible liquid and then mixed with a sodium alginate bath. Gelation involves Ca^{2+} cross-linking of alginate chains, which results in the formation of a film surface.

The details of the molecular events implicated in this process are still poorly understood, which in turn explains the increasing interest in gaining insight into the structural and energetic features of the interaction between calcium cation and alginate by means of theoretical studies. An aspect that has attracted much interest is the origin of the differential interaction of calcium cations with G and M units, as the more favourable interaction with G units is believed to arise from the 'egg-box' model proposed around 40 years ago.[59] According to this model, by adopting a 2_1 helical conformation, the alginate chain can pack assisted by the presence of the calcium cations between the distinct chains (**Figure 3**).

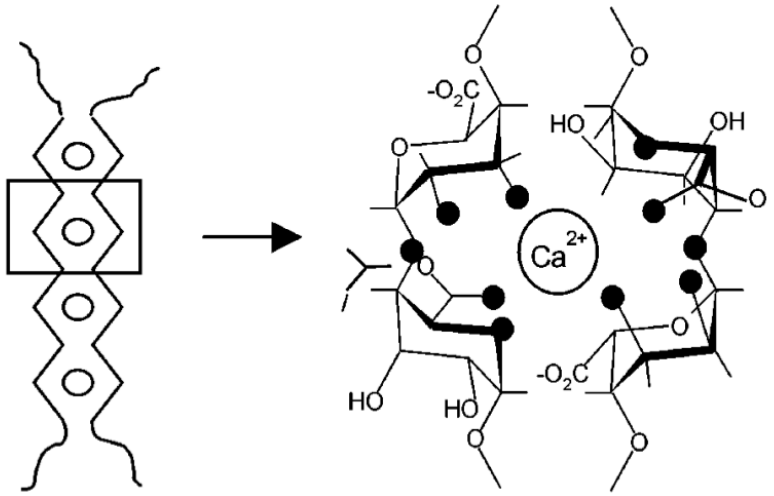


Figure 3. Schematic representation of the calcium coordination in the "egg box" model in calcium alginate junction zones (Reprinted with permission from Biomacromolecules 2001, 2, 1089-1096. Copyright 2011 American Chemical Society).

Several studies have examined the suitability of this model.[60-62] Recently, Plazinskin and Dracht[63] have used hybrid MD-DFT calculations to examine the interaction of calcium cations with a single G molecule or with an oligomeric chain of G units in order to disclose the interactions involved in cation chelation. They concluded that the preferential interaction of Ca^{2+} include the coordination with negatively charged carboxyl group and water molecules, while excluding a relevant contribution from hydroxyl groups or ring atoms. From an energetic point of view, both mono- and bidentate binding modes are found to be nearly equally probable. Finally, the results point out that only the carboxyl groups of two poly(guluronate) chains contribute to the binding of calcium and to the junction of these two chains.

Recently, by using atomistic MD simulations, Chipot and co-workers have examined the process of how liquid food can be shaped into spheres.[64] Specifically, they analysed the structure of the calcium alginate membrane and the role played by calcium ions, in comparison with sodium ions, the structural stability of the polysaccharide membrane, and the adaptability of the alginate membrane to encapsulate liquid foods of different nature.

The analysis of the trajectories showed that the alginate chains spontaneously form a net-like membrane on the surface of the liquid droplets (**Figure 4**). Furthermore, they also showed that the joints between adjacent alginate chains possess egg-box motifs, whereby the stabilization can be ascribed to the electrostatic interactions between calcium cations and the carboxylate moieties. Importantly, such stabilization is significantly reduced in the presence of sodium cations, thus pinpointing the crucial role played by the divalent cation in spherification.

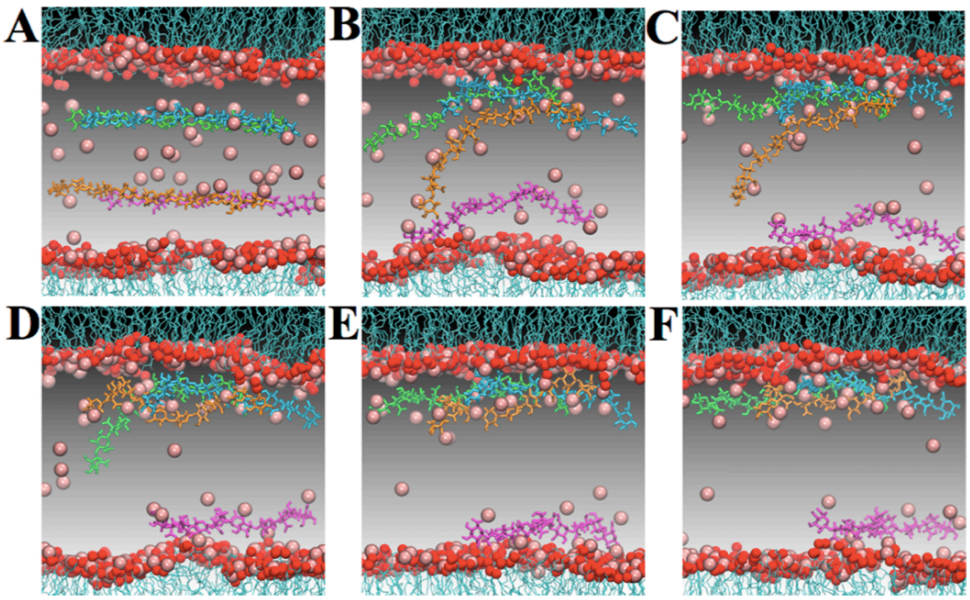


Figure 4. Representation of the spontaneous adsorption of four alginate chains at the oleate-water interface. Panels A-F denote selected snapshots along the 80 ns trajectory. Calcium ions are represented as pink spheres (Reprinted with permission from J. Phys. Chem. B 2014, 118, 11747-11756. Copyright 2011 American Chemical Society).

Green tea catechins: interaction with lipid bilayers

Catechins are polyphenols belonging to the class of flavan-3-ols typically found in cocoa, berries and green tea. Several beneficial effects, partly due to their antioxidant activity, have been reported.[65] Green tea (*Camellia sinensis*) contains high levels of flavan-3-ol monomers, the main components being (-)-epigallocatechin (**EGC**), (-)-epigallocatechin-3-O-gallate (**EGCG**), and (-)-epicatechin-3-O-gallate (**ECG**).[66] Fermentation of green leaves leads to reduction in flavan-3-ols content as a result of the action of polyphenol oxidase with a consequent accumulation of theaflavins and thearubigins in black tea.[67] Studies on animal models show that **EGCG** can inhibit carcinogenesis at all stages, viz. initiation, promotion and progression.[68]

Understanding the biophysical and biochemical implications of the interaction between catechins and specific cellular targets is of utmost relevance for their pharmacological use. Furthermore, even though catechins have proved to be effective in in vitro assays, their poor pharmacokinetic profile makes them unsuitable for direct in vivo applications. Thus, they are potentially good "lead" compounds for drug discovery in a nutraceutic scenario.

It has been speculated that the biological activities of catechins could be determined by their interaction with lipid membranes.[69,70] For instance, differential scanning calorimetry and spectroscopic studies by Catura *and co-workers*[71] have revealed that **ECG** and **EGCG** bind deep in bacterial model membranes. The interaction of tea catechins with lipid bilayers is influenced by physicochemical factors such as the number of hydroxyl groups on the B-ring, the presence of the galloyl moiety and the stereochemical structure. At this point, **EGCG** has been found to be the most effective in perturbing the membrane structure of bacteria-like model membranes.

Sirk and co-workers[72] performed atomistic MD simulations to gain insight into the role of hydrogen bonding of catechins with components of cell membranes. The simulation model consisted of a lipid bilayer containing a 1:1 mixture of 1-palmitoyl-2-oleoyl-phosphatidylcholine (POPC) and 1-palmitoyl-2-oleoyl-phosphatidylethanol-amine (POPE), which were considered to build up a membrane model that mimics the HepG2 membrane cell. Structural and dynamic properties for seven catechins (**C**, **EC**, **EGC**, **CG**, **ECG**, **EGCG** and **GCG**; **Figure 5**) interacting with the lipid bilayers, such as molecular orientation and hydrogen bonding, were monitored during 50 ns MD simulations.

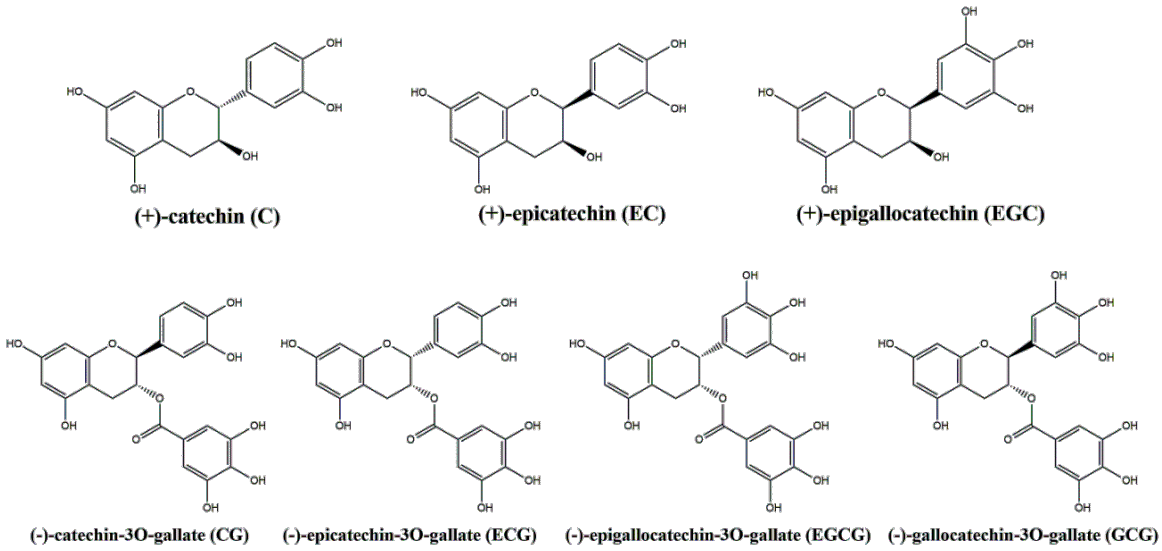


Figure 5. Chemical structures of the seven catechins studied by Sirk and co-workers (74).

Each catechin showed distinctive trends with regard to the catechin-bilayer interaction. All of them were able to interact with the lipid bilayer and were rapidly internalized. An important exception to this common behaviour was represented by **EGCG**, since its absorption into the lipid bilayer was preceded by adsorption and fluctuation on the lipid surface. More generally, the adsorption process was influenced by the size and conformational behaviour of compounds.

Once into the bilayer, all catechins maintained flexibility: their diffusion though the membrane was characterized by formation and breaking of hydrogen-bond interactions with the lipid headgroups. **EGCG** was the compound that formed a larger number of interactions. This effect was ascribed to the presence of the gallate unit and its *cis* configuration with regard to the B-ring, which promote multiple hydrogen-bond interactions. This unique behaviour among catechins might explain the ability of **EGCG** to target specific cells.[73] MD simulations also confirmed the higher affinity of **EGCG** and **ECG** for lipid bilayers compared to **EC** and **EGC**, providing important sub-molecular insights on the capability of catechins to interact with lipid bilayers.

Overall, these studies illustrate the potential impact of MD simulations in gaining insight into the physicochemical basis of the biological effects of polyphenols, which in turn should lead to a better comprehension of their nutraceutic role and the potential implementation in food sciences.

The ubiquity of cholesterol: the feared and crucial molecule

Even though most people relate excess of cholesterol with serious health problems, the importance of this molecule for our life is not generally well appreciated. Cholesterol is an essential component of the mammalian cell membranes needed for their structural integrity, fluidity and permeability. Additionally, it is the precursor molecule of vitamin D, bile acids and steroid hormones as cortisol, progesterone, estrogens and testosterone. Due to the significant health implications, cholesterol and its interactions with target molecules and different environments have been intensively studied in the last 30 years.[74]

The relationship between cholesterol and the properties of the lipid bilayer has been extensively studied by means of MD simulations. Among other topics, attention has been paid to the dependence of the structural and dynamical features of the lipid bilayer on the percentage of cholesterol, considering both interaction patterns[75-78] and the rigidity and membrane order.[79-89] These studies have provided fundamental information about the behaviour and properties of the cell membranes. Other relevant subjects are the effect of cholesterol on the permeability and the different hydration degree of the membrane,[90,91] and the influence on the formation and stability of membrane pores.[92-94] The effect of cholesterol analogs on the fluidity, order level and permeability of the membrane have also been examined by MD simulations, including compounds such as dehydroergosterol,[95,96] anosterol or ergosterol,[97-99] among others.[100-102]

Cholesterol may also form direct interactions with membrane proteins, particularly with G protein-coupled receptor (GPCRs), as recently reviewed by Oates and Watts.[103] GPCRs form a group of membrane proteins characterized by seven transmembrane (TM) helical segments connected by three extracellular and three intracellular loops and an alpha-helical C-terminus. The influence of cholesterol on GPCR activity is well known, but the precise nature of the interaction remains to be elucidated, although the discovery of bound cholesterol in the crystal structure of the β2-adrenergic receptor and the subsequent identification of a consensus cholesterol binding motif supports a direct effect on the receptor.[104] For instance, results from MD simulations of rhodopsin indicate that cholesterol influences the activation elements in the TM1-TM2-TM7-H8 functional network of the receptor, such as the relative angular motion between TM7 and H8, and the relative movement of TM1 and TM7, by affecting the helical kink parameters in TM1, TM2, and TM7.[105] In particular, the presence of cholesterol near

Val1.58, Tyr2.63, and Pro7.38 was found to correlate with the changes in the kink angle values in the TM1, TM2, and TM7, respectively.

Cholesterol also exerts effects in ligand-gate ion channels. In the glutamate-gated chloride channel, the X-ray structure of the open state has revealed the existence of a lipophilic agonist, ivermectin, bound to the subunit interface in the transmembrane domain.[106] MD simulations supported a stable binding, though colesterol exhibits some fluctuations between several orientations.[107] Furthermore, on the relatively short timescale spanned by simulations, cholesterol tends to promote pore opening.

In the case of the nicotinic acetylcholine receptor, MD simulations showed that cholesterol tend to concentrate around TM helix 4, in agreement with photoaffinity labeling studies, typically interacting via hydrogen bonding and hydrophobic interactions.[108] A putative cholesterol-binding pocket involving TM cysteine residues (Cys445, Cys220) of the β2 subunit was also identified in the open-channel system, while nonannular cholesterol binding sites were found in the closed-channel species (**Figure 6**). These findings suggest that cholesterol might be involved in the regulation of the channel gating process through direct interactions with critical residues.

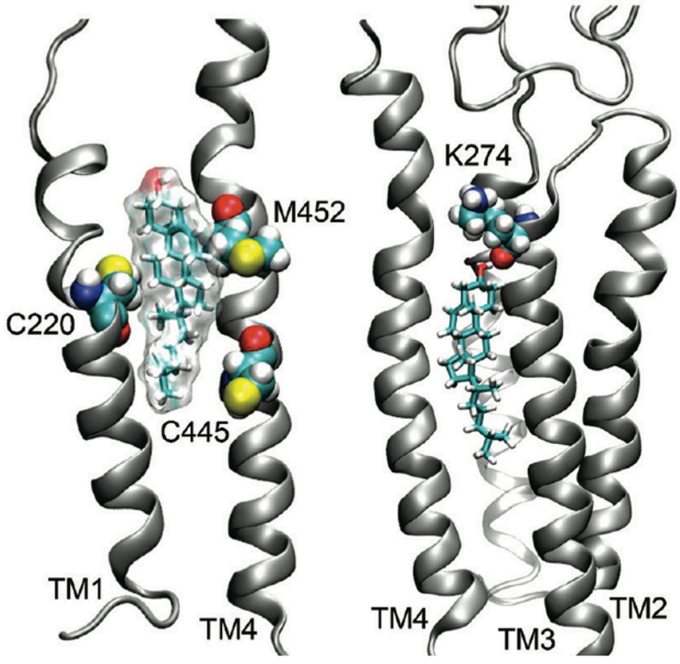


Figure 6. Representation of cholesterol binding sites in the open channel model of the nicotinic acetylcholine receptor. (Left) Cholesterol bound to the three residues containing sulfur atoms (yellow) in the b2 subunit. (Right) Hydrogen bonding of cholesterol to Lys274 in the b2 subunit (Reprinted with permission from J. Phys. Chem. B 2009, 113, 6964-6970. Copyright 2011 American Chemical Society).

Overall, these examples illustrate the potential contribution of MD simulations in clarifying the nature of the interactions formed between cholesterol and membrane proteins, and especially in disclosing the effect of these interactions on the biological function.

Olfactory receptors: the influence of food odour

Olfaction is a complex chemo-biological process that happens when odorants or pheromones bind to olfactory receptors (ORs) in the membrane of specialised sensory cells, thereby activating a signalling pathway that is finally transformed into an electrical signal by the opening of ion channels.[108,109] ORs belong to the superfamily of G protein-coupled receptors, and the binding of odorants activates the consequent signal pathway cascade inside the OR neurons. The relation between odorants and ORs could be defined as promiscuous in the sense that each OR can bind and respond to different odorant molecules, and a single odorant can activate multiple ORs that will recognise different structural features of the odorant. As a result, odours are the consequence of the superposition of several signals generated cooperatively by multiple ORs.

Several attempts have been made to establish correspondences between odorants and ORs as well as to classify ligands as agonists or non-agonists and to unravel the determinants of molecular recognition.[110-112] Singer was the first that used rhodopsin to model the rat's I7 olfactory receptor to analyse the interactions of octanal in its binding site by flexible ligand docking coupled to short (20 ps) restrained MD simulations,[113] in the first published work where MD was applied to model an OR. This model was later used to study the interaction of aldehydes with rat OR I7, leading to the identification of an exit pathway from the binding site through MD simulations in vacuum.[114] Identification of the pathway allowed to relate the existence of a clear transit path with OR activation in agreement with experimental data.[115]

With the recent advances in computing power, Anselmi *et al*. modelled the human OR17-40 using the crystal structures of rhodopsin and the ß2-adrenergic receptor.[116] By using free energy techniques, they showed that although different molecules can bind the OR binding site, only a few of them can bind strongly enough as to elicit an electrophysiological response. A similar strategy has been used to study how the human

OR1G1 is able to recognise a broad set of chemically diverse molecules,[117] including four known agonists (nonanal, 1-nonanol, 9-decen-1-ol and camphor) and one non-agonist (butanal). This study revealed the importance of the hydrophobic contribution to the binding free energy compared to the electrostatic term, which would explain the adaptability of the ORs to diverse chemical structures.

More recently, Lai et *al.* have run MD simulations of two mouse ORs, S97 and S86, forming a complex with different odorants to try to establish a binding pattern common to all of them. Their simulations showed that both ORs have two different binding sites (**Figure 7**) and that in each case binding to one but not the other has an enhanced activating effect on the receptor.[118]

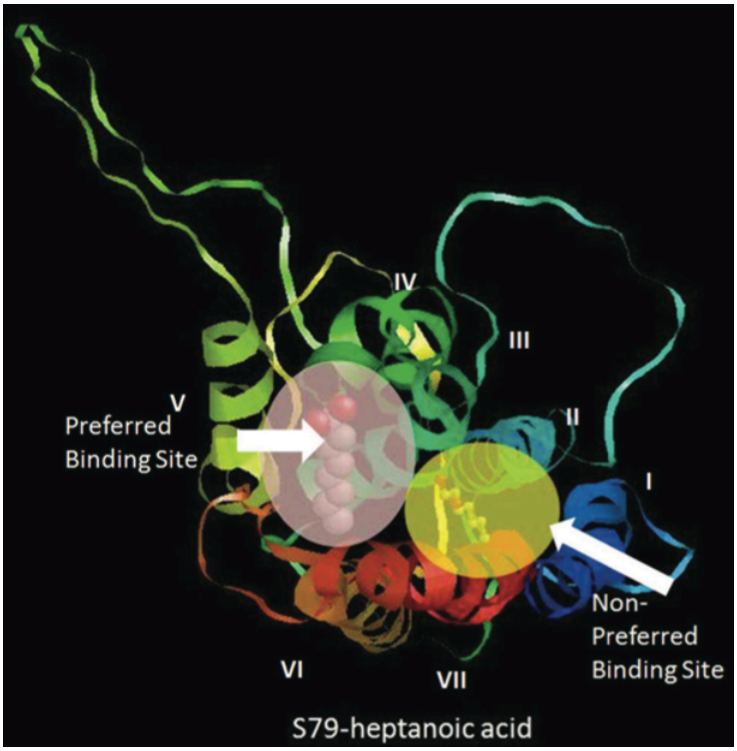


Figure 7. Representation of the structure of OR S70 showing the two putative binding pockets. The circle bound by TMs III, IV, V and VI is the region that favours OR activation, and the circle bound by TMs I, II, III and VII is the non preferred binding region (Reprinted with permission from Chem. Senses 2014, 39, 107-123. Copyright 2014 Oxford University Press).

Overall, MD-based simulation techniques are promising tools to decipher the intricate events of OR recognition and activation. However, further efforts need to be made to get a deeper insight into the promiscuity of ORs and the activation

mechanisms, which will be valuable to understand how odorants generate odours that we sense as smells.

Final remarks

As demonstrated by the preceding cases, MD simulations have all the potentialities to disclose the molecular details of relevant biochemical processes using more complete and realistic description of biomolecular systems. Predicting the impact of these technical progresses in food sciences is always difficult, especially keeping in mind that they are in a very incipient state. Nevertheless, it is reasonable to expect that MD simulations will have an increasing impact in understanding at the microscopic level the structure and properties of the basic constituents of foods. Up to now, the explicit description of food features at the microscopic scale has been poorly addressed, likely due to the high degree of molecular complexity, thus making it necessary to resort to models that incorporate these features in an average way. Nevertheless, we can be confident that the advances introduced in atomistic simulations will be able to gain deeper insight into the molecular properties and physics of foods and their transformations.

Multi-scale models are necessary to provide a range of sub-models for describing the intricacies of the material behaviour of foods and of the physical phenomena in food engineering at different spatial and time scales. In this context, it can be expected that the advances of MD simulations will facilitate gaining insight into the fine structure of food materials, which in turn will be valuable to develop more accurate representations for coarser descriptions of foods, thus providing bridges between the microscopic and mesoscopic models. Thus, even though the microscopic simulation of foods can still be considered to be in its infancy, we can be confident that the identification of the hidden relationships between microscopic and mesoscopic scales will provide the required background for food structural engineering.

Computational methods: from ligand- to target-based strategies

Computational methods: from Ligand- to Target-based strategies

Once the 3D structure of a molecule and all the parameters required for the atomic and molecular connectivities are known, the energy of the system can be calculated. In this context, the precision and the accuracy of a biomolecular simulation in reproducing experimental data is intrinsically related to (i) the quality of the starting material (crystallographic data as X-ray, NMR protein structures) and (ii) the accuracy of the mathematical equations used to model the biological event. In this scenario, in silico techniques can be ideally classified in ligand- and target-based approaches.

The former allows to study certain molecular events from a ligand point of view, whereas the latter focus the attention on the macromolecular counterpart. The election of the technique to be applied depends on the stating material available and purposes.

The basis of molecular simulations

The mechanical molecular (MM) model was developed with the objective to make *accessible* the description of molecular structures/properties of complex biomolecular systems in a practical manner, reaching a reasonable good balance between computational cost and accuracy in energetic evaluations.

It has become possible thanks to the introduction of some simplifications as the use of atomistic models and empirical energy functions. The ensemble of (i) these atomic/bond parameters for each type of atoms, bonds, dihedrals and (ii) mathematical equations used to describe the potential energy for a given molecular system constitute a force-field (FF), where parameters for energy functions are generally derived from experimental data or quantum mechanical (QM) calculations.

A potential energy function is a mathematical equation that renders the potential energy, V, of a chemical system as a function of its three-dimensional (3D) structure, R.[120-123] It can be considered as a sum of internal/covalent, $V(R)_{internal}$ and external/non-covalent, $V(R)_{external}$, potential energy, as reported in Eqs 1-3 (see **Figure 1**).

$$V(R)_{total} = V(R)_{internal} + V(R)_{external} \quad (1)$$

$$V(R)_{internal} = \sum_{bond} k_b (b - b_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\chi [1 + \cos(n\chi - \sigma)] \quad (2)$$

$$V(R)_{external} = \sum_{\substack{non-bonded \\ atom\ pairs}} \left(\varepsilon_{ij} \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - \left(\frac{R_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\varepsilon_D r_{ij}} \right) \quad (3)$$

Both (bond) stretching and (angle) bending terms are generally treated harmonically by applying the classical Hooke's law, which effectively keeps bonds (b_0) and angles (θ_o) near their equilibrium values. k_b and k_θ are the force constants associated with the bond and angle terms, respectively. The dihedral term refers to rotation of molecular groups adjacent to two bonded atoms; it implies a certain periodicity for the torsional term, generally treated with Fourier series. As a consequence, the equation used to describe this phenomenon includes parameters for the force constant, k_χ, the periodicity or multiplicity, n, and the phase, δ.

The most important term for computational studies of biological systems is the external/non-covalent interaction term and consists on summation of van der Waals and electrostatic interactions. The vdW (non-polar) interactions are generally expressed in term of Lennard-Jones 6-12 potential, where $1/r^{12}$ refers to the exchange repulsion between overlapping atoms and $1/r^6$ accounts for London's dispersion. Parameters are associated with the external terms are the well depth, ε_{ij}, between atoms i and j, and the minimum interaction radius, R_{ij}. The electrostatic or Coulombic potential describes attractive/repulsive behaviour between two atoms i and j with partial atomic charge q_i

and q_j. Also included is the dielectric constant, ε_D, which is generally treated as equal to 1, the permittivity of vacuum.

All the required MM parameters can be derived empirically (from experimental data) or from *ab initio* calculations. The quality of a FF depends on (i) the quality of the experimental data and the accuracy of the methods utilized to parameterize it, and (ii) on the set of equations used to evaluate both internal and external energies. Parameterization is performed on a "model-system" leading to a "limited" transferability among different biochemical contexts. Moreover, empirical derivation of some parameters may influence the suitability of the FF for the description of processes in different environments (i.e., in aqueous solution or in the interior of membranes). On the other hand, extensions to the classical minimal potential energy functions have been introduced to account for other important molecular phenomena, such as anharmonicity in bond stretching or polarizability. This has lead to formalization of the Class II (polarizable) FFs as MM3[124] and CFF93.[125] Nevertheless, most FFs limit the nonbonded terms to the combination of Coulombic (atomic charges) and Lennard-Jones parameters, which must provide a reasonable description of intermolecular interactions. Generally, the molecular net charge is distributed to all atoms, leading to an atomistic description of partial charges.

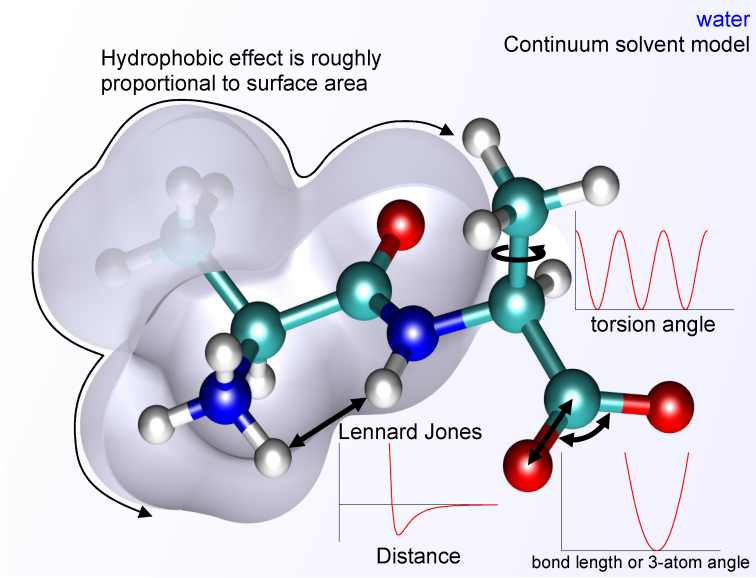


Figure 1. Representation of the internal (covalent) and external (non-covalent) interactions for a simple molecular model.

As reported by Cramer,[126] atomic charges can be derived empirically, semi-empirically (AMPAC/MOPAC) or quantum-mechanically, from partitioning of the wave function used to describe the molecular charge distribution.

The simulation of biological systems via MM has matured into routine application over the last decades with empirical FF becoming robust and reliable tools for the description of molecular structure and energies. The majority of biomolecular simulations are generally performed with the use of Class I (or Classical) FFs as the CHARMM[127-129], AMBER[130], and GROMOS packages. Other used FFs are Merck Molecular Force Field (MMFF)[131] and MM2 with its further implementations.[132]

Molecular Mechanics and Dynamics

MM and MD are the two mainstream techniques used to perform biomolecular simulations. Both rely on the selection and application of FFs with inclusion of a time factor for the latter.

Molecular Mechanics

The physico-chemical properties of a molecule are related to available conformational space in a given environment. MM describe molecular systems as a collection of spherical particles connected by springs that represent the constituent atoms and bonds. Accordingly, a specific potential energy value is associated to every conformational state as a function of the Cartesian coordinates of the atoms. Therefore, the energetic and conformational evolution of a molecule can be represented as a hypersurface of potential energy, generally characterized by local and global *minima.*[121-123] The global minimum is the lower energy value associated to the most stable conformation. The process that leads to exploration of the entire conformational space for a given molecule is called *conformational analysis*.

There are several methods to explore the conformational space of a molecule. In brief, they can be divided in:

- Non-derivative Methods
  - Simplex method
  - Systematic univariate methods
- Derivative Methods:
  - Steepest-descent
  - Conjugate gradient

The non-derivative methods directly explore the potential energy surface during application of geometrical modifications. Each new conformation is accepted only if it reaches an energy value lower that that of the previous conformation. The classification of the derivative methods depends on the order of derivation applied. Steepest-descent (SD) and Conjugate gradients (CG) are first derivative methods. In the former, the direction of the first derivative of the energy (the gradient) is parallel at each point of the surface whereas, in the CG approach, an orthogonal "conjugate" direction is selected during minimization. Given the mathematical derivation of the new minimum, the first method has to be used to perform a rapid, "*steep*" initial research, whereas the second one can be used to refine the local search in proximity of the minimum.

Molecular Dynamics

The time evolution of a molecular system can be followed by solving Newton's equations of motion, leading to a collection of snapshots that represent the structural fluctuations of the simulated system. Running a MD simulation requires the adoption of certain assumptions about the level of description of the biomolecular system, the interactions between the classical particles via suitable potential energy functions, and the algorithms utilized for sampling of the conformational landscape.

Description of the simulated system. Atomistic models provide a description of the simulated system at atomic detail, so that the trajectory will permit to follow the time evolution of each atom subject to appropriate constraints imposed by the molecular topology. Interactions between "classical" atoms are described by means of the FF, which determines the quality of the results. This has led to the introduction of continued

revision of current FFs, such as AMBER,[133,134] leading to successive refinements for the treatment of proteins,[135,136] nucleic acids,[137-139] glycans,[140] lipids[141] and even generalized FF for organic molecules.[142] Even though they are not explicitly described here, it is worth noting that similar refinements have been developed in other biomolecular FFs, such as CHARMM[143] and GROMOS.[144,145]

Current efforts are being made for the explicit inclusion of induction effects using a variety of polarization formalisms,[146] which should lead to a better representation of the molecular response to the electric fields created by the anisotropic biological environments. Thus, efforts have led to the implementation of the Drude model in CHARMM[147,148] and the induced dipole in AMOEBA[149] and AMBER.[150,151] We limit ourselves to remark the crucial effect played by the inclusion of induction effects to describe the lipid interface dipole potential, which is responsible for the difference in permeability between negative and positive hydrophobic ions.[152]

To set-up the simulation system, one has to specify the nature and composition of the environment, such as the nature of the solvent and the ionic atmosphere. It is now conventional to utilize periodic boundary conditions, which permits to replicate the system periodically in all directions, so that the atoms in the central simulation box are imaged in each of the surrounding boxes. This leads to the simulation of a pseudo-infinite system, as this procedure permits each atom to be surrounded by a sizable set of neighbouring atoms located either in the central box or in the replicated ones. Furthermore, the physical conditions adequate for the simulation, such as pressure or temperature, must be considered, which will define the ensemble of all microscopic states belonging to a single thermodynamic state, generally corresponding to the canonical (constant temperature and volume) or isobaric-isothermal (constant pressure and temperature) ensembles.

Sampling of the simulated system. The dynamical description of biomolecular systems encompass a large variety of time scales, ranging from extremely fast processes, such as the rotation of solvent-exposed side chains, to extremely slow motions, such as the folding of proteins. The fastest motions correspond to the vibration of bonds, which occur in the time scale of femtoseconds, thus limiting the time step to be used in the integration of Newton's equation of motion in atomistic MD simulations. Accordingly, solving the iterative numerical calculation of the instantaneous forces

must be repeated a huge number of times in order to built up a trajectory representative of the accessible microstates for the chosen simulation ensemble.

Freezing the fastest motions by using suitable constraints, such as those implemented in SHAKE[153] and LINCS,[154] is a widely used to increase the efficiency of computer simulations, as the bond vibrations have generally little interest in conformational analysis of biomolecules. On the other hand, a significant fraction of the computational cost comes from the evaluation of interatomic interactions, specifically due to the long-range nature of Coulombic interactions. As the size of the system is increased in order to incorporate more realistic descriptions of the physiological environment, the number of pairwise interactions to be considered may limit the overall computational efficiency. To overcome this limitation, the spherical truncation approach, by which interactions beyond a given cut-off distance was ignored, was first implemented. However, the requirement of energy conservation and the occurrence of unstable simulations led to the adoption of the Ewald summation method,[155] which provides a more rigorous framework for the calculation of electrostatic interactions in infinite periodic systems.

In addition to the developments and refinements in MD algorithms, the advances in computer power have contributed decisively to enhance the sampling capabilities of MD simulation codes. The last versions of MD software packages, such as NAMD[156] and GROMACS,[158] have achieved a high performance for exploiting parallelized computing. The development of MD-dedicated massively parallel supercomputers, such as the ANTON computer,[159] has also led to an outstanding enhancement in computing power. On the other hand, the high arithmetic performance and intrinsic parallelism of graphical processing units offers an alternative technological edge for enhancing the efficiency of MD simulations, as reflected in the GPU-oriented versions of AMBER[160,161] and ACEMD[162] codes.

Representative examples of the impact of improvements in FFs, computational algorithms and simulation speed come from the studies that allow direct observation of the dynamics of folding events in atomic detail.[163,164] These studies are informative about the structure and stability of the folded state, the heterogeneity of folding pathways, the origin of the barriers implicated in protein folding, and the nature of misfolded states. Similarly, MD simulations have shown the spontaneous binding of ligands to their targets without imposing any prior bias toward the binding pocket,[165-168] which are valuable not only to identify the binding pathway and energetic barriers that

determine binding kinetics, but also to disclose alternative binding sites that might facilitate the design of allosteric drugs.

Simplified treatments of biomolecular systems. Generally the time resolution of experimental techniques cannot go beyond the millisecond-microsecond resolution, while atomistic MD simulations typically cover hundreds of nanoseconds for systems that extend less than 10 nm. There is then urgent need for developing methods capable of bridging the time scale of experimental and computational techniques for realistic simulation systems.

A promising approach to reduce the cost of simulating large macromolecular systems consists in the use of coarse-grained models,[169,170] where 'classical' particles (beads) now represent typically clusters of atoms (e.g., an amino acid is represented by two beads, one for the backbone and the other for the side chain). These models lead to a significant reduction in the degrees of freedom of the system, thus enabling to overcome limitations in size and time scales encountered by atomistic simulations. Models have been developed for different kinds of constituent units in biomolecules within the residue-based coarse-grained approach, including nucleic acids, lipids, proteins and even water.[171-176] Nevertheless, at a coarser level of description, shape-based models, where beads represent protein segments, whose dynamics is described using MD simulations as well, have also been developed.[177,178]

A representative example of the coarse-grained force field is MARTINI.[179] It contains four main types of particles: polar, non-polar, apolar, and charged. Within each type, subtypes are distinguished depending on hydrogen-bond capabilities or the degree of polarity, giving rise to a total of 18 particle types. Non-bonded interactions are described by a Lennard-Jones potential, whereas charged groups also interact via a Coulombic energy function. The non-bonded interactions have been parameterized based on a systematic comparison to experimental thermodynamic data, including free energies of hydration and vaporization, and the partitioning free energies between water and a number of organic phases. Bonded interactions are treated by a standard set of potential energy functions, including harmonic bond and angle potentials, and multimodal dihedral potentials. In this case, parameterization is made using structural data derived from template geometries or from atomistic simulations. Applications of the MARTINI force field cover a wide range of topics, such as lipid polymorphism,

lipid membrane properties, membrane protein oligomerization, pore formation in membranes, and the interaction of nanoparticles with membranes.[180] Finally, it is also worth mentioning the package ESPResSO,[181] which has been designed as a versatile software for coarse-grained atomistic or bead-spring models of soft matter. The code has been utilized to study membrane remodeling and vesiculation by curvature-inducing model proteins.[182]

Ligand-based techniques: QSAR

> *"A pharmacophore is the ensemble of Steric and Electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response".*
>
> **IUPAC meeting, 1998**

Since its development about 50 years ago, the quantitative structure-activity relationship (QSAR) paradigm has revealed to be a very useful approach in many scientific fields such as medicinal chemistry, agrochemistry, food and environmental toxicology and generally, in the wide scenario of life sciences.[183] Its principal aim consists on modelling biological phenomena characterized by different level of complexity, from *relatively* simple local events[184,186] (inhibitory/binding activities in guest-host interactions) to more complex physio-pathological phenomena[4] (mutagenicity, cancerogenicity, metabolism, ADMET profiles…). The fate of this approach has been clearly determined by two important factors:

- ✓ The assumption that the biological effects exerted by a molecule can be related to its physico-chemical properties defined through the use of suitable, and case-specific molecular descriptors.
- ✓ The rapid evolution of software/hardware performances that allowed the advent of bioinformatics and computational techniques.

In this context, the work of Hansch and Muir[187] on the SAR of plant growth regulators and their dependency on Hammett constants and hydrophobicity can be considered the catalyst for a rapid development of new parameters and equations for QSAR application. The early approaches were based on molecular decomposition of such chemical properties as hydrophobicity (π), leading to fractional contributions (hydrophobic constants) to octanol/water logP. Combined with the previously discovered Hammett's electronic constants (σ),[188] this allowed to analyze the capability of polar and non-polar contributions in modeling certain biological phenomena, according to the following equation:

$$Log\frac{1}{C} = a\,\sigma + b\,\pi + ck \qquad (1)$$

Since the first crystallization of haemoglobin in 1864, optimization of X-ray and other spectroscopic techniques allowed to define and collect the structure of several thousands of biological targets (http://www.rcsb.org/pdb/statistics/holdings.do#), among proteins and nucleic acids, giving the possibility of examining the mutual complementarity in guest-host interactions. This has been not deleterious for QSAR approaches, but rather contributed to further legitimate its election as valuable and reliable approach in modelling real phenomena.

The pharmacophore approach

QSAR studies rely on the general concept of correlation between structure and biological activity of molecules: the pharmacophore approach. A pharmacophore has to be considered as a global 3D representation of the optimal steric and electrostatic properties for ligand-target interaction and not as a real molecule or association of functional groups. The generation of a QSAR model is done as a post-processing step of pharmacophore generation.

From 1D to xD-QSAR: what can be still implemented

Since their formalization, QSAR techniques were subjected to rapid evolution in terms of type of descriptors that can be used and methods that can be applied for data mining. The dimensionality of QSAR approaches is represented in **Table 1** (reproduced from Vedani and co-workers).

In 1D-QSAR, several molecular properties have to be detected to define those that are relevant descriptors of SAR studies. Various parameters are used to select the descriptors that define specific molecular properties, such as electronic, hydrophobic (ClogP) and steric (STERIMOL parameters or Molecular Volume as a measure of the overall bulk of a molecule) features.

Dimension	Method	Protein
1D-QSAR	Affinity correlated with pKa, logP, electronic properties,..	NO
2D-QSAR	Affinity correlated with structural patterns as connectivity, 2D pharmacophore..	NO
3D-QSAR	Affinity correlated with the 3D structure of the ligands	Possible
4D-QSAR	Ligands are represented as an ensemble of conformers, orientations, protomers, and stereoisomers	Typical
5D-QSAR	as 4D-QSAR π representation of different induced-fit models (dual-shell models in addition allow for anisotropic induced fit)	YES
6D-QSAR	as 5D-QSAR π representation of different solvation scenarios	YES

Table 1. Dimensionality of QSAR approaches.

1D-QSAR is typically focused on macroscopic properties as additive contributions to biological activity of a molecule. Conversely, 2D-QSAR attempts to analyse the specific molecular fragment's properties. Being a knowledge-based approach, it requires some constitutional information about the compound while constructing the model. The linear and nonlinear methods like multiple linear regression method, genetic

algorithm (GA), and partial least squares (PLS) techniques are then used to predict the QSAR model.[190,191]

Starting from 1980, grid-based approaches contributed to the development of 3D-QSAR techniques. For the first time, it became possible to describe molecular properties from a thee-dimensional and *atomic* point of view, reaching a more local and detailed level of representation. Cramer and co-workers[189] inaugurated fields-based *era* in 1988 with Comparative Molecular Fields Analysis (CoMFA), followed by Klebe and co-workers[192] with Comparative Molecular Similarity Analysis (CoMSIA). Generally speaking, these techniques allowed to (i) consider the molecules in their supposed bioactive conformation, and (ii) discern among favourable and unfavourable regions for ligand-receptor interactions from steric and electrostatic properties.

Starting from a defined molecular set, a preliminary setup of the system is needed in order to correctly calculate all the selected properties. The wide protocol is shown in **Figure 1** and consists on:

- Selection of the molecular set (training, validation and/or test sets)
- Geometrical optimization
- Conformational analysis
- Molecular alignment
- Grid generation
- Calculation of all values for the selected descriptors
- Regression analysis (PLS) and Validation (leave-one-out, LOO, and related techniques): evaluation of predictive performances in interpolation (training set) and extrapolation (test set) of biological data.[193]

The calculation of properties from suitable descriptors can be dependent or not on the alignment of bioactive conformers. For this reason, 3D-QSAR methods can be ideally divided in alignment-dependent and alignment-independent.

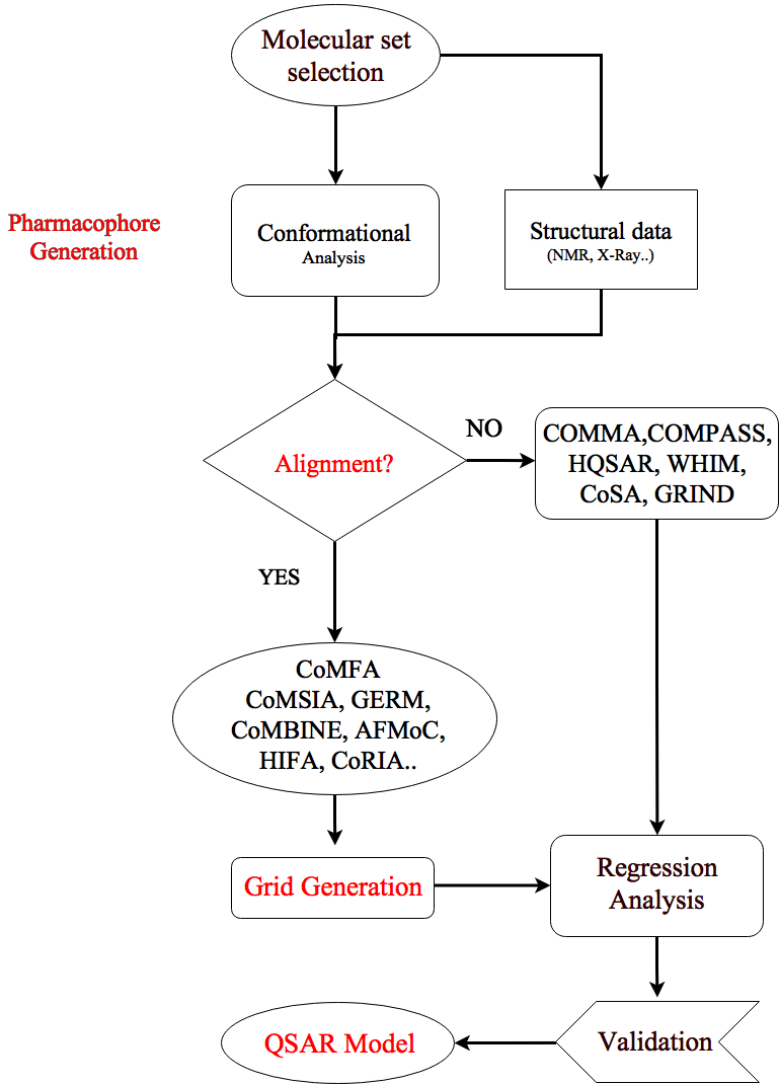


Figure 2. Flow chart of the main phases for a classical QSAR study.

- ✓ Alignment-dependent QSAR method are Comparative Molecular Field Analysis (CoMFA), Comparative Molecular Similarity Indices Analysis (CoMSIA), Genetically Evolved Receptor Modeling (GERM), Comparative Binding Energy Analysis (CoMBINE), Adaptation of the Fields for Molecular Comparison (AFMoC), Hint Interaction field analysis (HIFA) and Comparative Residue Interaction Analysis (CoRIA).[194-195]
- ✓ Alignment-independent QSAR methods are Comparative Molecular Moment Analysis (CoMMA), COMPASS, Holo-QSAR (HQSAR),

Weighted Holistic Invariant Molecular Descriptors (WHIM), Comparative Spectral Analysis (CoSA) and Grid Independent Descriptors (GRIND).

In order to reduce and extrapolate a reasonable and statistically significant number of variables to correlate with the biological activity, a statistical method is required. These methods attempt to find a correlation between activity (y-variable) and molecular descriptors (dependent or independent x-variables). Different methods have been utilized, like simple linear regression (LRA), multiple linear regression (MLR), principle component analysis (PCA), Principle Component Regression (PCR), PLS analysis, GFA, Cluster analysis, Artificial Neural Networks (ANN) and k-Nearest Neighbour (k-NN) method. The choice of the better techniques depend on the dimensionality of dependent and independent variables, and PLS is one of the most common adopted strategies in 3D-QSAR methods.

In general, the choice of conformations to be used for molecular alignment as well as the type of alignment are the main limitations of 3D-QSAR techniques. All grid-based methods are very sensible to some factors as grid dimension, grid spacing, alignment procedure, and orientation of aligned molecules into the grid.

Recently,[196] a new type of QSAR has been implemented to overcome the limitations of classical 3D-QSAR, named receptor independent (RI 4D-QSAR) and receptor-dependent (RD 4D-QSAR). These techniques are also called "ensemble sampling" and include other features (multiple conformations, orientation, and protonation state of ligand molecule) and parameters (MD simulations of reference molecule, temperature, size of the ensemble sampling). RI 4D-QSAR attempts to find a SAR starting from specific libraries of bioactive ligand conformations. Similarly to 3D-QSAR, the 4D pharmacophore model only collects information that comes from ligands. The basic aim of the RD 4D-QSAR studies is to map the ligand-receptor interaction mode: 3D pharmacophore is evaluated thermodynamically by accessing conformational states of the receptor ligand complex. In this context, other elements of the binding cavity as water molecules can be taken into account. The addition of a fifth (5D-QSAR)[197] and a sixth (6D-QSAR)[198] dimension is intended to account for solvation effects.

Statistical suitability of a QSAR model

For a linear model, the estimated activity (in log units), $y_{n,pred}$ of a molecule is a function of its geometric fit on the pharmacophore, $x_{n,}$ according to the following equation (Eq.2):

$$y_{n,pred} = ax_n + b \tag{2}$$

In a linear regression model, a typical index of how well the model fit the experimental data is represented by the squared value of the correlation coefficient (R^2). It goes from 0 (no fit) to 1 (perfect fit) and is sometimes defined also as the percentage of variance explained (ESS) by the statistical model respect to the total variance (TSS), expressed as Sum of Squares (Eq.3).

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{n=1}^{N}(y_{n,pred} - \bar{y})^2}{\sum_{n=1}^{N}(y_{n,obs} - \bar{y})^2} \tag{3}$$

The variable $y_{n,pred}$ refers to the predicted activity for each value (n), $y_{n,obs}$ refers to the real experimental data, and $\bar{y}$ is the arithmetic mean over all $y_{n,obs}$.

After model construction, a validation is needed in order to verify its robustness and stability. The majority of the QSAR models are validated through the application of the LOO, a cross-validation procedure where the statistical model is iteratively re-calculated after a random exclusion of one value a time. The cross-validation is measured by Q^2 (a sort of cross-validated R^2) and S_{PRESS} (Standard deviation error in prediction), which are commonly regarded as ultimate criteria of both robustness and predictive ability of the model. A more robust method is the LMO (leave-More-Out) where the model is recalculated after exclusion of a randomly defined percentage of values. Some studies have demonstrated that a real correlation between LOO cross-validated Q^2 for the training set and R^2 for test set compounds doesn't exist. Therefore, it is important to use both cross-validated models and external test set to correctly validate a QSAR model.[199]

Best model is selected following a criterion of lowest standard deviation error in prediction of the actual experimental values (PRESS) corrected by the number of latent values (n – p – 1) of the model, S_{PRESS}. S_{PRESS} represents not only an index of the predictive quality but also a measure of the parsimony of the model. The formula of PRESS and S_{PRESS} are reported in eq. 4 and 5, respectively:

$$\text{PRESS} = \sum_{n=1}^{N}\left(y_{n,obs} - y_{n,pred}\right)^2 \quad (4)$$

$$S_{PRESS} = \sqrt{\frac{PRESS}{n-p-1}} \quad (5)$$

As suggested by Alexander D.L. and co-workers,[200] a way to measure the level of precision in predicting the biological activity, especially for the test set compounds, is the RMSE (Root-Mean-Square Error). It expresses a standard deviation of residuals and is obtained by dividing PRESS (Predicted REsidual Sum-of-Squares) only for the number (n_o) of observations (Eq. 6).

$$\text{RMSE} = \sqrt{\frac{PRESS}{n_o}} \quad (6)$$

Finally, to be really descriptive and informative, a statistical model must demonstrate good predictive properties from retrospective (predictive performances for training set compounds) as well as prospective (predictive performances for test set compounds) points of view.

Target-based techniques: Ligand Docking

Definition: A structure-based approach that tries to predict the structure of the more favourable intermolecular complex formed between a relatively (potential) small ligand and a macromolecular target (Protein, DNA, RNA…) from both geometrical and energetic point of view.

In the past decades, the explosion of the protein structures directly retrievable from structural repositories as the Protein Data Bank (PDB)[201] have facilitated the development of computational techniques to be used in a biological/biomedical context.

Computational (target-based) simulations help in the identification of relevant "bioactive regions" of the chemical space, allowing to (i) discover structural features implicated in ligand-target interactions and (ii) analyse them at sub-molecular level. The correct prediction of the binding mode for a ligand-protein complex is of great importance in Structure-Based Molecular Modeling (SBMM) studies in order to perform a preliminary filtering of molecules generally stored into ligand databases as DUD,[202] DrugBank,[203] ZINC[204] and select only the best candidates to carry out a specific biological effect. These compounds are selected as starting "points" for a further structural optimization and *in vitro* testing. This process is generally known as *lead discovery and optimization*.

SBMM have had a slow start due to the limitation of the computational power (hardware/software facilities) and high-resolution 3D structures. In particular, two important steps can be identified during the *evolutionary path* of Molecular Docking. In 1985, Goodford[205] presented the software GRID. The method allowed to (i) study ligand-protein interactions and (ii) discern among favourable and non-favourable regions by using chemical probes that explore the protein binding sites to define the Molecular Interaction Fields (MIFs). In 1988, Kuntz and co-workers[206] provided the first rigid body docking based simply on the geometrical ligand-target complementarity. Since that time, a vast number of docking programs have been developed.

Docking is a computational technique aimed at the prediction of the most favourable ligand–target spatial configuration and an estimate of the corresponding

binding affinity, although as stated at the beginning accurate scoring methods remain still elusive. In the first step a conformational search algorithm explores the possible ligand conformations (poses) inside the protein-binding pocket. In the second step, a scoring function is applied to evaluate and select the most favourable pose.[207]

These two steps represent the main challenge in lead discovery and optimization. Different search algorithms and scoring functions were developed and can be chosen and applied. They allow to (i) generate suitable conformations for the ligand into the binding site and (ii) rank them in order to elect the best geometric/energetic compromise for ligand-target complex.

Some important preliminary steps have to be respected to perform a proper docking simulation, such as the definition of protomers and tautomers, stereoisomers, and conformations. Depending on the level of flexibility achieved during the simulation, docking can be classified in:

- Rigid Docking: both ligand and target are considered as rigid bodies (only the six degrees of freedom of the ligand, corresponding to rotation and translation, are taken into account).
- Semi-flexible Docking: the conformational flexibility of ligands is also taken in account while the protein is kept rigid.
- Flexible Docking: it includes flexibility of a limited number of amino acid side-chains by exploiting rotamers libraries. Flexibility at the target level can also be included coupling the docking with a MD program. The flexibility of the backbone in such protocols is included, but with the risk of possible inaccuracies due to the FFs.

During docking, the search algorithms aim to explore the conformational space of the ligands inside the protein active site in an efficient and fast fashion. Simulating ligand flexibility is a computationally expensive process. In order to reduce the computational costs, several strategies were developed. They can be classified in:

- Site-Point Search: An example of this method is Glide.[208] It approximates a complete systematic search of the conformational, orientational, and positional space of the docked ligand. In this search, an initial rough positioning and scoring phase that dramatically narrows the search space is

followed by flexible energy optimization on an OPLS-AA non-bonded potential grid for a few hundred surviving candidate poses. The very best candidates are further refined via a Monte Carlo sampling of pose conformation.

- Fragment-based methods: the ligand is divided into fragments and incrementally built into the binding site. A more efficient incremental method is introduced, for instance, in FlexX.[209]
- Genetic Algorithms: they work by representing the ligand conformation in a modular way, using operations similar to mutations and crosses. The quality of the results is a function of the starting genes, the number of evolutionary events, i.e., the mutations and crosses, and the scoring function to pick the more favourable conformers.[210] An example of this approach can be found in GOLD[211] docking suite, whereas a modification of the classical search algorithm (Lamarckian-based genetic algorithm) that allow a more rapid research has been integrated in the latest versions of AutoDock.[212]

Consensus scoring

Scoring functions can be grouped in three families: MM FF, empirical and knowledge-based scoring functions.

Energy scoring functions pursue to estimate the binding free energy of the ligand to the protein target active pocket. Unfortunately, scoring is the weakest step in docking methodologies for, at least two reasons. First, for a given list of molecules to be ranked, in the majority of the cases it is unable to even rank-order a hit list. Secondly, it is very difficult to discriminate binding modes among the poses proposed for a given molecule.

Charifson and co-workers[213] conducted an extensive computational study in which they showed that combining-scoring functions (*consensus scoring*) in an intersection-based consensus approach results in an enhancement in the ability to discriminate between active and inactive enzyme inhibitors. An analysis of two different docking methods and 13 scoring functions provides insights into which functions perform well, both singly and in combination. The consensus scoring further provides a dramatic reduction in the number of false positives identified by individual scoring functions, leading to a significant enhancement in hit-rates.

The success of a docking-based virtual screening drastically depends on the accuracy and ability of a scoring function in correctly ranking ligand poses generated from the search algorithms. More efforts have been made to reach to a higher level of efficiency and accuracy in high-throughput screening. They can be rapidly listed as follows:

- Use of non-linear scoring function[214]
- Ensemble docking
- Re-ranking
- Application of knowledge-based strategies[215]

Different studies have shown that docking methods perform well in reproducing ligand-binding poses for experimentally derived structures; however, they can fail if a protein structure was solved in the presence of a very different compound.[216] This is probably due to the fact that receptor conformational changes (induced-fit) could be necessary to correctly accommodate these ligands, leading to incorrect evaluations.

Ensemble docking came from the necessity to account for protein flexibility during pose generation: a ligand is docked into several protein structures/conformations in order to identify the best-scored pair of protein conformation and ligand binding mode.[217] Re-ranking/re-scoring strategies are also applied to improve the success rate in lead discovery. They include re-evaluation of the generated ligand poses through application of more sophisticated methods as linear interaction energy (LIE), molecular mechanic/Poisson−Boltzmann (generalized Born) surface area (MM-PB(GB)SA), free-energy perturbation (FEP), and thermodynamic integration (TI).[218-221]

HINT: hydrophobicity as driving force for estimation of ligand-target interactions

It is very difficult to capture the physico-chemical essence of hydrophobicity. Since its first introduction in QSAR by Hansch and colleagues,[222] the concept of hydrophobicity has proved to be very useful in describing the ability of a molecule to "interact" with a polar phase. From a biochemical point of view, it is an important driving-force in protein folding and protein binding interactions.[223] In this context, a

empirically derived re-scoring method, also implemented in Sybyl (TRIPOS) platform is HINT (Hydrophatic INTeraction) function, which uses empirically-derived atomic hydrophobic constants and allows to consider both enthalpic and entropic effects, generally neglected by other types of scoring functions.

HINT model[224] scores each atom–atom interaction within (*IntraMolecular*) or between (*InterMolecular*) biological molecules with the following equation (Eq.1):

$$b_{ij} = a_i S_i a_j S_j T_{ij} R_{ij} + r_{ij} \quad (1)$$

where $\mathbf{b_{ij}}$ is the interaction score between atoms i and j, a is the hydrophobic atom constant, **S** is the solvent accessible surface area (H_2O probe), $\mathbf{T_{ij}}$ is a logic function described below, and $\mathbf{R_{ij}}$ and $\mathbf{r_{ij}}$ are functions of the distance between atoms i and j (i.e. r). Generally, the hydropathic-dependent function, $\mathbf{R_{ij}}$, is the simple exponential $\mathbf{e^{-r}}$ and $\mathbf{r_{ij}}$ is an implementation of the Lennard- Jones potential function. The $\mathbf{r_{ij}}$ term is mostly a penalty function to flag van der Waals violations. The double sum, $\mathbf{\Sigma\Sigma b_{ij}}$, is the total interaction score for the system. The HINT convention is that favourable interactions are scored with $\mathbf{b_{ij}} > 0$ and unfavourable interactions are scored with $\mathbf{b_{ij}} < 0$. The logic function $\mathbf{T_{ij}}$ returns a value of 1 or −1 depending on the character of the interacting polar atoms (i.e.,. $a < 0$). There are three possibilities: acid–acid, acid–base, or base–base; only acid–base is scored favourably. $\mathbf{T_{ij}}$ also flags hydrogen bonds that are in the HINT model a special case of acid–base interactions. The hydrophobic atom constant, **a**, is the key parameter in the HINT model: it is calculated with an adaptation of the CLOGP[225] method of Hansch and Leo.

The hydrophobic effect is computed by the hydrophobic atom constant, a_i. When the key HINT parameter, a_i (the hydrophobic atom constant) is greater than zero the atom will favourably interact with another atom whose a_j is greater than zero, a hydrophobic–hydrophobic interaction. If a_i and a_j are both positive, implying that they are both hydrophobic, this is scored favourably as a hydrophobic interaction. Similarly, if a_i and a_j are both negative, and one is a Lewis acid while the other is a Lewis base, this i–j interaction would also be scored favourably like a hydrogen bond donor or acceptor group on a ligand establishing interactions and solubility in water. Similarly, energetic effects related to solvation/desolvation must be implicit in $LogP_{o/w}$ data.

The strength of hydrophobic interaction depends on distance between atoms i and j. HINT allows to manually define i-j distance with a selection of a specific "Distance Function", generally considered as exponential in default calculations. According to Israelachvili and Pashley theory,[226] hydrophobic interactions decay exponentially with distance.

The hydrophobic field is calculated according to Eq.2:

$$A_t = \sum_{i=1}^{M} s_i a_i R_{it} \quad (2)$$

where s_i is the solvent accessible surface area, a_i is the hydrophobic atom constant for atom i and R_{it} is the function of distance between i and the test point t; in this case, $R_{ij} = e^{-r}$.

Given that $LogP_{o/w}$ is derived almost directly from the results of a real experiment, it includes all energetic effects (Eq.3).

$$logP_{o/w} = -\frac{\Delta G}{2.303} RT \quad (3)$$

Since

$$\sum a_i = logP_{o/w} \quad (4)$$

then a_i (the hydrophobic atom constants including Leo factors) are also directly related to ΔG, i.e. *having both enthalpic and entropic components*. The a_i values are dimensionless parameters directly related to the free energy of atom transfer (as a part of a specific solute molecule) between two solvents, water and n-octanol.

In conclusion, still today, docking strategies suffer of some important indirect/direct limitations as:

- Inaccuracies from crystallographic structures: the identity of the isoelectric nitrogen and oxygen of the side chains of asparagine and glutamine, the position of whole flexible residues, like lysine and glutamate, especially at the protein surface, of mobile loops, and even flexible parts of the ligand, the identification and location of water molecules, which are often isoelectronic to common buffer constituents in crystallization media, the influence of crystal packing on the target structure, and the ionization state of key residues or the ligand due to the difficulty in identifying the hydrogen atoms in X-ray structures.
- Scoring function inaccuracies: only enthalpic (binding) effects are considered while entropic ones are generally neglected, and the difficulty in describing non-classical types of interaction (cation-π interactions, charge transfer interactions, hydrogen bonding to π-systems, halogen bonding, orthogonal dipolar alignment, dipolar antiperiplanar interactions, π-stacking, π edge-to-face contacts, hydrogen bonding involving CH groups).
- As discussed by Tirado-Rives and Jorgensen,[227] these limitations also originate from the "experimental" fields: the so-called tiny "window of activity" for *in vitro* tested compounds makes very difficult to effectively model and discriminate the binding free energy of compounds within a ligand-target biological context.
- In general, docking programs do not allow to fully address the need of a deep analysis of ligand-target reciprocal adaptation also known as *induced-fit*. They only offer a snapshot of a time-dependent phenomenon and does not account for global stability of the proposed interaction.

Quantum Mechanics (QM): ab initio, Density Fuctional Theory and Semi-Empirical calculations.

Definition: theory of matter that is based on the concept of the possession of wave properties by elementary particles, that affords a mathematical interpretation of the structure and interactions of matter on the basis of these properties, and that incorporates within it quantum theory and the uncertainty principle.

QM methods can be partitioned into four classes: (i) standard SCF methods like Hartree-Fock (HF) and Density Functional Theory (DFT), (ii) the so-called post-HF methods, which include correlation effects on top of a HF calculation, and (iii) semi-empirical (SE) methods, which can be formally derived from the HF method by approximating and neglecting many of the costly integrals and which usually apply a minimal basis set.

To provide a proper contextualization of the *work environment,* the effective discussion of the methodologies enumerated before will be preceded by a rapid overview on the basic physico-chemical and mathematical concepts of QM calculations.

In QM the state of a chemical system can be described by a wave function as a solution of the time-independent Schrödinger Equation (Eq.1).[228]

$$\hat{\mathrm{H}}\Psi = E\Psi \qquad (1)$$

where the Hamiltonian operator, $\hat{\mathrm{H}}$, is associated to the total energy E, of the system as a sum of kinetic and potential component.

To solve the Schrödinger equation would mean to reach a complete physico-chemical description of the system but this is impossible for polyatomic systems. Therefore, some simplifications have to be introduced. A commonly made simplification is the Born-Oppenheimer approximation, which separates the motions of nuclei and electrons. Thus, electrons are assumed to move around the fixed nuclei and only their energetic contribution is taken into account, while the kinetic energy of the nuclei is neglected. In this context, electrons re-adapt their position as a function of the nuclear configuration, creating a direct correspondence between the electron distribution and the position of nuclei in the system

It is also assumed that each electron moves into the average field of all other electrons, generating the self-consistent field (SCF), generally described by the following HF equation (Eq.2).

$$f_i\chi_i = \varepsilon_i\chi_i \quad (2)$$

where χ_i refers to the wave function for each electron, *i wi*th energy ε_i, whereas the Fock operator, f_i, is a one-electron Hamiltonian operator that depends on its solutions, χ_i and is expressed as a sum of kinetic and potential energies, according to Eq.3.

$$f_i = -\frac{1}{2}\nabla_i^2 - \sum_A^M \frac{Z_A}{r_{iA}} + v_i^{SCF} \quad (3)$$

where the first term represents the kinetic energy of the electron, *i*, the second term account for its potential energy in the field of the nuclei (*M* nuclei with nuclear charges Z_A) and finally, the third one, v_i^{SCF}represents the potential energy in the self-consistent field of the other electrons.

The Hartree-Fock method

The HF method is applied to solve the simplified version of the Schrödinger equation as reported in the Born-Oppenheimer approximation. For multi-electronic systems, each electron is described as an asymmetric product of the one-electron wave function (orbitals), generally known as Slater determinant. In this way, a molecule can be described as a linear combination of atomic orbitals. The so described atomic orbitals are subjected to an iterative minimization process in order to find the minimal energy value for each Slater determinant. The process stops when a convergence threshold value is reached. At the end of the process, a set of "stable and minimized" self-consistent one-electron orbitals is obtained for the electronic ground state. In brief, the process can be summarized as follows: initial guess of molecular orbitals, Fock matrix formation, diagonalization and, if the SCF convergence is achieved, molecular properties calculation.

The assumption of an average field of the electron neglects correlation between the electrons. However, the HF method is a good starting point for either refined post-HF methods, which yield more accurate results at the cost of increased computational effort, or for simplified treatments aimed at reducing the computational expense.

Post-HF methods

This term encodes a variety of methods designed to take into account the electron correlation as a correction to the HF solution. Possibly Configuration Interaction is the most natural method to account for electron correlation. To this end, a HF calculation is performed. From the resulting determinant, a series of excited states generated by promoting electrons from occupied to virtual orbitals in all possible ways are generated. If only one electron is moved, then singly excited determinants are obtained, if two electrons are promoted, then doubly excited determinants result, and so on. The Hamilton matrix is set up on the basis of these determinants. The matrix is diagonalised to obtain the eigenvalues (energies) and eigenfunctions (wave functions).

Since the complete set of basis is enormous, this procedure is generally not practicable, and then one has to resort to other strategies, such as perturbative methods, where electron correlation is treated as a perturbation to the HF solution. This strategy has led to the series of MPx methods, which have been widely used in the study of (bio)organic compounds.

Density Functional Theory

The majority of biologically relevant systems have "atomic" dimensions that make impossible the application of classical *ab initio* QM calculations. In these cases, the use of a more simplified method is needed. The DFT represents the workhorse for large-scale correlated QM studies. DFT uses electron density to derive the QM properties of the system instead of the N-dimensional wave functions. In the early sixties, Hohenberg and Kohn[229] demonstrated that the general properties of a given ground-state molecule could be described as a function of its electronic density (*functional*), which only depends on three spatial coordinates for *all* the electrons. In this context, the total energy of a system in its ground state can be described as a *functional* of the electron density.

The basis of most functionals is the Local Density Approximation (LDA). In this approach, only the electron energy at specific points is considered. One example of this approximation is the simple Thomas-Fermi model.[230,231] According to this theory, the exchange/correlation energy functional in a uniform electron gas for each electron is given by the Eq.4.

$$E_{XC}^{LDA}[\rho] = \int \rho\,(r)\varepsilon_{XC}^{unif}\,(\rho)\,dr \qquad (4)$$

where ε_{XC}^{unif} represents the energy for each electron.

Another example of a semi-local functional is the Generalized Gradient Approximation (GGA), which also consider the changes in electron density at specific points.

The HF theory can in principle provide the exact exchange energy when correlation effects are ignored. So, in order to reach a higher level in orbitals description, the HF expression for the exchange energy can be combined with the density functionals. This led to the development of the *hybrid functionals.*[232] For chemical applications, many hybrid functionals have been proposed. One of the most popular hybrid functionals is B3LYP (Eq.5), based on the Becke's[233] three-parameter functional, B3, in combination with the LYP correlation functional.

$$E_{XC}^{B3LYP} = a_0\,E_X^{HF} + (1-a_0)E_X^{LSDA} + a_X\,\Delta E_X^{B88} + (1a_c)E_C^{VWN} + a_C E_C^{LYP} \qquad (5)$$

where a_0, a_X and a_C are the three parameters from the B3 formulation.

Semi-empirical Quantum Chemistry

The SE methods were developed with the aim to reduce the computational costs of HF calculations and are based on the HF formalism. Generally, the following concepts are applied:

- ✓ Reduction of the basis set. Only the valence electrons are explicitly treated while inner electrons are described via a semi-empirical atom core. For the valence electrons, a minimal basis set is used.

- ✓ Neglect of differential overlap, meaning the basis functions do not overlap under certain circumstances. In turn, many of the cumbersome two-electron integrals are neglected.
- ✓ Replacement of most of the remaining integrals by simple parameterised functions. A sensible parameterisation should compensate for the simplifications made before.

By taking these as starting points, several strategies have been developed. Many SE methods are based on the Neglect of Diatomic Differential Overlap (NDDO) approximation,[234] which has given rise to methods such as the Modified Neglect of Differential Overlap (MNDO),[235] the Austin Method 1 (AM1)[236] and the Parameterised Model 3 (PM3),[237] of which the most representative is the PM6 parameterisation.[238] PM6 is parameterised based on a large experimental data set of over 9000 compounds and includes currently 83 elements, up to bismuth. The other levels of approximations are Intermediate Neglect of Differential Overlap (INDO) and Complete Neglect of Differential Overlap (CNDO). Despite the mature age of these approximations, for example, Zerner's Intermediate Neglect of Differential Overlap, ZINDO[239] has recently found new applications in the study of electronic coupling terms in biological electron transfer processes[240,241] as well as energetics of spin state[242] and optical properties.[242,243]

The QM Continuum Solvation Models

As discussed by Tomasi and co-workers,[244] the solvation models can be probably considered the second most important innovation in the field of the QM approaches for condensed systems. This kind of approaches has in fact made possible the treatment of the solvent effects in a QM context.

The basic model is based on the following assumptions:

- The solute is described at given QM level.
- The solute-solvent interactions are limited to the electrostatic component.
- The system is very diluted: in generally it is assumed to be constituted by one solute molecule surrounded by a continuum solvent.
- No dynamic effects are considered.

All continuum models are linked to the cavity concept. The system is composed by a molecule (the solute) that is placed in a cavity, obtained within a continuous dielectric medium (the solvent). Definition of the shape and size for that cavity depends on the method applied and tries to (i) reproduce as well as possible the molecular shape and (ii) include all solute charge distribution. In this context, the contributions to the solute-solvent (A-B) interaction energy can be considered as a sum of three components: electrostatic, cavitation, and repulsion/dispersion.

Several models were proposed to describe the solute-solvent interactions: PCM[245] (Polarizable Continuum Model) and a successive reformulations as CPCM and IEFPCM (integral Equation Formalism), MST[246-249] (Miertus-Scrocco-Tomasi), BKO (Born-Kirkwood-Onsanger), COSMO[250] (Conductor-like Screening Model), COSMO-RS[251] (Conductor-like Screening Model for Real Solvents), and SS(V)PE.[252]

The MST method,[253] which has been parameterized for water, octanol, chloroform and carbon tetrachloride, computes the solvation free energy as a sum of three terms: cavitation (cav), van der Waals (vW) and electrostatic (ele; Eq. 6).

$$\Delta G_{solv} = \Delta G_{Cav} + \Delta G_{vW} + \Delta G_{ele} \quad (6)$$

The cavitation term is the work needed to generate the solute cavity into the continuum dielectric solvent. The vW term considers dispersion/repulsion interactions between solute and solvent, whereas the electrostatic is the work necessary to generate the charge distribution for the solute immersed into the solvent.

The cavitation term is a derivation of the Pierotti's scaled particle theory.[254] The vW parameter is considered linearly dependent to the solvent-exposed atomic surface (SAS). Finally, the electrostatic term is determined considering the interaction between the solute charge distribution and the reaction field's elements (the surface generated by each atom). The method gives the possibility to perform the partition of each (polar/non-polar) contribution at the atomistic level, providing a local description of the solvation properties of a given molecule. Thus, interactions are computed at the surface and then centred on each atomic nuclei. The method will be discussed in detail later.

Part 1. Preliminary in silico evaluation of endocrine disrupting effects for thioxanthone photoinitiators

Preliminary in silico evaluation of endocrine disrupting effects for thioxanthone photoinitiators

Food safety exists when all people, at all times, have physical and economic access to sufficient safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life.

1996 World Food Summit

Background

Food safety and security policy is focused on two important things: the guarantee of a proper nutrition and the minimization of the food-related toxicological risk. The latter could come from different biotic-related and non-biotic sources as microbial algal, fungal and bacterial toxins, food additives, pesticide residues and (plastic/non-plastic) food contact materials.[255] While most of them are known and regularly monitored by European Union (EU) through the application of specific guidelines for data collection and harmonization,[256,257] others have not been fully characterized yet and still lack of such regulatory policy.

Food contact materials are all materials and articles intended to come into contact with food, such as packaging and containers, kitchen equipment, cutlery and dishes. There are many types of non-plastic materials that come into contact with food, including coatings, paper and board, adhesives, printing inks and rubber.

The safety of food contact materials requires evaluation as chemicals can migrate from these materials into food. The materials should be manufactured in compliance

with common EU regulations, including good manufacturing practices (GMP), so that any potential transfer to foods does not raise safety concerns, change the composition of the food in an unacceptable way or have adverse effects on quality (for instance, taste and/or odour). However, most non-plastic food contact materials are not currently covered by specific European legislation. Toxicological issues can also arise from their *in vivo* metabolism: the prediction of these metabolites is of critical importance, as their presence in the body may give an unexpected rising of direct or indirect toxic effects. The experimental determination of these metabolites is very resource intensive. Within the past few years a wide variety of computational approaches and tools have been developed in an attempt to pinpoint the most likely site of metabolism (SOM) of a molecule and the resulting metabolites.[258-260]

In 2003, EC adopted a legislative proposal for a new chemical management system called REACH (Registration, Evaluation and Authorization of CHemicals).[261,262] One of the main points of the proposal was the development of alternative no-testing methods to avoid, or at least limit, the use of animal testing. In this context, *in silico* techniques could be well suited to screen compounds with unknown toxic profile due to their "relative" rapidity in predictions and the limited physical resources needed.[263]

It should be noted that, for a given substance, *in silico* toxicology generally does not take into account the dose and the exposure levels. Therefore, these methods have to be considered as a useful support in preliminary stages of an overall food-related risk assessment protocol. Put together, *in silico* metabolite predictors and computational tools could be an efficient strategy for preliminary evaluation of potential toxicological effects that could be exerted by chemicals and their metabolites.

In conclusion, experimental toxicity screening should be considered a multi-tier approach, with *in silico* models towards the front, followed by chemical and *in vitro* biological assays, and finally, *in vivo* animal studies.

The Case Study

In 2005, EFSA reported the case of *infant formulas* contamination by thioxanthone (TXs) photoinitiators.[264]

As reported in **Table 1**, liquid products rich in fats are more subjected to migration of isopropyl-thioxanthone (ITX). In the case of non-fatty liquid foods such as fruit juices, fruit nectars and drinks, other factors such as the presence of citrus oils, fruit fibres and pulp, could facilitate the migration of ITX by acting as carriers or co-solvents. In general, contamination seems to increase with fat content and decrease with increasing pack size. The EFSA Panel[264] noted that due to their high consumption of food per kg body weight, infants exclusively fed with infant formulae packed in cartons printed with UV-cured inks are potentially more exposed to ITX than other population groups. Despite *in vivo* genotoxicity studies, no further data on potential toxicity profile for TXs are available.

Product	Fat content (%)	Pack size (ml)	ITX* (μg/l)
UHT milk	3.8	1000	71
UHT milk	1.5	1000	92
UHT milk	0.1	1000	27
Soy milk	1.5	1000	134
Soy milk vanilla	1.5	1000	90
Soy milk & juice	0.6	1000	71
Chocolate milk	2.9	200	148

Table 1. ITX contamination levels for some milk and soy beverages..

Endocrine disrupting chemicals (EDCs) and potential EDCs are mostly man-made, found in various materials such as pesticides, metals, additives or contaminants in food, and personal care products. EDCs have been suspected to be associated with altered reproductive function in males and females, increased incidence of breast cancer, abnormal growth patterns and neurodevelopmental delays in children, as well as changes in immune function. Human exposure to EDCs occurs via ingestion of food, dust and water, via inhalation of gases and particles in the air, and through the skin. EDCs can also be transferred from the pregnant woman to the developing fetus or child through the placenta and breast milk. Pregnant mothers and children are the most vulnerable populations to be affected by developmental exposures, and the effect of

exposures to EDCs may not become evident until later in life (taken from WHO: http://www.who.int/ceh/risks/cehemerging2/en/).

This work aimed to show how *in silico* tools could be a suitable approach for prediction of toxicological behaviour exerted by food contaminants, also contributing in the explanation of their sub-molecular implications. A computational model was set up to predict the putative Androgen Receptor (AR) mediated endocrine disrupting effects of TXs chemicals and their *in silico* predicted phase I metabolites.

The applied protocol can be summarized as follows (**Figure 1**):

- Validation: evaluation of metabolite prediction and docking performances.
- Analysis: phase I metabolites prediction, computational simulation of interaction (docking) and energetic (rescoring) evaluations.
- Final report: prediction of potential AR-mediated endocrine disrupting effects for phase I metabolites and their parents.

Previously reported data on 2-ITX metabolism[265] and anti-androgen activity[266] was used as validation set respectively for making metabolic and androgen receptor binding-affinity predictions.

Metaprint2D (http://www-metaprint2d.ch.cam.ac.uk) was used for metabolic predictions, whereas GOLD docking software and HINT (Hydropathic INTeraction) scoring function[267] were used respectively for protein-ligand pose generation and their energetic evaluation.

The so generated protein-ligand complexes were then energetically re-evaluated with the HINT scoring function. Competitive ability (measured as *affinity ratio*) towards the endogenous ligand (testosterone) was determined by scaling the best interaction score found for each compound on the reference score for testosterone.

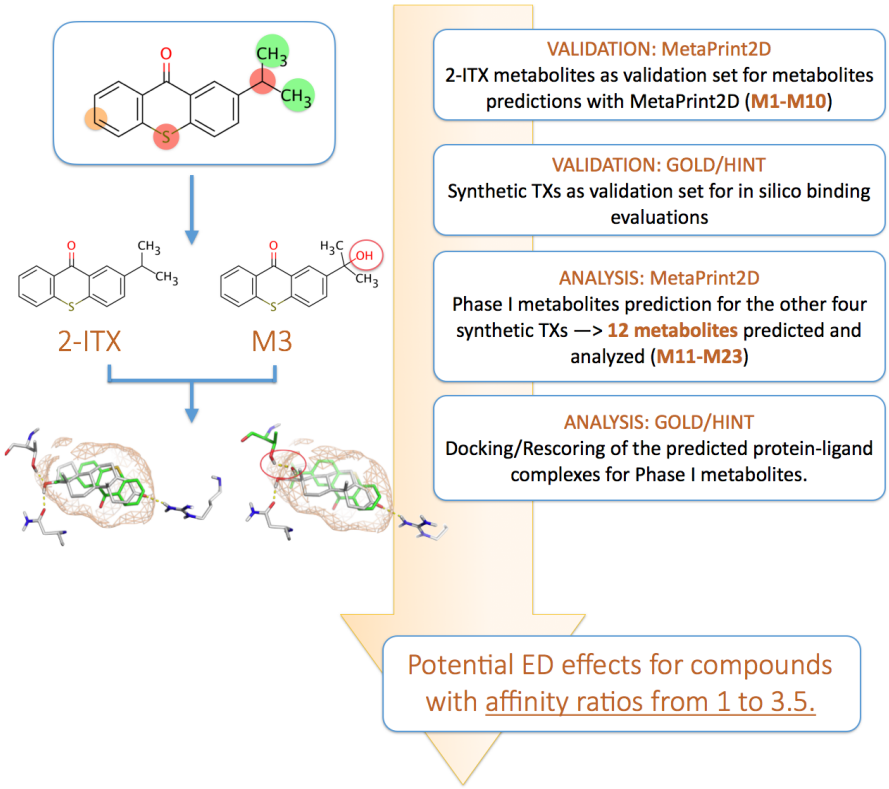


Figure 1. Schematic representation of the computational protocol applied in this study. The putative sites of metabolism are circled in red (more plausible), orange and green (less plausible).

At least six of the twelve analysed compounds are found to be capable to interact with AR-LBP.

From a structural point of view, AR-LBP is characterized by a highly hydrophobic proteic environment, delimited by some polar contacts necessary for stabilization of interaction with the natural ligand, testosterone (see **Figure 2**). These features explain the good (geometrical/energetic) accommodation of TX molecules.

In this context, *in silico* hydroxylation (see **Figure 1**) at the isopropyl moiety of 2-ITX could favour the stabilization of ligand-protein complex through the formation of H-bond interactions. Conversely, sulfoxidation is poorly tolerated because of negative hydrophobic-polar interactions generated between ligand and receptor into the binding cavity.

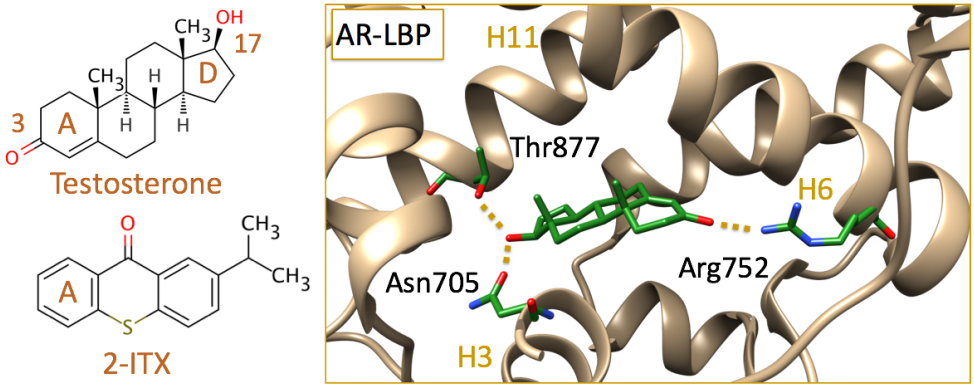


Figure 2. AR-LBP with endogenous ligand, testosterone (in green); residues involved in H-bond interactions are in green sticks.

These preliminary results suggest that an hazard evaluation focused not only on unaltered TXs, but also on their metabolites could be necessary to avoid a potential underestimation of the toxicological risk for this class of chemicals.

Twelve phase I metabolites were predicted starting from the five photoinitiators reported by Reitsma and co-workers.[266] All these structures were docked into the AR ligand-binding pocket (AR-LBP; PDB ID: 2AM9).

Paper

Preliminary hazard evaluation of Androgen receptor-mediated endocrine-disrupting effects of thioxanthone metabolites through structure-based molecular docking

Preliminary Hazard Evaluation of Androgen Receptor-Mediated Endocrine-Disrupting Effects of Thioxanthone Metabolites through Structure-Based Molecular Docking

Tiziana Ginex [‡], Chiara Dall'Asta [§] and Pietro Cozzini [‡*]

[‡]Molecular Modelling Laboratory, Department of Food Science, University of Parma, Parco Area delle Scienze, 17/A – 43124, Parma, Italy.

[§]Department of Food Science, University of Parma, Parco Area delle Scienze, 59/A – 43124, Parma, Italy.

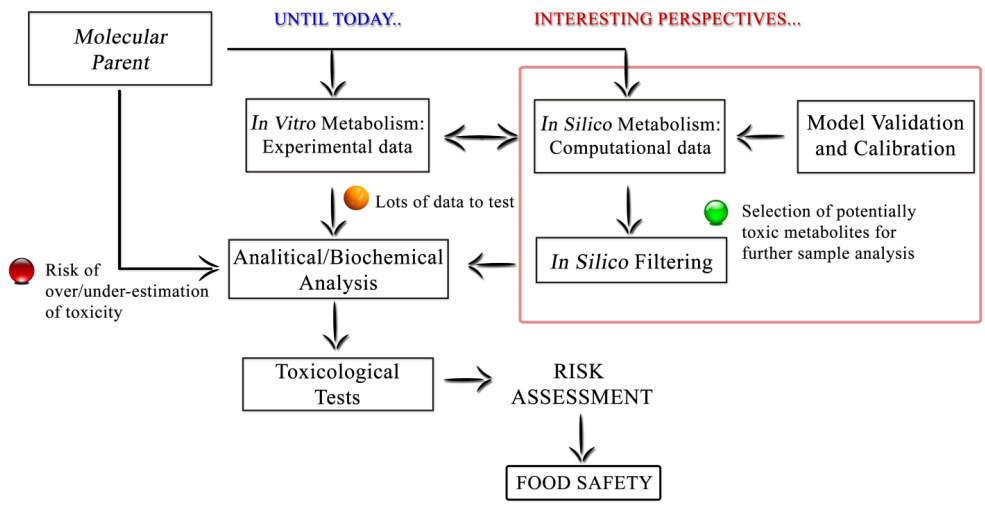
Abstract

Foodstuff could be a vector for naturally occurring and/or unwanted dangerous substances that can act either as they are or after their bioactivation. The scientific community agrees that the metabolic activity of chemicals should be taken into account for proper risk assessment. Unfortunately, the in vitro evaluation of a metabolic panel and analytical/biochemical detection in food-safety assessment are very expensive and challenging because of the abundance of data to analyze. In this context, properly validated computational protocols could be a useful tool for making metabolic and binding/activity predictions. This strategy has been applied to thioxanthone photoinitiators (TX), identified as food contaminants, especially in infant formulas, as reported by the European Food Safety Authority in 2005. Their lipophilicity suggests rapid hepatic metabolism, but the currently available data only concern 2-ITX. We have predicted phase I metabolites for the TX class of compounds and defined their binding affinity for the AR ligand-binding pocket using a local model based on available information about metabolism and AR activity. Some metabolites should undergo further in vitro or/and in vivo toxicological evaluations because they have proved to be suitable as ligands for AR.

Keywords: Androgen receptors, structure-based molecular docking, food toxicology, photoinitiators, food packaging, ITX metabolites.

Introduction:

To function correctly, the human body needs functional nutrients retrieved from a wide variety of foods. However, foodstuff could be a vector for naturally occurring and/or unwanted dangerous substances that can act either as they are or after their bioactivation. The scientific community agrees that the metabolic activity of chemicals should be taken into account for proper risk assessment. Unfortunately, the in vitro evaluation of a metabolic panel and analytical/biochemical detection in food-safety assessment are very expensive and challenging because of the abundance of data to analyze. Moreover, the scientific community needs new technologies for a faster, cheaper, easier, and more accurate evaluation of metabolites and their parents. In this context, in silico technologies could provide a valid support for metabolic and binding/activity predictions.



Scheme 1. Schematic representation of traditional safety assessment and implementation with *in silico* techniques.

If properly validated, such approaches could allow the detection, quantification, and toxicological evaluation to be focused only on those metabolites previously identified as being potentially toxic, in accordance with EFSA policies. The traditional approach and new perspectives offered by in silico techniques are summarized in Scheme 1.

Here, we report a case study of the structure−toxicity relationship for tioxanthone photoinitiators. Thioxanthones are widely used in food packaging. They are added in pigmented formulations as photoinitiator/sensitizers during UV-curing ink-polymerization procedures. Because of their long triplet lifetimes, which can lead to quenching reactions, thioxanthone (TX) compounds promote efficient and rapid ink polymer- ization even at low temperatures, which preserves the lifetime of UV-LED lamps. Despite their good chemical properties, some cases of food contamination have occurred. In 2005, the European Food Safety Authority (EFSA) reported the results of analytical tests on the levels of ITX and EHDAB in a number of food products packaged in cartons printed with UV-cured inks containing isopropyl thioxanthone (ITX) and 2-ethylhexyl-4- dimethylaminobenzoate (EHDAB) as photoinitiators.[1] In infant formulas, ITX has been found at levels ranging from 120 to 305 μg/L, whereas in milk- and soy-based foods not specifically intended for babies, ITX reached levels ranging from 54 to 219 μg/L.

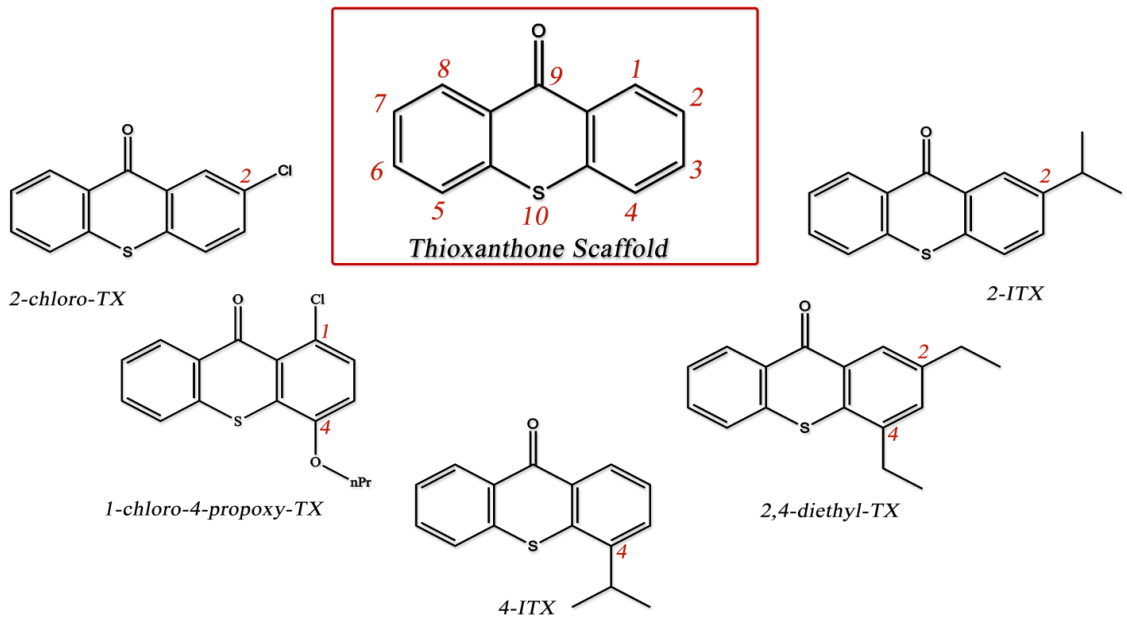


Figure 1. Chemical structures of the five synthetic tioxanthones analyzed by Reitsma and co-workers.[3]

The panel noted that infants exclusively fed with infant formulas packed in cartons printed with UV-cured inks are potentially exposed to more ITX and EHDAB than other population groups because of their high consumption of food per kilogram of body weight. From a toxicological point of view, Peijnenburg, Reisma, and respective co-workers[2,3] (personal communication) previously analyzed AhR-agonistic, -antiandrogenic, and -antiestrogenic effects of some TX compounds using a dioxin receptor chemical-activated luciferase gene-expression assay (DR CALUX) and yeast-based androgen and estrogen bioassays, respectively. The chemical structures of 2- and 4-ITX, 2,4- diethyl-TX, 2-chloro-TX, and 1-chloro-4-propoxy-TX are reported in Figure 1. When tested alone, none of these five compounds have shown agonistic activity. With the exception of 1-chloro-4-propoxy-TX, all tested compounds have been capable of displacing the endogenous ligand,

testosterone, in concentrations of 0.07 and 1 μM, thus showing antiandrogenic behavior with no cytotoxic effects.

The androgen receptor (AR) is a 3-ketosteroid receptor and belongs to a large family of ligand-inducible transcriptional factors identified as nuclear receptors.[4] AR has four main structural and functional domains. Among them, the AR ligand- binding pocket (LBP) is located in the C-terminal domain (CTD). The binding of the endobiotic ligand testosterone or its more potent 5α-reductase-derived metabolite DHT to the LBP induces protein homodimerization and translocation into the nucleus, where the recruitment of cofactor proteins and the activation of transcriptional machinery takes place.[5,6]

From a metabolic point of view, Aprile and co-workers[7] (personal communication) investigated the metabolic stability of 2-ITX in human and liver subcellular fractions in vitro by incubating the chemical with rat liver microsomes (RLM) and human liver microsomes (HLM). After reaction and purification steps, each sample was injected into an LC/MS column for analysis. The structures of eight metabolites were defined by comparing their mass spectra with those of synthetic reference compounds. When 2-ITX was incubated in the presence of RLM, the formation of the 2-(1,2-diol) derivative was observed; the oxidation of the 2-(propen-2-yl) derivative could lead to the generation of an epoxide intermediate, which is completely hydrolyzed by epoxide hydrolases. High TX lipophilicity implies rapid human hepatic metabolism, but this does not always result in detoxification. Currently, despite the androgen-disrupting effects of the five synthetic TX compounds and the metabolic pathway of 2-ITX, no further metabolic/toxicological considerations are available.

Previously reported data on 2-ITX metabolism and AR activity was used as a validation set for making metabolic and binding-affinity predictions, respectively. In this way, a computational approach was performed to predict putative metabolites and to define the potential AR-related endocrine- disrupting effects of the TX class of compounds. Our final results suggest that further in vitro studies could be necessary for a more concrete risk assessment.

Materials and Methods

Protein and Chemicals

For the receptor, the crystal structure of the agonist-bound androgen receptor in complex with its endogenous ligand, testosterone,[8] was used in these studies (PDB: 2AM9, resolution (Å), 1.64; mean B value, 27.38; pH 7.6. 2AM9 was selected because of its good crystallographic properties in terms of resolution, protonation states (pH), and B-factor (index of structural stability), which are crucial to have for a starting material to be suitable for in silico analysis. Moreover, 2AM9 was able to reproduce correctly experimental data on the AR activity of TX compounds. Among the 52 available crystallographic structures of AR, there is no structural model of WT antagonism. We tried to model AR competitive antagonism starting from a crystallographic structure of 2AM9; the assumption is that compounds with an affinity greater than that of the natural ligand for the AR-LBP could be able to displace it in a competitive manner, influencing physiological AR activity. Hydrogen atoms were added in Sybyl v.8.1 (Tripos, Inc., St. Louis, MO) and were energy-minimized using the Powell algorithm with a convergence gradient of 0.5 kcal (mol Å)$^{-1}$

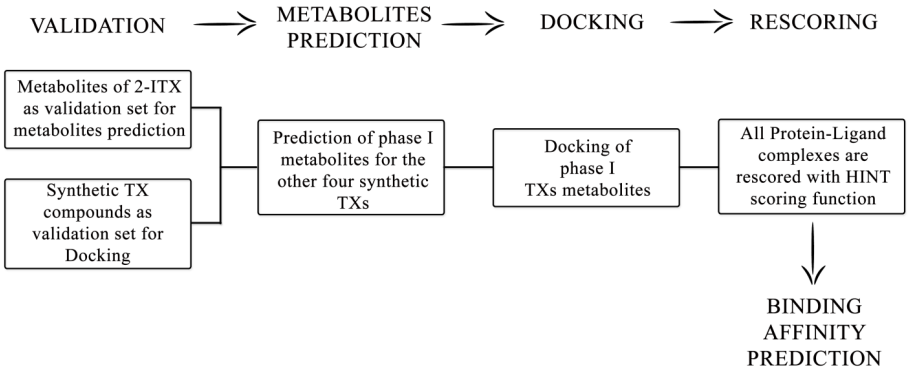
for 1500 iterations. All water molecules, other solvents, and the crystallographic ligand were deleted.

For chemicals, 2-ITX, 4-ITX, chlorinated (2-chloro-TX), alkyl derivative (2,4-diethyl-TX), 1-chloro-4-propoxy-TX, (R)- and (S)-2- (1-hydroxypropan-2-yl)-TX (R-M1; S-M2), 2-(2-hydroxypropan-2-yl)-TX (M3), 2-(prop-1-en-2-yl)-TX (M4), (R)- and (S)-2-(2-methyloxiran-2-yl)-TX (R-M5; S-M6), 2-(2-hydroxypropan-2-yl)-TX- 10-oxide (M7), 2-isopropyl-TX-10-oxide (M8), (R)- and (S)-2-(1,2-dihydroxypropan-2-yl)-TX (R-M9; S-M10), 1-chloro-4-hydroxy-TX (M11), 1-chloro-6-hydroxy-4-propoxy-TX (M12), 1-chloro-4-propoxy-TX 10-oxide (M13), 2,4-diethyl-TX-10-oxide (M14), 2-chloro-TX-10-oxide (M15), 2-chloro-6-hydroxy-TX (M16), 6-hydroxy-2-isopropyl-TX (M17), 4-(2-hydroxypropan-2-yl)-TX (M18), 4-iso-propyl-TX-10-oxide (M19), 2,4-diethyl-6-hydroxy-TX (M20), 6- hydroxy-4-isopropyl-TX (M21), 2-ethyl-4-((R)-1-hydroxyethyl)-TX (M22), and 2-ethyl-4-((S)-1-hydroxyethyl)-TX (M23) were built in Sybyl and then energy-minimized with a convergence gradient of 0.05 kcal (mol Å)$^{-1}$ for 100 iterations. All structures (protein and chemicals) were checked for the correct assignment of atom and bond type and then saved as .mol2 files.

Computational Protocol

The entire protocol is summarized in Scheme 2. Briefly, it consists of validation, metabolites prediction, docking, rescoring, and binding-affinity prediction. A validation step was performed to test the suitability of our approach in metabolite and binding predictions. Metabolites identified by Aprile and co-workers for 2-ITX (M1−M10) were used to test the performance of MetaPrint2D in site of metabolism (SOM) predictions,

whereas the five TX compounds of Reitsma and co-workers were used as a validation set for docking predictions.



Scheme 2. Schematic representation of computational protocol.

Validation of MetaPrint2D Predictive Performance. MetaPrint2D phase I predictions for the thioxanthone scaffold, which are based on a P450's general model, are reported in Figure 2; the C2-isopropyl moiety, S10, and C6 were identified as the sites of the molecular scaffold that are the most likely to undergo phase I metabolism.

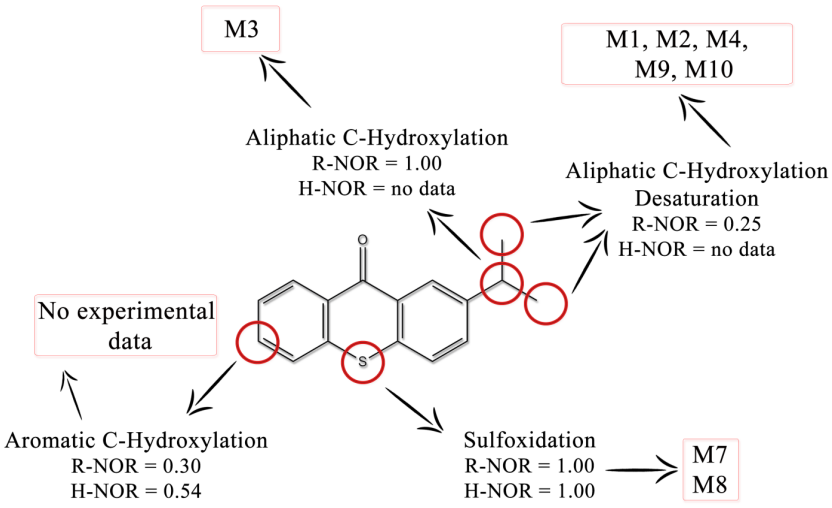


Figure 2. MetaPrint2D predictions for thioxanthone scaffold. As showed, the isopropyl moiety, S_{10} and C_6 were identified as these sites of a molecule that are most likely to undergo phase I metabolism. R-NOR and

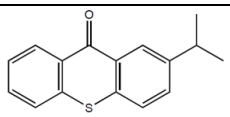
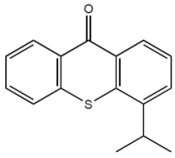
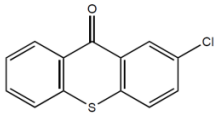
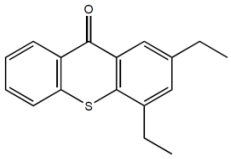
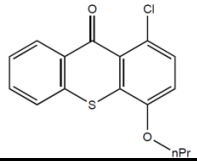
H-NOR that stands for NOR value in respectively Rat and Human model of metabolism were calculated in order to virtually reproduce the same experimental conditions. In case of TX compound, the model is able to correctly predict oxidative metabolism. No experimental data is provided for C_6 aromatic C-hydroxylation.

Aliphatic C-hydroxylation, desaturation, and sulfoxidation as well as aromatic C-hydroxylation are the most likely mechanisms. The NOR value in rat and human models of metabolism (R-NOR and H-NOR, respectively) were calculated to reproduce the same experimental conditions virtually. In the case of the TX scaffold, the model was able to predict oxidative metabolism correctly. Moreover, the in silico predictions agree with the experimental results that suggest a particular metabolic regioselectivity that is mainly focused on the isopropyl moiety and sulfur atom. Aliphatic C-hydroxylation of the 2-isopropyl moiety was correctly predicted (M1−M3, M9, and M10). Desaturation of the isopropyl moiety leads to the formation of olefin M4. M7 and M8 result from sulfoxidation in both human and rat models (NOR = 1.00). Computational predictions also suggest aromatic C-hydroxylation in C6, especially in the human model, but there was no experimental confirmation of this. Epoxide derivatives M5 and M6 are not predicted by MetaPrint2D-React. However, Aprile and co-workers proposed the formation of an instable epoxide from the olefin moiety of M4, but they point out that it was not found in extracts because of its rapid hydrolysis to the corresponding 1,2-diol derivatives (M9 and M10).

Validation of Docking Performance. A comparison between in vitro ligand-mediated transcriptional activity and in silico binding affinity for the validation set is reported in Table 1. Testosterone was considered as a reference compound (HS = 350). Binding affinity is expressed both in absolute (HS) and relative (affinity ratio) terms. Compounds

with HS values greater than testosterone show a higher affinity for AR-LBP and might be able to compete with the endogenous ligand for AR-LBP binding. The affinity ratio expresses the index of the binding affinity relative to testosterone; in other words, it is the ratio between the HS of a tested compound and the HS of testosterone. 2-ITX, 4-ITX, and 2-chloro-TX show an affinity 2 to 3 times greater than testosterone; 2,4-diethyl-TX has an affinity ratio less than 1, and it is considered either to not or to be less capable of significantly competing with endogenous ligand testosterone.

Table 1. Androgen receptor-mediated endocrine disrupting effects of validation set: comparison between *in vitro* ligand-mediated transcriptional activity and *in silico* binding affinity for TX chemicals.

COMPOUND	STRUCTURE	In vitro ACTIVITY IC50 (μM)**	HS	AFFINITY RATIO (HSComp/HSRef*)
2-ITX[3]		1	1005	2.87
4-ITX[3]		1	1014	2.90
2-Chloro-TX[3]		2,5	888	2.54
2,4-Diethyl-TX[3]		9	306	0.87
1-Chloro-4-propoxy-TX[3]		-	No Likely Pose	-

*** Reference compound:** Testosterone (HS = 350 pt.). Affinity Ratio is expressed as $\mathbf{HS}_{Comp}/\mathbf{HS}_{Ref}$. Affinity ratio for Testosterone is 1. ****** Experimental activity has been reported as concentration of tested compound that induces halving in fluorescence signal, in presence of agonist testosterone (Reitsma et al., 2012). Androgen-mediated transcriptional activity has been tested on yeast hormone bioassays that express h-AR and green fluorescent reporter protein (Bovee et al., 2004, 2007). Except for 1-chloro-4-propoxy-TX, all tested compounds of reference set are capable to displace endogenous ligand testosterone in concentrations of 0.07 and 1μM. [3]Already tested *in vitro* (Reitsma et al., 2012).

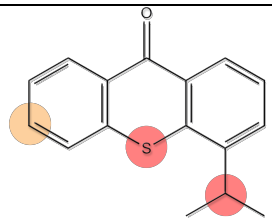
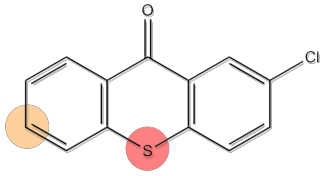
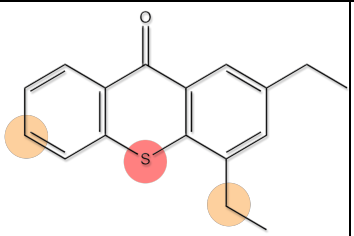
There are no likely poses for 1-chloro-4-propoxy-TX because of its bulky propoxy moiety. Docking results for tested compounds are in good agreement with available experimental data where the clear sigmoidal response of testosterone is inhibited by TX compounds in the following order: 2-ITX, 4-ITX, 2-chloro-TX, and 2,4-diethyl-TX. No androgenic activity was registered for 1-chloro-4-propoxy-TX.

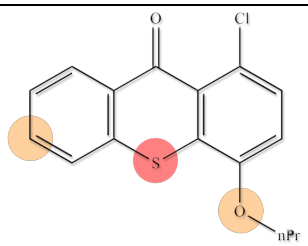
Metabolite Prediction. After validation, the web interface of MetaPrint2D-React was used to predict phase I metabolites for 4- ITX, 2-chloro-TX, 2,4-diethyl-TX, and 1-chloro-4-propoxy-TX; each chemical was uploaded as an isomeric SMILES string generated by ChemBioDraw Ultra 12.0. Reaction types, phase I metabolites, and respective NOR values for the four synthetic TX compounds are summarized in Table 2. Only chemicals with NOR values greater than 0.5, referring to more plausible metabolites, were taken into account for docking. In particular, oxidative metabolism on S10 (sulfoxidation) was predicted for all thioxanthones (NOR = 1.00); in this way, M13, M14, M15, and M19 were obtained. Aromatic C6-hydroxylation (NOR = 0.5) for all tioxanthones leads to the formation of M12, M16, M17, M20, and M21. In vitro metabolic studies of 2-ITX, mainly focused on rat metabolism, have not shown the formation of C6-hydroxyl derivatives; however, C-hydroxylation for polyaromatic scaffolds is frequently reported in the scientific literature, as in the case of PCBs.[11−13] For this reason, we decided to include these

derivatives (M12, M16, M17, M20, and M21) in our study as well. Metabolic predictions for 2,4-diethyl-TX also suggest the formation of M22 and M23 as a result of aliphatic C-hydroxylation at the level of the C4-isopropyl moiety (NOR = 0.5). O-Dealkylation was also predicted for 1-chloro-4-propoxy-TX (M11; NOR = 0.5).

Docking and Rescoring. Each predicted chemical (M11−M23) was docked into AR-LBP. After docking, the GOLD output was energetically revalued and properly ranked with the use of the HINT algorithm and scoring function. Generally, a high HINT score is associated with a good binding result (see the following section).

Table 2. MetaPrint2D predictions: hypothesized mechanism, metabolites and respective NOR values for the four synthetic TX compounds of validation set.

Compound	Structure	Hypothesized Mechanism	Predicted Metabolites	NOR value
4-ITX		**S:** Sulfoxidation **C6:** Aromatic C-hydroxylation **C1':** Aliphatic C-hydroxylation	**M19** **M21** **M18**	1.00 0.50 1.00
2-Cl-TX		**S:** Sulfoxidation **C6:** Aromatic C-hydroxylation	**M15** **M16**	1.00 0.50
2,4-diEt-TX		**S:** Sulfoxidation **C6:** Aromatic C-hydroxylation **C3':** Aliphatic C-hydroxylation	**M14** **M20** **M22/M23**	1.00 0.50 0.50

1-Cl-4PrO -TX		**S:** Solfoxidation **C6:** Aromatic C-hydroxylation **nPrO-C4:** O-Dealkylation	**M13** **M12** **M11**	1.00 0.50 0.50

A suitable positioning of each chemical into the binding site as well as the energetic ranking proposed by the HINT scoring function were taken into account for final considerations.

Software and Procedures

Sybyl v.8.1 was used for protein and chemical preparations. MetaPrint2D-React was used as the metabolite and reaction-type predictor tool. The software predicts the sites of a molecule that are most likely to undergo phase I metabolism based on their similarity to known sites of metabolism and to sites that are known not to be metabolized. The method is based on a database of atom environments found in molecules known to undergo metabolic transformation, such as the data found in the Symyx(R) (previously MDL) metabolite database that has collected over 80 000 metabolic transformations of xenobiotics curated from reports in the scientific literature. This database contains information about phase I additions (C-hydroxylations and other oxidations), eliminations (dealkylations and amide/ester hydrolysis reactions), and phase II additions (conjugations, etc.).

For metabolic predictions, the atom environments of a query molecule are calculated, and the database is searched for similar environments. Then, an occurrence ratio is calculated for each atom in the molecule; this measures how often this or a similar environment has

been found at a reaction center relative to how many times it has been observed in total. To have a standard output, these ratios are then scaled so that the molecule's most likely site of metabolism is given a normalized occurrence ratio (NOR) of 1.[14,15] The derived NOR value is represented with a customizable color code that goes from white (NOR = 0) to red (NOR = 1). This method was capable of correctly identifying the first three most likely sites of metabolism in 87% of the external validation molecule set with a confidence of 90% in an internal validation procedure. Figure 3 provides an example of MetaPrint2D-React metabolite prediction for 2-ITX.

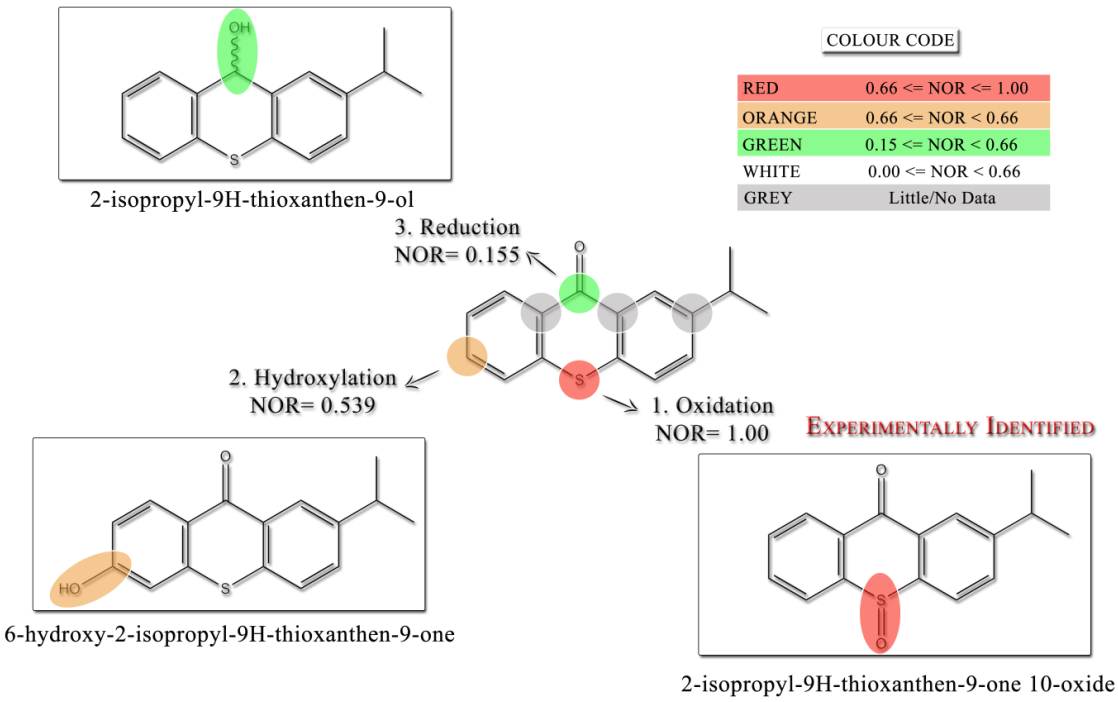


Figure 3. Example of MetaPrint2D-React molecular data mining: the predicted SOM (sites of metabolism) for 2-ITX are reported. As shown, the computationally proposed 2-isopropyl-TX-10-oxide (**M8**) has been also experimentally identified. Only chemicals with NOR values greater than 0.5 that refer to more plausible metabolites was taken into account for *in silico* evaluations.

A semiflexible docking was performed with GOLD suite v. 5.1 (CCDC, Cambridge, UK). AR .mol2 files and all chemicals were loaded into the GOLD docking program. A default radius of 10 Å, from the coordinates of the central atom of crystallographic ligand testosterone, was used to direct the site location into the AR binding cavity. In this way, a suitable coverage of the active site was reached. Ligand flexibility was added before computation. The number of genetic algorithm (GA) runs was set to a maximum of 50 poses for each molecular candidate. For the genetic algorithm run, a maximum number of 100 000 operations were performed on a population of 100 individuals with a selection pressure of 1.1. Operator weights for crossover, mutation, and migration were set to 95, 95, and 10, respectively. The number of islands was set to 5, and the niche size was set to 2. The default GOLDScore fitness function was applied for performing the energetic evaluations.[16,17] The distance for hydrogen bonding was set to 2.5 Å, and the cutoff value for the van der Waals calculation was set to 4.0 Å. To define side-chain flexibility on AR, available crystals were subjected to geometrical superposition (alignment by secondary structure). All of the information collected on the stability/flexibility of some amino acid side chains (Leu701, Asn705, Gln711, Trp741, Met745, Met780, Thr877, and Met895) was used to perform a proper computational approach to reach accurate predictions. To provide a rapid explanation of a more accurate comparison process, we have selected the three most representative crystals. As shown in Figure 4, the superposition of AR cocrystallized with a full agonist (PDB 2AM9), a nonsteroidal modulator[18] (PDB 3RLJ), and a steroidal modulator[19] (PDB 2PNU) revealed important ligand-induced crystallographic flexibility for Leu701, Asn705, Gln711, Trp741, Met745, Met780, Thr877, and Met895.

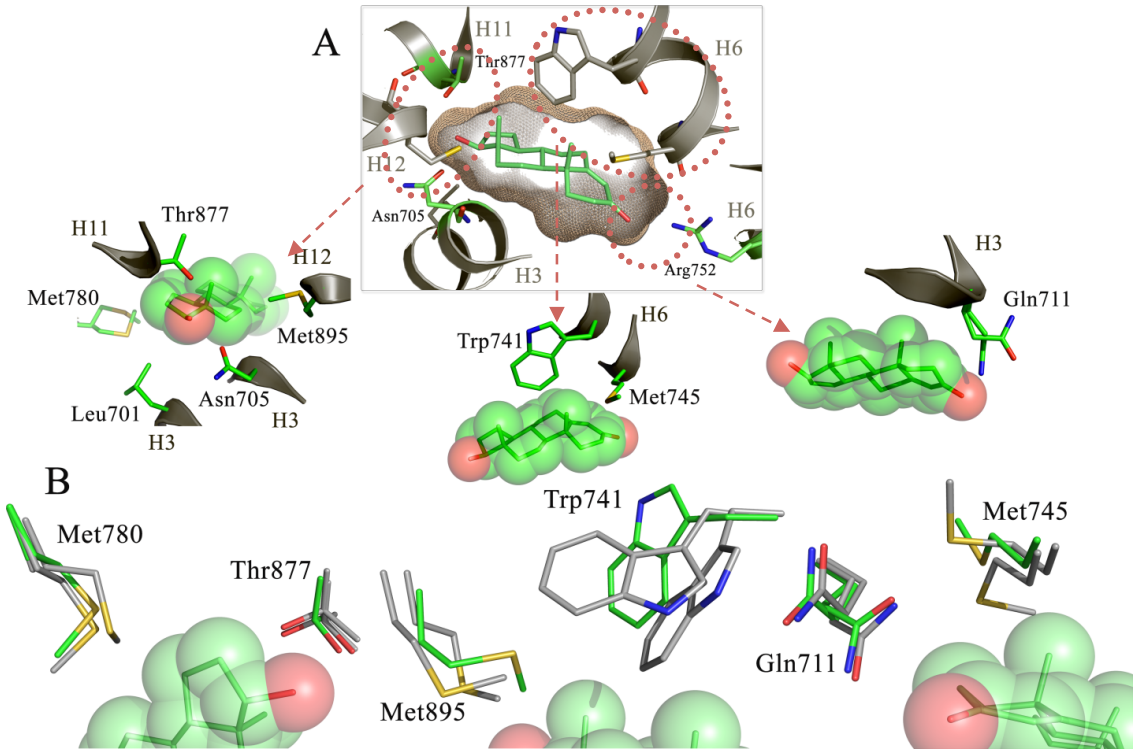


Figure 4. (A) Representation of AR-LBP (PDB ID: **2AM9**): H-bond residues are highlighted in green sticks (Thr877, Asn705 and Arg752). The three red circles identify regions considered flexible during docking simulations: Leu701 (H3; C-terminal), Asn705 (H3; C-terminal), Met780 (H8; C-terminal), Thr877 (H11; C-terminal) and Met895 (H12; N-terminal) lies in front of 17β-OH moiety, Trp741 (H6; N-terminal) and Met745 (H6) are over the steroid scaffold whereas Gln711 (H3) is placed in front of the 3-keto carbonyl of testosterone. Cavity shape is reported in wheat wireframe. (**B**) Superposition of AR-LBP for **2AM9** (full agonist testosterone), **3RLJ** (non-steroidal modulator, SARM S22), **2PNU** (steroidal modulator, EM-5744): side chains orientations for 2AM9 are reported in green whereas side chains orientations for 3RLJ and 2PNU are reported in grey.

Three main red- circled regions can be defined: Leu701 (H3; C-terminal), Asn705 (H3; C-terminal), Met780 (H8; C-terminal), Thr877 (H11; C-terminal), and Met895 (H12; N-terminal) are in front of 17β-OH moiety and Trp741 (H6; N-terminal) and Met745 (H6) are over the steroid scaffold, whereas Gln711 (H3) is placed in front of the 3-keto carbonyl of testosterone. These conformational rearrangements may critically affect ligand accommodation into the binding site, so their side-chain flexibility was taken into account

during docking simulations. GOLD is more binding-site-dependent than other docking software. It has revealed well the exploration of conformational space and the generation of protein−ligand complexes, but it is significantly poorer when binding is predominantly driven by hydrophobic interactions,[20] as in the case of AR-LBP. Because of this, we used GOLD for geometrical definition and HINT for energetic evaluation of protein− ligand complexes generated by GOLD. Despite the enthalpic contribution to the free-energy calculation of protein−ligand interactions already considered by GOLD, the HINT scoring function allows the entropic contribution to the free-energy calculation associated with the displacement of water molecules from hydrophobic binding site by the ligand to be also taken into account. It is based on experimental Log $P_{o/w}$ values and includes the effects of solvation, enthalpic terms such as hydrogen bonding, Coulombic attractions, and hydrophobic attractions. The suitability of the algorithm was tested on a set of 53 protein−ligand complexes with and without water-molecule contributions.[21,22] A linear regression was found between HINT score values and $\Delta G_{binding}$; the correlation between HS and $\Delta G_{binding}$ can be summarized by the following equation

$$\Delta G = - a H_{TOT} - b$$

where the values of a and b depend on the case study. As shown, H_{TOT} is inversely correlated to $\Delta G_{binding}$; negative ΔG values that are associated with a stable and energetically favored protein−ligand complex are identified by high HS values. In an example of virtual screening[23] where the model structures of 26 cyclin-dependent kinase inhibitor complexes were generated by docking (and thus not subject to crystallographic

errors, etc.), HINT results were shown to correlate highly with inhibition. The regression of ΔG versus H_{TOT} for these data reflects the equation reported above with a SE of ±0.3 kcal mol^{-1} and r^2 of 0.94.

Therefore, to allow a proper evaluation of binding free energy in the case of the TX compounds, GOLD output was rescored with the HINT scoring function (hydrophatic interactions; http://www.tripos. com/). High HS (best interaction) are associated with low-free-energy levels for protein−ligand complexes.[24,25] The HINT model describes specific atom−atom interactions between two molecules using the equation

$$H_{TOT} = \Sigma\Sigma\, b_{ij} = \Sigma\Sigma\, (a_i S_i\, a_j S_j R_{ij} T_{ij} + r_{ij})$$

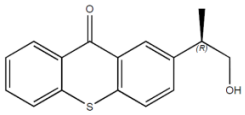
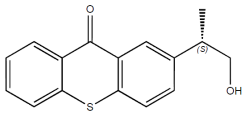
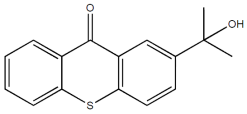
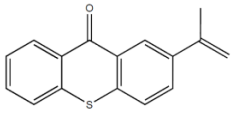
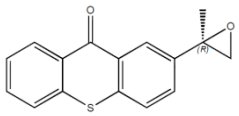
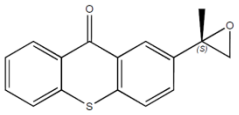
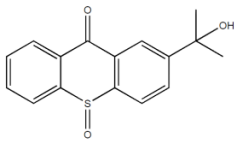
where a is the hydrophobic atom constant (derived from Log $P_{o/w}$), S is the solvent-accessible surface area, T is a function that differentiates polar−polar interactions (acid−acid, acid−base, or base−base), and R and r are functions of the distance between atoms i and j.[26] Each specific intra- and intermolecular atom(i)−atom(j) interaction is defined as the interaction score, b_{ij}, and gives an additive contribution to the global HINT score of protein−ligand association.

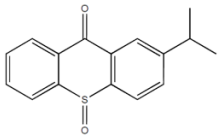
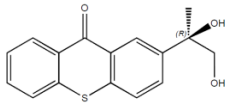
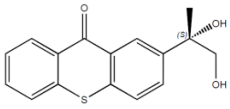
Results and Discussion

As for other 3-keto nuclear receptors, the AR-LBD is configured as a hydrophobic site of accommodation for steroid ligands. As reported in Tables 3 and 4, HS values calculated for metabolites (M1−M10 = experimental but also predicted in silico; M11−M23 = only

predicted in silico) are in close agreement with CoMSIA studies on steroidal and nonsteroidal chemicals as AR ligands by Wang and co-workers.[27]

Table 3. Computational prediction of androgen-mediated endocrine disrupting effects for experimental metabolites.

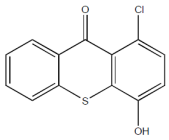
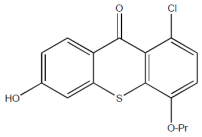
COMPOUND	ABBREVIATION	STRUCTURE	HS	AFFINITY RATIO (HSComp/HSRef*)
(R)-2-(1-hydroxypropan-2-yl)-TX[9]	**M1**		383	1.09
(S)-2-(1-hydroxypropan-2-yl)-TX[9]	**M2**		373	1.07
2-(2-hydroxypropan-2-yl)-TX[9]	**M3**		404	1.15
2-(prop-1-en-2-yl)-TX[9]	**M4**		866	2.47
(R)-2-(2-methyloxiran-2-yl)-TX[9]	**M5**		669	1.91
(S)-2-(2-methyloxiran-2-yl)-TX[9]	**M6**		579	1.65
2-(2-hydroxypropan-2-yl)-TX-10-oxide[9]	**M7**		155	0.44

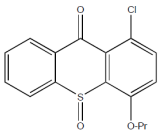
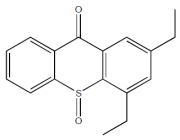
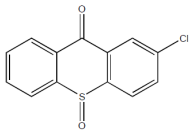
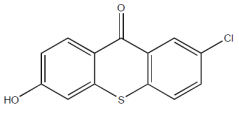
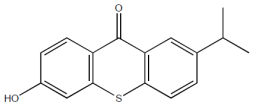
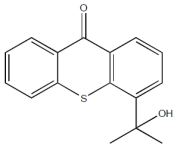
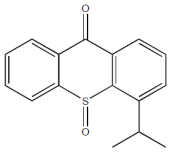
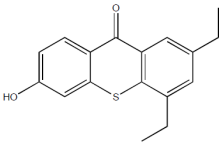
2-isopropyl-TX-10-oxide[9]	**M8**		176	0.50
(R)-2-(1,2-dihydroxypropan-2-yl)-TX[9]	**M9**		Negative	-
(S)-2-(1,2-dihydroxypropan-2-yl)-TX[9]	**M10**		126	0.36

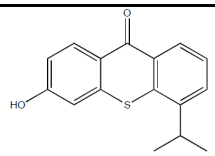
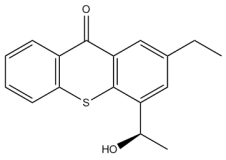
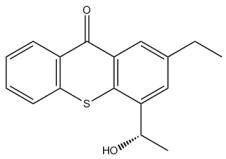
***Reference compound:** Testosterone (HS = 350 pt.). Affinity Ratio is expressed as $\mathbf{HS}_{Comp}/\mathbf{HS}_{Ref}$. Affinity ratio for Testosterone is 1. [9] From Aprile et al., 2010.

If the key phenomena for a good binding affinity of steroid derivatives are electrostatic (especially H-bonds), then the driving force of the binding process for nonsteroidal ligands is mainly hydrophobic. In Figure 5, comparisons between the binding orientations for endogenous AR ligand testosterone (A), one synthetic (2-ITX; B1−2), and one metabolic M3; C1−2) candidate are reported.

Table 4. Computational prediction of androgen-mediated endocrine disrupting effects for MetaPrint2D-React metabolites.

COMPOUND	ABBREVIATION	STRUCTURE	HS	AFFINITY RATIO (HSComp/HSRef*)
1-chloro-4-hydroxy-TX	**M11**		493	1.41
1-chloro-6-hydroxy-4-propoxy-TX	**M12**		453	1.29

1-chloro-4-propoxy-TX 10-oxide	**M13**		No Likely Pose	-
2,4-diethyl-TX 10-oxide	**M14**		341	0.97
2-chloro-TX-10-oxide	**M15**		88	0.25
2-chloro-6-hydroxy-TX	**M16**		1045	2.99
6-hydroxy-2-isopropyl-TX	**M17**		1138	3.25
4-(2-hydroxypropan-2-yl)-TX	**M18**		No Likely Pose	-
4-isopropyl-TX 10-oxide	**M19**		No Likely Pose	-
2,4-diethyl-6-hydroxy-TX	**M20**		879	2.71

6-hydroxy-4-isopropyl-TX	**M21**		1008	2.88
2-ethyl-4-((R)-1-hydroxyethyl)-TX	**M22**		125	0.36
2-ethyl-4-((S)-1-hydroxyethyl)-TX	**M23**		Negative	-

*** Reference compound:** Testosterone (HS = 350 pt.). Affinity Ratio is expressed as HSComp/HSRef. Affinity ratio for Testosterone is 1.

The AR−testosterone binary complex is stabilized by hydrophobic interactions with the steroid scaffold. 17β-OH in ring D acts as a H-bond donor for Asn705(H3) and as a H-bond acceptor for Thr877(H11), whereas the carbonyl oxygen in position 3 of ring A acts as a H-bond acceptor for the guanidine group of Arg752. Because of their planar structure, the binding poses for the TX compounds and testosterone are very closely related.

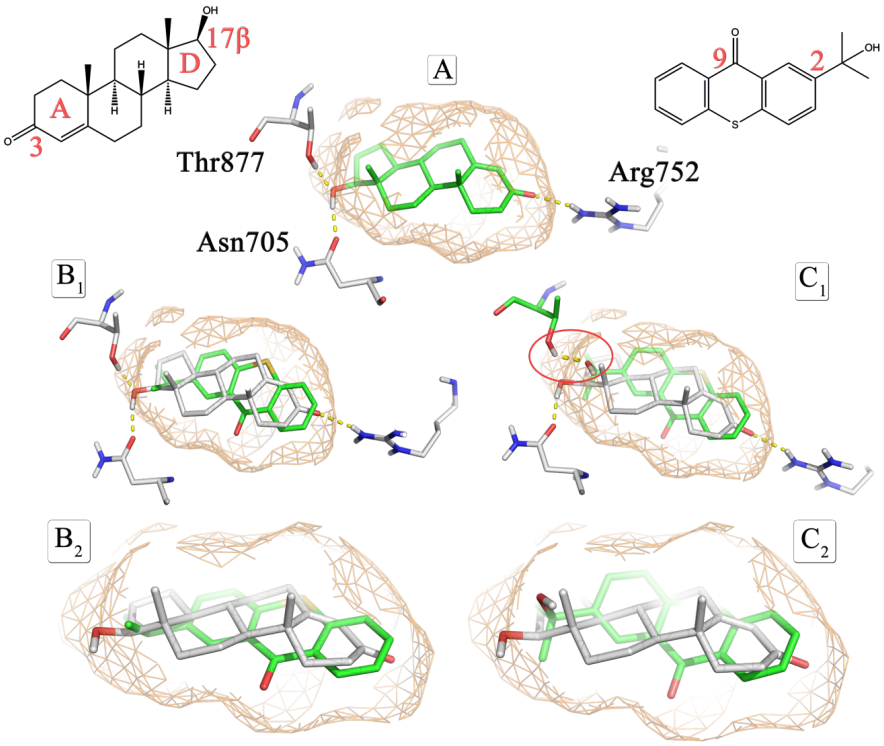


Figure 5. Comparison of binding models for natural ligand testosterone (T; **A**), 2-ITX (**B1-2**), and **M3** (**C1-2**). (A) The 17β-OH edge of Testosterone acts as H-bond donor for Asn705(H3) and H-bond acceptor for Thr877(H11) whereas the C3=O edge acts as H-bond acceptor for guanidine group of Arg752. Both 2-ITX and **M3** mimic binding orientation of testosterone, giving a suitable occupation of the active site. The cavity shape is reported in wheat "wireframe".

A suitable occupation of the AR cavity is also observed by docking simulations. M7, M8, M9, and M10 are expected to be less capable of competing with testosterone for a stable accommodation into the binding site. In the case of the two 10-oxide derivatives, it could be due to steric and electronic clashes generated by the sulfoxide placed in front of Val746(H6), Met787(H8), and Leu873(H11). It is interesting to note that the 1,2-diol moiety in M9 and M10 and the ternary alcohol in M1 and M2, which differ only by an accessory hydroxyl functional group placed in C1′, gives a significant lowering of the affinity.

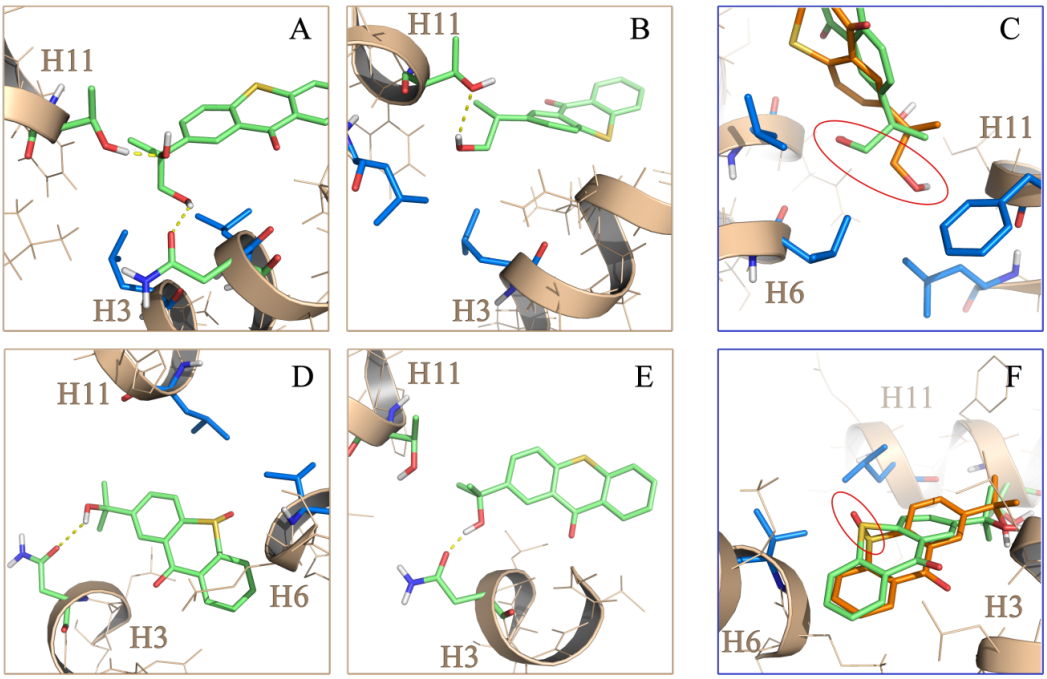


Figure 6. Graphic representation of predicted binding poses for **M9** (**A**), **M1** (**B**), **M7** (**D**), **M3** (**E**), analysis of negative effects generated by α-oriented hydroxyl groups (**C**) for 1,2-diol (in orange) and 2-hydroxyl derivatives and by the 10-sulfoxide moiety (in orange; **F**). Side chains that give ligand-receptor negative interactions are highlighted in blue sticks.

In Figure 6, the predicted binding models for M9 (A), M1 (B), M7 (D), and M3 (E) are reported. The C1′−OH and C2′−OH of M9 act as a H-bond acceptor and H-bond donor for Thr877(H11) and Asn705(H3), respectively (Figure 6A). M1 gives only one H-bond with Thr877 (Figure 6B). A comparison between the 1,2-diol (in orange) and primary alcohol (in lime) (Figure 6C) reveals how the presence of polar groups lying in the N-terminal hydrophobic region of H11 can negatively affect binding affinity: the red-circled hydroxyl groups are in front of the Phe876, Leu873, Leu880 (H11), and Val746 (H6) hydrophobic

side chains. Indeed, as demonstrated by Fang and co-workers[28] and Hong and co-workers,[29] the 17β-OH configuration (over the steroid plane) of testosterone is important for binding affinity. α-Epitestosterone (OH lies under the steroid plane) shows a relative binding affinity (RBA) three times lower than the 17β diastereoisomer. The tertiary alcohol of M7 (in orange) and M3 (in lime) act as a H-bond donor for Asn705 (H3) (Figure 6D,E). A comparison between these two compounds (Figure 6F) reveals how the red-circled sulfoxide is poorly tolerated by the binding pocket: it lies in front of the hydrophobic side chains of Leu873 (H11) and Val746 (H6), giving negative hydrophobic/polar interactions. No likely pose was reached for M13, M18, and M19 because of electronic and steric clashes generated by the sulfoxide and hydroxyl groups in C4. HINT scores under the cutoff and affinity ratios less than 1 are predicted for M15, M22, M14, and M23. A C6-hydroxyl group is revealed as being good at promoting ligand−receptor interaction. With the exception of M12, affinity ratios greater than 2 were obtained for M17, M21, M20, and M16. The bulky C4-propoxy group significantly affects HS values for M12, giving an affinity ratio of 1.29. The C6-hydroxyl group H-bonds with Gln711(H3), Met745(H6), and Arg752(H6) (data not shown).

Briefly, the in silico results can be summarized as follows: a 10-oxide group is poorly tolerated, whereas a 2-(1,2- dihydroxypropan-2-yl) substituent gives steric and electronic clashes that reduce binding affinity. Except for M7, M8, M9, M10, M14, M15, and M22 and those that were negative controls, the affinity ratios for all of the tested compounds go from 1 to 3.5, showing that these compounds could act as competitors of testosterone for the steroid-binding site.

Conclusions

The case study of TX compounds and their metabolites is an application of a more general computational approach to some phases of toxicokinetic/toxicodynamic studies. We are conscious of the small size of the validation set, but no other data are currently available. Additionally, these data have allowed us to generate a good local model to study the thioxanthone scaffold and were able to predict correctly the experimental results. A dimensionally restricted and chemically homogeneous validation set can allow for the building of accurate, simple, and plastic local models for efficient preliminary filtering. The approach is not intended as a stand-alone protocol but rather as a valid and innovative support for use in the preliminary stages of safety assessment.

Because no toxicological data on TX metabolites is available, our results suggest that hazard evaluation could be necessary to avoid a potential underestimation of the toxicological risk for this class of chemicals.

Author Information

Corrsponding author: *E-mail pietrocozzini@unipr.it; Phone: +39 0521 905669; Fax: +39 0521 905556.

Acknowledgment

We thank Dr. M. Reitsma of the RIKILT-Institute of Food Safety, Wageningen UR, Wageningen and Dr. S. Aprile of the Dipartimento di Scienze Chimiche, Alimentari, Farmaceutiche e Farmacologiche and Drug and Food Biotechnology Center, Università degli Studi del Piemonte Orientale A. Avogadro for the scientific material that they kindly

provided. We also thank Prof. Glen E. Kellogg, Department of Medicinal Chemistry, Virginia Commonwealth University, for the use of the HINT software.

References

(1) EFSA (European Food Safety Authority) (2005) Opinion of the scientific panel of food additives, flavourings, processing aids and materials in contact with food on a request from the commission related to 2-isopropylthioxanthone (ITX) and 2-ethylhexyl-4-dimethylaminobenzoate (EHDAB) in food contact materials. EFSA J. 293, 1−15.

(2) Peijnenburg, A., Riethof-Poortman, J., Baykus, H., Portier, L., Bovee, T., and Hoogenboom, R. (2010) AhR-agonistic, anti- androgenic, and anti-estrogenic potencies of 2-isopropylthioxanthone (ITX) as determined by in vitro bioassays and gene expression profiling. Toxicol. In Vitro 24, 1619−1628.

(3) Reitsma, M., Bovee, T. F., Peijnenburg, A. A., Hendriksen, P. J., Hoogenboom, R. L., and Rijk, J. C. (2013) Endocrine-disrupting effects of thioxanthone photoinitiators. Toxicol. Sci. 132, 64−74.

(4) Germain, P., Staels, B., Dacquet, C., Spedding, M., and Laudet, V. (2006) Overview of nomenclature of nuclear receptors. Pharmacol. Rev. 58, 685−704.

(5) Claessens, F., Verrijdt, G., Schoenmakers, E., Haelens, A., Peeters, B., Verhoeven, G., and Rombauts, W. (2001) Selective DNA binding by the androgen receptor as a mechanism for hormone-specific gene regulation. J. Steroid Biochem. Mol. Biol. 76, 23−30.

(6) Li, J., and Al-Azzawi, F. (2009) Mechanism of androgen receptor action. Maturitas 63, 142−148.

(7) Aprile, S., Del Grosso, E., and Grosa, G. (2011) In vitro metabolism study of 2-isopropyl-9H-thioxanthen-9-one (2-ITX) in rat and human: evidence for the formation of an epoxide metabolite. Xenobiotica 41, 212−225.

(8)PereiradeJeśus-Tran,K.,Côte,́P.L.,Cantin,L.,Blanchet,J., Labrie, F., and Breton, R. (2006) Comparison of crystal structures of human androgen receptor ligand-binding domain complexed with various agonists reveals molecular determinants responsible for binding affinity. Protein Sci. 15, 987−999.

(9) Bovee, T. F. H., Helsdingen, J. R., Rietjens, I. M. C. M., Keijer, J., and Hoogenboom, L. A. P. (2004) Rapid yeast estrogen bioassays stably expressing human estrogen receptors α and β, and green fluorescent protein: A comparison of different compounds with both receptor types. J. Steroid Biochem. Mol. Biol. 91, 99−109.

(10) Bovee, T. F. H., Helsdingen, J. R., Hamers, A. R. M., Duursen, M. B. M., Nielen, M. W. F., and Hoogenboom, L. A. P. (2007) A new highly specific and robust yeast androgen biosassay for the detection of agonist and antagonists. Anal. Bioanal. Chem. 389, 1549−1558.

(11) McLean, M. R., Bauer, U., Amaro, A. R., and Robertson, L. W. (1996) Identification of catechol and hydroquinone metabolites of 4- monochlorobiphenyl. Chem. Res. Toxicol. 9, 158−164.

(12) Amaro, A. R., Oakley, G. G., Bauer, U., Spielmann, H. P., and Robertson, L. W. (1996) Metabolic activation of PCBs to quinones: Reactivity toward nitrogen and sulfur nucleophiles and influence of superoxide dismutase. Chem. Res. Toxicol. 9, 623−629.

(13) Lehmler, H. J., and Robertson, L. W. (2001) Synthesis of hydroxylated PCB metabolites with the Suzuki-coupling. Chemosphere 45, 1119−1127.

(14) Boyer, S., Arnby, C. H., Carlsson, L., Smith, J., Stein, V., and Glen, R. C. (2007) Reaction site mapping of xenobiotic biotransformations. J. Chem. Inf. Model. 47, 583−590.

(15) Carlsson, L., Spjuth, O., Adams, S., Glen, R. C., and Boyer, S. (2010) Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. BMC Bioinf. 11, 362.

(16) Jones, G., Willett, P., and Glen, R. C. (1995) A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. J. Comput.-Aided Mol. Des. 9, 532−549.

(17) Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997) Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol. 267, 727−748.

(18) Duke, C. B., Jones, A., Bohl, C. E., Dalton, J. T., and Miller, D. D. (2011) Unexpected binding orientation of bulky-B-ring anti- androgens and implications for future drug targets. J. Med. Chem. 54, 3973−3976.

(19)Cantin,L.,Faucher,F.,Couture,J.F.,deJeśus-Tran,K.P., Legrand,P.,Ciobanu,L.C.,Frećhette,Y.,Labrecque,R.,Singh,S.M., Labrie, F., and Breton, R. (2007) Structural characterization of the human androgen receptor ligand-binding domain complexed with EM5744, a rationally designed steroidal ligand bearing a bulky chain directed toward helix 12. J. Biol. Chem. 282, 30910−30919.

(20) Perola, E., Walters, W. P., and Charifson, P. S. (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 56, 235−249.

(21) Cozzini, P., Fornabaio, M., Marabotti, A., Abraham, D. J., Kellogg, G. E., and Mozzarelli, A. (2002) Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water. J. Med. Chem. 45, 2469−2483.

(22) Fornabaio, M., Cozzini, P., Mozzarelli, A., Abraham, D. J., and Kellogg, G. E. (2003) Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 2. Computational titration and pH effects in molecular models of neuraminidase-inhibitor complexes. J. Med. Chem. 46, 4487−4500.

(23) Gussio, R., Zaharevitz, D. W., McGrath, C. F., Pattabiraman, N., Kellogg, G. E., Schultz, C., Link, A., Kunick, C., Leost, M., Meijer, L., and Sausville, E. A. (2000) Structure-based design modifications of the paullone molecular scaffold for cyclin-dependent kinase inhibition. Anti-Cancer Drug Des. 15, 53−66.

(24) Fornabaio, M., Spyrakis, F., Mozzarelli, A., Cozzini, P., Abraham, D. J., and Kellogg, G. E. (2004) 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.

(25) Marabotti, A., Spyrakis, F., Facchiano, A., Cozzini, P., Alberti, S., Kellogg, G. E., and Mozzarelli, A. (2008) Energy-based prediction of amino acid-nucleotide base recognition. J. Comput. Chem. 29, 1955− 1969.

(26) Kellogg, G. E., and Abraham, D. J. (2000) Hydrophobicity: Is LogP(o/w) more than the sum of its parts? Eur. J. Med. Chem. 35, 651−661.

(27) Wang, X., Li, X., Shi, W., Wei, S., Giesy, J. P., Yu, H., and Wang, Y. (2012) Docking and CoMSIA studies on steroids and non-steroidal chemicals as androgen receptor ligands. Ecotoxicol. Environ. Saf. 89, 143−149.

(28) Fang, H., Tong, W., Branham, W. S., Moland, C. L., Dial, S. L., Hong, H., Xie, Q., Perkins, R., Owens, W., and Sheehan, D. M. (2003) Study of 202 natural, synthetic, and environmental chemicals for binding to the androgen receptor. Chem. Res. Toxicol. 16, 1338−1358. (29) Hong, H., Fang, H., Xie, Q., Perkins, R., Sheehan, D. M., Tong, W. Comparative molecular field analysis (CoMFA) model using a large diverse set of natural, synthetic and environmental chemicals for binding to the androgen receptor. SAR QSAR Environ. Res. 14, 373-388.

Part 2. New QM-derived molecular descriptors for 3D-QSAR

New QM-derived molecular descriptors for 3D-QSAR

Background

Within a micro/meso-scale context, QSAR approaches[268] are very useful in disclosing sub-molecular implications of structure-activity/property relationships for small molecules,[269,270] sensory compounds,[271] bioactive chemicals,[272] antimicrobial[273] peptides, food proteins[274] and other important food-related phenomena. These techniques allow to deeply investigate the physico-chemical determinants (charge, hydrophobicity, volume…) of specific properties, such as stability, texture, biological effects, and aroma of food products.

As already mentioned in the Introduction, a correct evaluation of the physico-chemical properties of food components is important to reach a proper knowledge at the macroscopic level. In this context, QSAR models could be valuable for explaining specific nutritional and toxicological properties of foods. Furthermore, these methods could be useful for developing and predicting peptide and protein structures with specific technological/food related properties such as

- ✓ Functional foods with peptides as active compounds
- ✓ Antimicrobial peptides as alternatives to other preservatives
- ✓ Proteins and protein hydrolysates with specific physical properties.
- ✓ Monitor the formation of undesirable peptide structures.
- ✓ Regulate enzyme activity in fermented foods by introduction of designed inhibitory peptides.

This work aimed to develop and validate new physico-chemical descriptors to be applied in a 3D-QSAR context. Unlike the MM-derived physico-chemical descriptors (partial charges, vW radius,…) used in classical 3D-QSAR CoMFA and CoMSIA methods, our atomic log*P*-based parameters are derived from QM calculations. *This constitutes two important advantages: the intrinsic accuracy of QM calculations and the physico-chemical features of logP-derived parameters allow us to reach a higher level of accuracy in the description of molecular properties, and a more easy interpretation of molecular determinants in host-guest interactions.*

The HyPhar (log*P*-based) descriptors proposed in this study have been derived from the (re-parameterized) MST version of the IEF-PCM continuum solvation model at the B3LYP level of theory. For each molecule immersed in a defined solvent, the free energy of solvation is defined as summation of the electrostatic and non-electrostatic terms. When determined for water and octanol, they can be combined to derive electrostatic ($\log P_{ele}$) and non-electrostatic ($\log P_{n\text{-}ele} = \log P_{cav} + \log P_{vW}$) contributions to the molecular hydrophobicity. The 3D-distribution of these components can then be utilized to understand the differences in biological activity.

	D2/D4 inhibitors		**Antifungal 2-aryl-4-Chromanones**	**GSK3 inhibitors**	**Cruzain inhibitors**	**Thermolysin inhibitors**
Training						
N	32	32	27	56	26	61
Mean	7.28	8.21	5.69	6.73	6.72	4.96
St. Dev.	1.25	0.66	0.40	0.41	0.71	2.09
Test						
N	6	6	7	18	6	15
Mean	6.93	8.16	6.51	6.61	6.65	5.22
St. Dev.	1.10	0.86	0.42	0.44	0.75	1.91

Table 1. Number of compounds, mean activity and standard deviation for the five training and test sets in this study.

To this end, five molecular sets (see **Table 1**) have been selected from the literature to validate the suitability of these new descriptors. Statistical performances for the new HyPhar models (H1-H4; more details in the text) generated by using the log*P*-derived

descriptors have been compared with the results obtained from classical CoMFA and CoMSIA methods.

All the work can be schematized as follows (see **Figure 1**):

- ✓ Molecular set selection and preparation
- ✓ Original (CoMFA/CoMSIA) model reproduction
- ✓ Calculation of logP-derived parameters
- ✓ HyPhar models generation (HyPhar H1-H4)
- ✓ Statistical analysis: comparison of statistic performances for HyPhar and CoMFA/CoMSIA models
- ✓ Graphical analysis: comparison of the isocontour maps for HyPhar and CoMFA models

As a general trend, this study has shown that the statistical results obtained from HyPhar H1-4 models are in good agreement with the reference data.

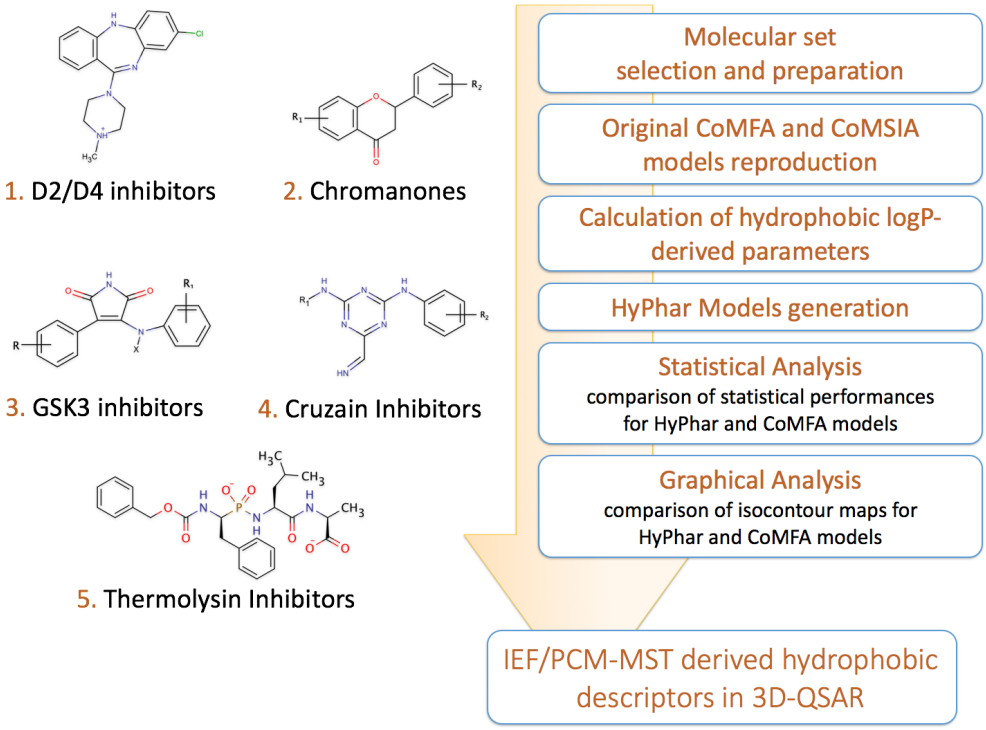


Figure 1. Schematic representation of the validation protocol applied in this study.

HyPhar methodology has demonstrated to compare well with the standard CoMFA/CoMSIA models for the five molecular sets analysed. The advantages in application of these new hydrophobic descriptors can be summarized as follows:

- a larger consistency in description of molecular properties
- a more easy interpretation of the logP-based parameters with a direct linkage with experimentally measurable data
- a proper accounting of both enthalpic and entropic contributions to solvation free energy in ligand-receptor interactions
- the predictive accuracy reported for classical CoMFA/CoMSIA methods is achieved with a lower number of statistical principal components and, in case of CoMSIA, with a lower number of property fields.

Furthermore, the graphical representation of the pharmacophoric fields closely agrees with the key features derived from CoMFA models, suggesting a good agreement between our log*P*-derived electrostatic and non-electrostatic descriptors and those of the reference 3D-QSAR methods. Overall, the results confirm the suitability of these new parameters in modeling the activity for specific systems in a 3D-QSAR context.

Within the context of structure-activity relationships analyses, another important challenge consists in addressing the selectivity issue. This term refers to the possibility to correctly model the selectivity of a given set of molecules across different but closely (structurally) related receptors or, more generally, biological systems. The selectivity issue is a very important and well-known problem in medicinal chemistry, but it also affect the selective interactions of food components against different biomolecular targets. Thus, there are some examples reported in the literature, in which the variation in chemical structure leads to changes in the activity profile for instance, in specific antimicrobial activity for some food additives (essential oil compounds),[275] or in Topoisomerase I/II inhibitory activity for polyphenols.[276]

In this regard, the capability of our new hydrophobic descriptors to address the selectivity issue has been tested for a set 88 inhibitors of Thrombin, Trypsin and Factor Xa. The analysis has been focused on three selectivity models:

- Thrombin/Trypsin
- Thrombin/Factor Xa
- Trypsin/Factor Xa

In all cases, the selectivity was modelled as the difference in inhibitory activities of the 88 molecules toward two systems at a time. As before, four HyPhar models were calculated for each selectivity model and compared with those of the original work by Klebe.[277] The molecular features associated to the selectivity toward each system have been analysed and discussed in detail in the paper.

Paper 1

Development and validation of hydrophobic molecular fields derived from the Quantum Mechanical IEF/PCM-MST Solvation Models in 3D-QSAR

Development and Validation of Hydrophobic Molecular Fields Derived from the Quantum Mechanical IEF/PCM-MST Solvation Models in 3D-QSAR

Tiziana Ginex,[a] Jordi Muñoz-Muriedas,[b] Enric Herrero,[c] Enric Gibert,[c] Pietro Cozzini,[a]* and F. Javier Luque[d]*

[a] Dipartimento di Scienze degli Alimenti, University of Parma, Parco Area delle Scienze 59/A, 43121 Parma, Italy

[b] GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, United Kingdom

[c] Pharmacelera, Jordi Girona 1-3, Campus Nord Universitat Politècnica de Catalunya, Edifici K2M, 08034 Barcelona, Spain

[d] Department of Chemical Physics and Institut de Biomedicina (IBUB), Faculty of Pharmacy, University of Barcelona, Av. Prat de la Riba 171, 08921 Santa Coloma de Gramenet, Spain

* E-mail: pietro.cozzini@unipr.it (PC) or fjluque@ub.edu (FJL)

Abstract

Since the development of structure-activity relationships about 50 years ago, 3D-QSAR methods belong to the most refined ligand-based *in silico* techniques for prediction of biological data using physicochemical molecular fields. In this scenario, this study reports the development and validation of quantum mechanical (QM)-based hydrophobic descriptors derived from the parametrized MST continuum solvation model to be used in 3D-QSAR studies within the framework of the Hydrophobic Pharmacophore (HyPhar) method. To this end, five sets of compounds reported in the literature (dopamine D2/D4 antagonists, antifungal 2-aryl-4-chromanones, and inhibitors of GSK-3, cruzain and thermolysin) have been revisited. The results derived from the QM/MST-based hydrophobic descriptors have been compared with previous CoMFA and CoMSIA studies, and examined in light of the available X-ray crystallographic structures of the targets. The analysis reveals that the combination of electrostatic and non-electrostatic components of the octanol/water partition coefficient yields pharmacophoric models fully comparable with the predictive potential of standard 3D-QSAR techniques. Moreover, the graphical representation of the hydrophobic maps provides a direct linkage with the pattern of interactions found in crystallographic structures. Overall, the introduction of the QM/MST-based descriptors, which could be easily adapted to other continuum solvation formalisms, paves the way to novel computational strategies for disclosing structure-activity relationships in drug design.

Keywords: hydrophobic molecular field · continuum solvation model · quantum mechanics · 3D-QSAR

Introduction

Three-dimensional quantitative structure-activity relationships (3D-QSAR) methods have been fundamental for disclosing practical computer-aided guidance in drug discovery.[1-4] One of the relevant goals to be attained in a 3D-QSAR model is the derivation of a statistically significant and predictive model that allows to rank the potency of new compounds, thus providing a quantitative correlation between molecular structure and biological activity. Moreover, a 3D-QSAR model should assist the provision of a graphical representation of the topological distribution of physicochemical properties relevant for the affinity and selectivity of compounds, thus enabling an easy interpretation of the hidden relationships between differences in chemical structure and biological activity.

A cornerstone contribution to 3D-QSAR was Comparative Molecular Field Analysis (CoMFA),[5] where Lennard-Jones (L-J) and Coulomb potentials are mapped onto regularly spaced grid points surrounding the mutually aligned molecules. Although CoMFA has proven to be extremely useful, it suffers from a number of practical limitations. L-J and Coulomb potentials are believed to encode enthalpic contributions to the binding affinity, while entropic components appear to be inadequately covered. Further, the two fields exhibit different distance dependence, and since the slope of the L-J potential is very steep close to the van der Waals surface, there are drastic changes in grid points located around the molecular surface (i.e., the most interesting region for non-covalent interactions). Hence, apparently minor deviations in the alignment or the conformation of compounds may trigger chemically unphysical changes in the fields around the molecules. In addition, both L-J and Coulomb fields show singularities at atomic positions, and arbitrary cutoff values are then introduced to avoid extremely large values. Finally, the isocontour maps often exhibit discontinuties due to the

steepness of the potentials close to the molecular surface and the cutoff settings, thus making their interpretation more difficult.

Several strategies have been proposed to alleviate these limitations. Comparative Molecular Similarity Index Analysis (CoMSIA) relies on similarity indices determined using a gaussian-type functional form, which leads to smoother distance dependence, avoid singularities and provides improved contour maps.[6] The effect of smoothing functions to correct exceedingly large values of molecular fields has been explored in other studies.[7] Alternatively, the L-J potential has been replaced by other steric descriptors, such as an atom-based indicator field,[8] which indicates the presence of specific atoms in predefined volume elements, or the volume of the intersection between van der Waals envelopes of ligand and a probe atom.[9] Further, extraction of a minimum set of representative *pseudo*-atoms have been proposed in Comparative Molecular Active Site Analysis (CoMASA) to eliminate lattice-based problems,[10] while Compass limits the computation of the physical properties only near the molecular surface.[11]

With regard to the molecular fields, other physicochemical descriptors have also been examined in 3D-QSAR studies. CoMSIA was originally formulated considering electrostatic, steric and hydrophobic fields, which were later complemented by hydrogen-bond donor and acceptor properties.[6,12,13] The role of lipophilicity of the fragments present in a molecule on intermolecular interactions has also been accounted for in the Molecular Lipophilicity Potential (MLP)[14] and Hydropathic Interactions (HINT)[15,16] methods, which use empirically contributions to the molecular hydrophobicity. On the other hand, VolSurf extracts simpler, lattice-independent descriptors related to the hydrophilic and hydrophobic regions of molecules by using the molecular fields determined for water and hydrophobic probes from GRID

calculations,[17-19] In this line, other descriptors have been proposed in order to eliminate the need for molecular alignment, such as Comparative Molecular Moment Analysis (CoMMA),[20] which utilizes moments of molecular mass and charge distribution, while other strategies have exploited molecular spectra as descriptors invariant to molecular orientation, such as EVA[21,22] and Comparative Spectra Analysis (CoSA),[23] which exploits normal coordinate eigenvalues as well as NMR and IR spectra, respectively.

In order to improve the accuracy of the molecular fields determined from classical expressions and expand the number of descriptors, quantum mechanical (QM) methods are being utilized in 3D-QSAR studies.[24-26] Thus, QM-based semiempirical methods have been used for the computation of probe intermolecular energies at grid point around the molecules, so that the fields would account for a better description of the electronic structure of molecules and the interactions with suitable probes.[27-29] QM-based fields also allow the description of unusual hydrogen- and halogen-bond interactions.[30] Finally, QM-based methods allow the extension of the properties to be considered for the calculations of 3D molecular fields, including typical reactivity indexes such as frontier molecular orbitals or paremeters derived from Fukui function.[31-33]

In this context, the aim of this study is to explore the suitability of QM-derived descriptors of hydrophobicity for 3D-QSAR studies. This work is motivated by three major reasons. First, (de)solvation is a major contribution to the binding affinity of drugs, as this term accounts from the differential interactions formed by the ligand upon transfer from aqueous solution to the lipophilic environment of the binding pocket. Second, hydrophobicity is generally estimated from the octanol/water partition coefficient, a property that encodes both enthalpic and entropic contributions. Third, the advances in refined parametrization of QM-Self Consistent Reaction Field (QM-SCRF)

codes provide a basis to evaluate the solvation free energy in a variety of solvents, and hence the partition coefficient of solutes. Therefore, it can be expected that QM-SCRF methods can be useful for providing novel descriptors relevant for understanding the molecular determinants of ligand binding. The computational procedure presented here, denoted Hydrophobic Pharmacophore (HyPhar), relies on a rigorous partitioning scheme[34] of the solvation/transfer free energy into fragment contributions within the framework of the parametrized MST versions[35,36] of the QM solvation continuum IEF-PCM method.[37] The partitioning scheme divides the solvation in water and octanol into contributions assigned to the surface elements that define the solute/solvent interface, which can be subsequently integrated to derive atomic or group contributions to the octanol/water partition coefficient. The suitability of the QM/MST hydrophobic fields is calibrated through comparison with the results obtained by CoMFA and CoMSIA methods for five molecular systems. The results support the potential use of the QM/MST-based hydrophobic contributions to provide novel molecular fields in ligand-based drug design.

Materials and Methods

Atomic contributions to the hydrophobicity

The hydrophobicity of a molecule is typically determined from the partitioning between octanol and water,[38] but fractional descriptors are needed to evaluate the hydrophobic complementarity between a given molecule and its biological target. In this regard, the 3D hydrophobicity pattern of a molecule can be defined from the atomic contributions to the $\log P_{o/w}$ taking advantage of the decomposition scheme formulated for the solvation free energy within the MST version of the IEF-PCM solvation model.[37]

In the MST method the solvation free energy (ΔG_{sol}) is calculated by adding three contributions (Eq. 1).[39] The first one is the cavitation term (ΔG_{cav}), which is the work required for creating a cavity shaped to accommodate the solute in the solvent. The second component is the van der Waals term (ΔG_{vW}), which accounts for dispersion-repulsion between solute and solvent. Finally, the third component is the electrostatic term (ΔG_{ele}), which measures the work needed to build up the solute charge distribution in the solvent.

$$\Delta G_{sol} = \Delta G_{ele} + \Delta G_{cav} + \Delta G_{vW} \quad (1)$$

Following the IEF-PCM formalism, the reaction field generated by the solvent consists of a set of imaginary charges located on the solute cavity. This strategy allows to partition ΔG_{ele} into atomic contributions following a perturbative description of the solute-solvent electrostatic interaction (Eq. 2).[40]

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

where N is the total number of atoms, M is the total number of reaction field charges (q_j^{sol}) spread over the cavity surface, and Ψ^o is the wave function of the solute in the gas phase. Let us note that the electrostatic contribution of atom i, $\Delta G_{ele,i}$, is calculated taking into account the interaction of the molecular electrostatic potential with the set of reaction field charges pertaining to the solvent-exposed surface of atom i ($M \in i$).

Both ΔG_{cav} and ΔG_{vW} are evaluated using expressions that depend linearly on the solvent-exposed surface of each atom in the molecule, and hence can be directly decomposed into atomic contributions (Eqs. 3 and 4).

$$\Delta G_{cav} = \sum_{i=1}^{N} \Delta G_{cav,i} = \sum_{i=1}^{N} \frac{S_i}{S_T} \Delta G_{P,i} \quad (3)$$

where $\Delta G_{P,i}$ is the cavitation free energy of atom *i* determined using Pierotti's formalism,[41] whose contribution is weighted by the contribution of the solvent-exposed surface (S_i) of atom *i* to the total surface (S_T).[42]

$$\Delta G_{vW} = \sum_{i=1}^{N} \Delta G_{vW,i} = \sum_{i=1}^{N} \xi_i S_i \quad (4)$$

where ξ_i denotes the atomic surface tensión of atom *i*, which is determined by fitting the experimental free energy of solvation.[35,36]

The total solvation free energy is obtained as a sum of the three contributions for all atoms in the molecule (Eq. 5).

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

Application of Eq. 5 to the solvation of a given compound in water and octanol leads to octanol/water transfer free energy ($\Delta G_{tr}^{o/w}$), which can be expressed as the sum of atomic contributions (Eq. 6).

$$\Delta G_{tr}^{o/w} = \sum_{i=1}^{N} \Delta G_{tr,i}^{o/w} = \sum_{i=1}^{N} (\Delta G_{ele,i}^{o/w} + \Delta G_{cav,i}^{o/w} + \Delta G_{vW,i}^{o/w}) \quad (6)$$

Finally, the hydrophobicity of a molecule, expressed as logarithm of the octanol/water partition coefficient ($\log P$), can be related to the sum of the atomic contributions (Eqs. 7 and 8).

$$\log P = \sum_{i=1}^{N} \log P_i = \sum_{i=1}^{N} (\log P_{ele,i} + \log P_{cav,i} + \log P_{vW,i}) \quad (7)$$

$$\log P_X = \sum_{i=1}^{N} \log P_{X,i} = \sum_{i=1}^{N} -\frac{\Delta G_{X,i}^{o/w}}{2.303RT} \qquad (X : ele,\ cav,\ vW) \quad (8)$$

Molecular systems

Five molecular systems were selected from the literature to calibrate the suitability of the QM/MST-based atomic hydrophobicity descriptors in HyPhar. These systems, which were chosen to encompass a diverse range of targets as well as biological activities, had been previously examined by means of CoMFA/CoMSIA analysis, thus providing suitable test systems to validate the 3D-QSAR model built up from HyPhar. The molecular systems comprise a set of dopamine D2/D4 receptor antagonists,[43] antifungal chromanones,[44] glycogen synthase kinase-3 (GSK-3) inhibitors,[45] cruzain inhibitors,[46] and finally thermolysin inhibitors.[1,47] A graphical representation of the main prototypical and pharmacophoric elements for the five molecular sets is given in Figure 1, while a detailed description of the chemical structures and experimental activities is provided in Supporting Information (Figures S1-S6 and Tables S1-S5). Molecular properties (molecular weight, hydrogen-bond donor/acceptor, clogP, and

number of rotatable bonds) for the five systems analysed in this study have been calculated using DataWarrior program.[49]

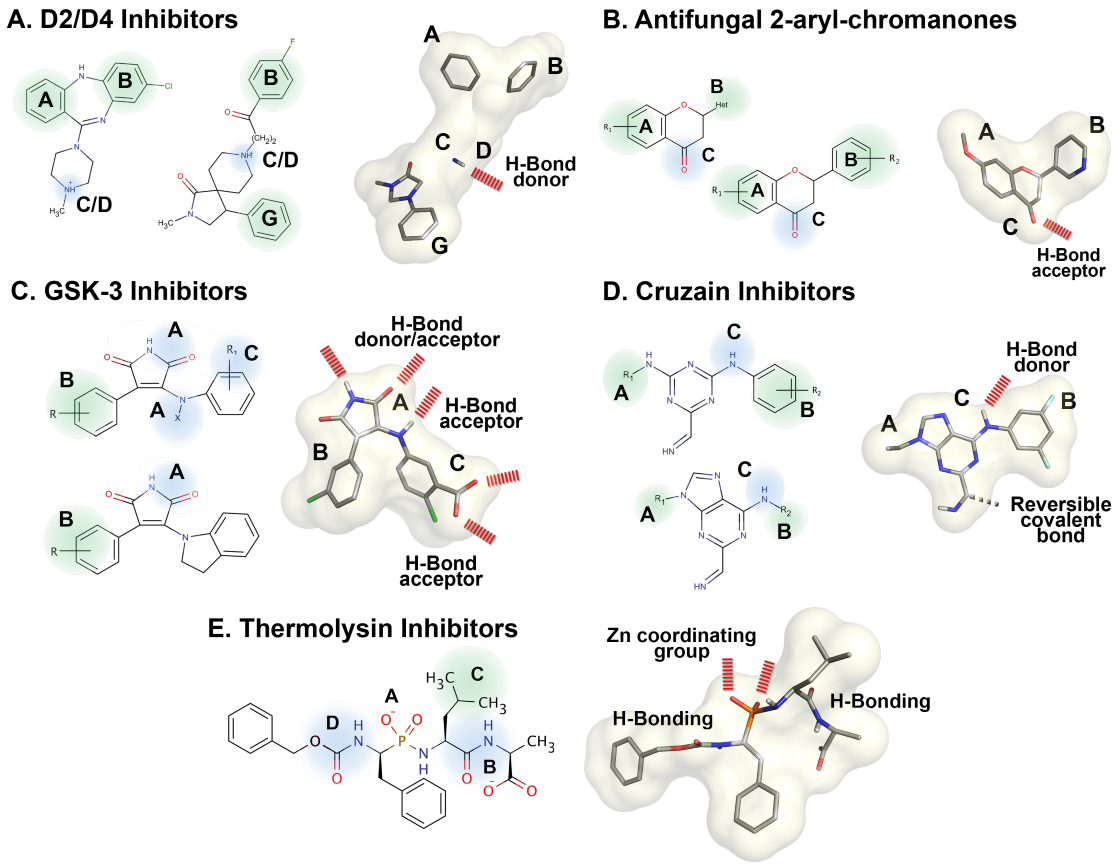


Figure 1. Prototypical structures for the molecular systems analyzed in this study, and schematic representation of topological pharmacophore elements identified from previous 3D-QSAR studies: regions relevant for electrostatic/steric contributions are shown in green, and key hydrogen-bond donor/acceptor are indicated in blue (see refs. 43-47 for details). The white isocontour denotes the shape of the whole set of aligned molecules for each test system.

A graphical representation of the distribution of these properties for each set of compounds is provided in Supporting Information (Figure S10-14 in Supporting Information), showing that the compounds cover a wide range of values for these properties. For the comparative purposes of this work, we have adopted the same distinction between training and test subsets of compounds adopted in the reference

works taken from the literature. Finally, since a correct molecular alignment is of utmost importance for the derivation of 3D-QSAR models, we have also reproduced as close as possible the alignment procedure reported in the original works.

a) Dopamine D2/D4 antagonists. A set of 41 diverse antagonists were analyzed in the original work.[43] The training set included 32 structurally diverse compounds, covering a range of inhibitory potencies (measured as pK_i) ranging from 5.66 to 10.30 for the D2 receptor, and from 7.28 to 10.30 for the D4 one (Figures S1 and S2). Compounds were selected according to pharmacophoric criteria reported previously for the D4 receptor antagonists,[48] which involves the presence of three aromatic rings and an ammonium nitrogen at specific positions, and a site-point in the N-H direction (Figure 1). The molecular geometries of the antagonists and their relative alignment were directly retrieved from the supporting information in ref. 43.

b) Antifungal 2-aryl-4-chromanones. Out of the 34 compounds included in the original work (Figure S3),[44] 27 were used for the training set, which covered a range of inhibitory potencies (measured as pI_{50}) from 4.98 to 6.29. Among the three different alignment procedures examined by the authors, procedure II (see Figure 2 in ref. 44), which relies on the superposition of the 4-chromanone moiety, led to the overall best predictive model. For our purposes, this procedure is also convenient in order to minimize the uncertainty in the molecular alignment, which was carried out with Sybyl 8.1[51] as reported in ref. 44.

c) GSK-3a inhibitors. The original set consisted of 74 4-arylmaleimides, which share the 3-phenyl-2,5-dihydro-1H-pyrrole-2,5-dione scaffold (Figure S4).[45] The training set comprised 56 compounds, with experimental potencies (measured as pIC_{50}) ranging from 5.58 to 7.70. Amongst the two template-based alignments reported by the authors, we have followed the aligment to the ligand (compound 52) extracted from the X-ray

structure of the GSK-3β complex (PDB ID 1Q4L), which led to the best QSAR models. Accordingly, alignment was reproduced using the automated alignment procedure implemented in Sybyl, as reported in ref. 45.

d) Cruzain inhibitors. The original study examined a set of 32 purine carbonitriles identified as reversible covalent inhibitors of cruzain (Figure S5).[46,52] The training set (26 compounds) covered a range of pIC_{50} values ranging from 5.80 to 8.00. Among the two alignments examined in the reference work, we have adopted the procedure shown in Figure 1B in ref. 46, whereby molecules were aligned to compound 23 (as found in the X-ray structure with cruzain; PDB ID 3I06) superposing the six-membered ring of the purine ring under the constraint that the phenyl groups of all molecules were oriented to the same side of the purine scaffold.

e) Thermolysin inhibitors. 76 peptidomimetics have been analysed for their inhibitory effect (measured as *pKi*) on thermolysin. Original partial charges (Gasteiger-Marsili method) and molecular alignment (SEAL alignment) used by Klebe and co-workers[47] have been used in conjunction with the molecular geometries retrieved from the comparative work of Sutherland.[1] Calculations have been performed following the partitioning of the compounds between training (61 molecules) and test (15 molecules) sets adopted in ref. 47. The training set covered a wide range of inhibitory activities (from 0.52 to 10.17), with a mean value of 4.97 and a standard deviation of 2.07.

With the exception of D2/D4 antagonists and thermolysin inhibitors, whose molecular structure was taken respectively from the original data in ref. 43 and 1 (see above), the geometry of all compounds was optimized at the B3LYP/6-31G(d) level of theory. The final geometries were examined to rule out the occurrence of drastic conformational changes that might affect molecular alignment, and subsequently used in single-point continuum solvation calculations in water and octanol performed with the parametrized

B3LYP/6-31G(d) version of the IEF-PCM MST model.[36] Though this procedure differs from the level of theory utilized for geometry optimization in refs. 43-46, it was deemed necessary in order to keep the consistency with the parametrization of the MST model and to account for the conformational-dependence of the atomic hydrophobic contributions. All calculations were performed with Gaussian 09.[53]

Model generation and statistical analysis

In order to extract the 3D-QSAR HyPhar model, all analyses were perfomed using the in-house PharmQSAR software.[54] For each test system, the compounds were aligned using the guidelines indicated in the reference works (see above), and then were placed in a lattice of 0.5 Å grid spacing with boundaries chosen to allow a minimun of 3 Å extension from the atoms in the molecules (for the sake of comparison with the results reported in ref. 47 a grid spacing of 1 Å was used for the thermolysin inhibitors). The total number of points in the grids were 9200, 880, 1210, 8075 and 18720 for the series of D2/D4 antagonists, 2-aryl-4-chromanones, GSK-3a inhibitors, cruzain and thermolysin inhibitors, respectively. The atomic hydrophobicities were projected into the grid using the similarity index function (A^q) implemented in CoMSIA (Eq. 9).

$$A^q(j) = \sum_{i=1}^{N} w_{probe} w_i e^{-\alpha r_{iq}^2} \qquad (9)$$

where i stands for the summation index for all atoms of molecule j, w_i is the actual value of the atomic hydrophobicity of atom i, w_{probe} is the hydrophobicity of the probe atom, which is taken as +1, α is the attenuation factor, and r_{iq} is the distance between the probe atom at grid point q and atom i of the test molecule.

Test calculations performed to choose the optimal value of the attenuation factor supported the suitability of the value recommended in CoMSIA, which was then set to 0.3. Each projected field was stored as an *M*x*Ng* matrix, where *M* is the number of molecules and *Ng* denotes the number of grid points. PharmQSAR uses partial least squares (PLS) to extract the hidden relationships between biological data and the hydrophobic field using an algorithm based on NIPALS.[55] To this end, pretreatment of the *M*x*Ng* matrix involved normalization of the field values, which were corrected by the mean value and normalized to unit variance, and removal of columns with a standard deviation lower than a certain threshold (typically 0.1). No scaling was applied to the molecular fields obtained for each descriptor in the derivation of the 3D-QSAR models.

The quality of the best hydrophobic model was assessed from standard statistical descriptors in 3D-QSAR studies,[56] which allowed direct comparison with the results obtained from the original CoMFA/CoMSIA results. The optimum number of components was selected on the basis of the leave-one-out cross-validation, the lowest standard deviation error in prediction of the actual experimental values corrected by the number of degrees of freedom of the model, S_{PRESS}, and the ability of the model to predict the biological activity of the external test set of compounds utilized in the reference studies.

Finally, let us note that CoMFA calculations were performed for each molecular system to test the internal accuracy of the PharmQSAR software, and particularly to check the consistency of the molecular alignment performed for this work. To this end, attention was paid to the origin of the atomic charges and L-J potentials used for each set of compounds in order to guarantee the maximum reproducibility of the results.

Results and Discussion

The ability of the partitioning scheme to decompose the octanol/water $\log P$ into electrostatic ($\log P_{ele}$), cavitation ($\log P_{cav}$) and van der Waals ($\log P_{vW}$) contributions (Eqs. 7 and 8) allowed us to explore the relationships with the biological activity using different combinations of descriptors. In particular, since both $\log P_{cav}$ and $\log P_{vW}$ molecular fields ultimately depend on the solute-exposed surface of atoms (Eqs. 3 and 4), they are expected to reflect the size and shape of the molecule, and accordingly they are expected to encode the information related to the steric field. In fact, these two fields are highly correlated, indicating that they provide similar information about the steric features of compounds (see Figure S12). On the other hand, since $\log P_{ele}$ is determined from the solvent reaction field induced from the solute charge distribution (Eq. 2), it should encode information related to the electrostatic features of molecules. Furthermore, the nonredundancy of the information encoded by this field is noted in the lack of an apparent correlation with either $\log P_{cav}$ or $\log P_{vW}$ (Figure S12). Accordingly, for our purposes we have devised several combinations of descriptors: i) $\log P_{ele}$ and the cube of the atom radii, which was taken as a simple measure of the atomic size,[43] ii) $\log P_{ele}$ and $\log P_{cav}$, iii) $\log P_{ele}$ and $\log P_{vW}$, and, finally iv) $\log P_{ele}$ and the total non-electrostatic component, $\log P_{n-ele}$, which was obtained as the addition of $\log P_{cav}$ and $\log P_{vW}$. In the following the models obtained from the preceding combinations of descriptors will be denoted H1-H4, respectively. In addition, CoMFA models obtained by combining Coulomb and L-J potentials as reported in refs. 43-47 were determined for the sake of comparison.

Overall analysis of hydrophobic pharmacophores

The statistical parameters obtained for the different HyPhar models (H1-H4) for the five molecular systems are shown in Table 1, which also displays the corresponding values taken from CoMFA and CoMSIA models in refs. 43-47.

Let us first note that CoMFA calculations performed in this study reproduce the results reported for the five systems (see Table 1 and Figure S13 in Supporting Information).

Table 1. Summary of statistical parameters obtained for the HyPhar models derived from the QM MST-based hydrophobic descriptors (models H1-H4) and CoMFA/CoMSIA results taken from refs. 43-47 for the compounds in the training sets.

	CoMFA [a]	**CoMSIA**	**H-1**	**H-2**	**H-3**	**H-4**
D2 (32 compounds) [b]						
r^2	0.88 (0.85)	0.92	0.93	0.94	0.94	0.95
S	0.45 (0.40)	0.37	0.27	0.26	0.24	0.23
q^2	0.68 (0.67)	0.75	0.75	0.77	0.79	0.81
S_{press}	0.74 (0.74)	0.63	0.64	0.61	0.59	0.56
Nc [c]	3 (3)	3	3	3	3	3
Fields (%) [d]						
	E 34 (41)	E 12	$\log P_{ele}$ 62	$\log P_{ele}$ 69	$\log P_{ele}$ 64	$\log P_{ele}$ 63
	S 68 (59)	S 9	R^3 38	$\log P_{cav}$ 31	$\log P_{vW}$ 36	$\log P_{n\text{-}ele}$ 37
		H 24				
		D 14				
		A 41				
D4 (32 compounds)						
r^2	0.74 (0.72)	0.77	0.71	0.73	0.73	0.73
S	0.35 (0.55)	0.33	0.56	0.54	0.54	0.53
q^2	0.49 (0.41)	0.51	0.45	0.45	0.44	0.44
S_{press}	0.49 (0.53)	0.48	0.45	0.50	0.50	0.50
Nc	2 (3)	2	2	2	2	2
Fields (%)						
	E 29 (40)	E 14	$\log P_{ele}$ 47	$\log P_{ele}$ 56	$\log P_{ele}$ 52	$\log P_{ele}$ 54
	S 71 (60)	S 11	R^3 53	$\log P_{cav}$ 44	$\log P_{vW}$ 48	$\log P_{n\text{-}ele}$ 46
		H 19				
		D 19				
		A 38				
2-Aryl-4-chromanones (27 compounds)						
r^2	0.95 (0.94)	0.92	0.82	0.78	0.77	0.82
S	0.10 (0.25)	0.13	0.44	0.49	0.51	0.43
q^2	0.72 (0.77)	0.74	0.49	0.52	0.51	0.56
S_{press}	0.24 (0.21)	0.24	0.30	0.29	0.29	0.28
Nc	6 (5)	6	3	3	3	3
Field (%)						
	E 35 (28)	E 58	$\log P_{ele}$ 51	$\log P_{ele}$ 41	$\log P_{ele}$ 34	$\log P_{ele}$ 40
	S 65 (71)	S 42	R^3 49	$\log P_{cav}$ 59	$\log P_{vW}$ 66	$\log P_{n\text{-}ele}$ 60
GSK-3a (56 compounds)						
r^2	0.94 (0.91)	0.91	0.90	0.91	0.90	0.89

S	0.10 (0.31)	0.13	0.32	0.31	0.32	0.33
q^2	0.84 (0.81)	0.78	0.79	0.80	0.78	0.77
S_{press}	- (0.18)	-	0.19	0.19	0.19	0.20
Nc	5 (3)	5	3	3	3	3
Field (%)						
	E 49 (45)	E 53	$logP_{ele}$ 59	$logP_{ele}$ 55	$logP_{ele}$ 52	$logP_{ele}$ 52
	S 51 (56)	S 9	R^3 41	$logP_{cav}$ 45	$logP_{vW}$ 48	$logP_{n\text{-}ele}$ 48
		H 38				
Cruzain (26 compounds)						
r^2	0.98 (0.93)	0.90	0.86	0.81	0.83	0.83
S	0.11 (0.26)	0.25	0.38	0.45	0.43	0.43
q^2	0.74 (0.67)	0.56	0.56	0.54	0.56	0.55
S_{press}	0.41 (0.43)	0.52	0.49	0.49	0.48	0.49
Nc	5 (4)	5	3	2	2	2
Field (%)						
	E 18 (39)	E 23	$logP_{ele}$ 23	$logP_{ele}$ 53	$logP_{ele}$ 46	$logP_{ele}$ 47
	S 82 (61)	S 34	R^3 77	$logP_{cav}$ 47	$logP_{vW}$ 54	$logP_{n\text{-}ele}$ 53
		D 35				
		A 8				
Thermolysin (61 compounds)[h]						
r^2	0.94[e] (0.92)	0.78[f]/0.89[g]	0.85	0.85	0.89	0.89
S	0.55[e] (0.28)	1.00[f]/0.71[g]	0.40	0.39	0.34	0.34
q^2	0.51[e] (0.48)	0.51[f]/0.58[g]	0.55	0.57	0.62	0.64
S_{press}	1.53[e] (1.56)	1.51[f]/1.41[g]	1.44	1.40	1.34	1.31
Nc	7[e] (5)	5[f]/7[g]	4	4	5	5
Field (%)						
	E 64[e] (85)	E 54[f]/30[g]	$logP_{ele}$ 37	$logP_{ele}$ 52	$logP_{ele}$ 47	$logP_{ele}$ 52
	S 36[e] (15)	S 46[f]/26[g]	R^3 63	$logP_{cav}$ 48	$logP_{vW}$ 53	$logP_{n\text{-}ele}$ 48
		H -[f]/44[g]				

[a] Data taken from the original works. Data in parenthesis refer to CoMFA analysis performed in this study. [b] Number of compounds in the training set. [c] Number of principal components. [d] Fraction of the field (in percentage). E: electrostatic; S: steric; H: hydrophobic; D: H- bond donor; A: H-bond acceptor; R^3: cube of the atomic radius. [e] Statistical results obtained for models with a grid spacing of 1 Å. CoMSIA results obtained by using two (electrostatic + steric) and three (electrostatic + steric + hydrophobic) grids are reported (data taken from ref. 47).

This agreement indicates that potential uncertainties arising from the bioactive conformation of compounds, molecular alignment, assignment of partial charges and L-J parameters, and evaluation of molecular fields carried out here with regard to the reference works should be marginal. In fact, the most significant difference is found for cruzain inhibitors (q^2 values of 0.74 in ref. 46 *vs*. 0.67 in this study), but this may be attributed to the different origin of the atomic charges.

Thus, unlike the original work,[46] where each molecule was optimized taking into account the interaction with residues in the enzyme binding site, we derived partial

charges from in vacuo calculations. Overall, the agreement found between the CoMFA results in refs. 43-47 and the present CoMFA model lends support to the in-house computational procedure utilized in this study. Data in Table 1 point out that the HyPhar methodology leads to models of statistical quality comparable to both CoMFA and CoMSIA. As a general rule, the results are closer to CoMSIA than to CoMFA, a trend that likely reflects the use of the similarity index functions utilized in the calculation of the QM/MST-based hydrophobic molecular fields.

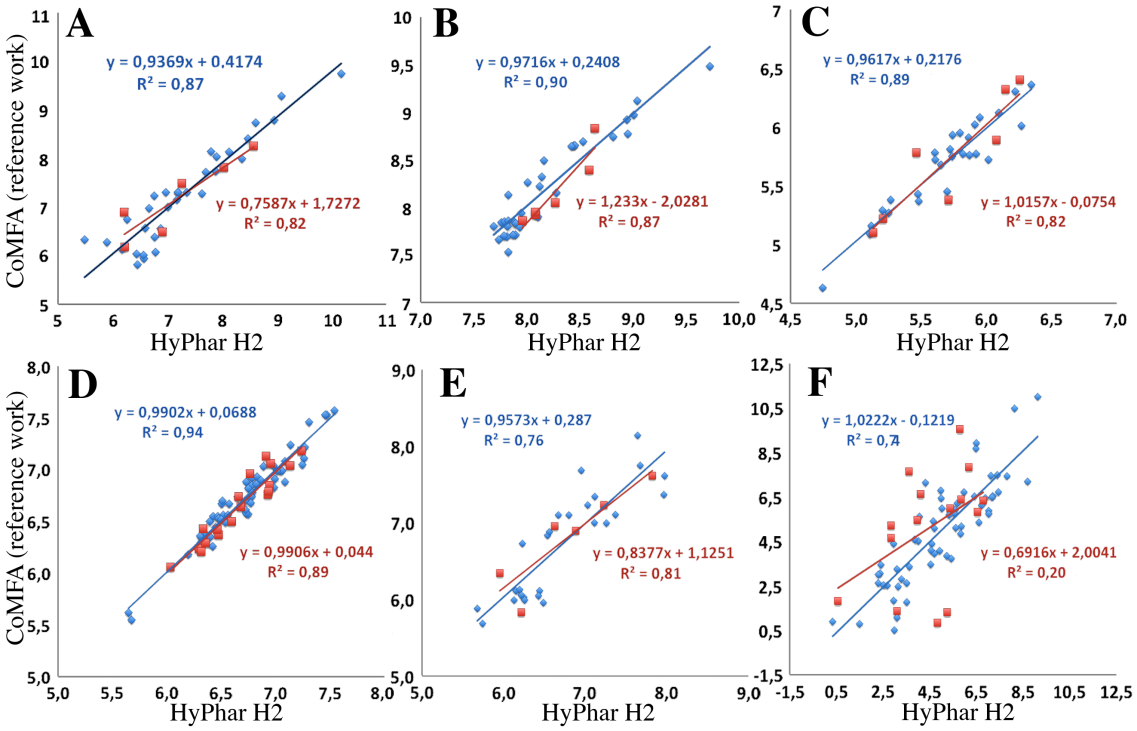


Figure 2. Comparison of the results obtained from standard CoMFA models (data taken from refs. 43-46) and the MST-based hydrophobic H2 model for (A) D2, (B) D4, (C) 2-aryl-4-chromanones, (D) GSK-3α, (E) cruzain and (F) thermolysin systems. Compounds of the training/test set are shown in blue/red, respectively.

Furthermore, very similar results are obtained for models H2-H4, which can be ascribed to the dependence of the corresponding atomic contributions on the van der Waals aurface of atoms in the molecule (Eqs. 3 and 4). Finally, the resemblance observed for

models H1 and H2-H4 indicates that the non-electrostatic terms effectively encode the steric information of molecules.

Comparison of the results obtained from the QM/MST-based hydrophobic H2 model and the standard CoMFA models[43-47] is shown in Figure 2 (comparison with CoMSIA results is shown in Figure S14; see Supporting Information).

In all cases there is a nice agreement between CoMFA and HyPhar models not only for the compounds pertaining to the training set, but also for molecules in the test set. Thus, the values of q^2 and S_{press} obtained from models H2-H4 outperforms (D2) or compares well (D4, GSK-3a) with the performance of CoMFA. The agreement is slightly less satisfactory for 2-aryl-4-chromanones and cruzain inhibitors. Thus, the S_{press} values increase from 0.24 (2-aryl-4-chromanones) and 0.41 (cruzain inhibitors) to 0.28-0.29 and 0.48-0.49, respectively, for the H2-H4 models. At this point, it has to be noticed that the set of 2-aryl-4-chromanones is the most challenging, as the range of biological activities differ by only 1.31 pI_{50} units. On the other hand, the results obtained for cruzain inhibitors may reflect the influence played by the binding site on the partial charges used in CoMFA (see above). Nevertheless, it is worth noting that the number of principal components required for H2-H4 models is limited to 2 or 3, while CoMFA and CoMSIA models included up to 5 (cruzain inhibitors) and 6 (2-aryl-4-chromanones) principal components. With regard to the thermolysin inhibitors, the statistical data derived from both CoMFA and CoMSIA calculations, including in this latter case models with two (steric and electrostatic fields) and three (steric, electrostatic and hydrophobic) fields, have been reported in Table 1 (data taken from Table 3 in ref. 47). Considering statistics for the two fields models, HyPhar H1-H4 models perform slightly better (q^2 of 0.55-0.64; S_{press} of 1.31-1.44) that CoMFA and CoMSIA (q^2 of 0.51; S_{press} of 1.51-1.53). Moreover, HyPhar models are obtained with fewer components (4-5) than

those required for the reference CoMFA and CoMSIA (5-7) models in ref. 47. This may likely be attributed to the more accurate description of the compounds afforded by the QM MST calculations compared to the point charges and L-J parameters assigned in standard CoMFA and CoMSIA analyses.

Table 2. Summary of statistical parameters obtained for the HyPhar models derived from the QM MST-based hydrophobic descriptors (models H1-H4) and CoMFA/CoMSIA results taken from refs. 43-47 for the set of test compounds.

	CoMFA [a]	CoMSIA	H1	H2	H3	H4
D2 (6 compounds)						
r^2	0.56 (0.68)	0.72	0.78	0.78	0.76	0.76
S_{PRESS}	0.81 (0.69)	0.64	0.58	0.51	0.59	0.60
RMSE [b]	0.66 (0.56)	0.53	0.47	0.47	0.49	0.49
D4 (6 compounds) [c]						
r^2	0.05 (0.25)	0.01	0.10	0.10	0.05	0.04
	0.24 (0.55)	*0.10*	*0.49*	*0.48*	*0.41*	*0.36*
S_{PRESS}	0.94 (0.83)	0.96	0.91	0.92	0.93	0.94
	0.86 (0.66)	*0.93*	*0.70*	*0.71*	*0.75*	*0.79*
RMSE	0.76 (0.68)	0.78	0.74	0.75	0.76	0.77
	0.66 (0.51)	*0.72*	*0.54*	*0.54*	*0.58*	*0.61*
2-Aryl-4-chromanones (7 compounds)						
r^2	0.91 (0.91)	0.93	0.86	0.81	0.82	0.88
S_{PRESS}	0.13 (0.14)	0.12	0.17	0.20	0.19	0.16
RMSE	0.11 (0.12)	0.10	0.14	0.17	0.16	0.14
GSK-3a (18 compounds)						
r^2	0.80 (0.74)	0.82	0.77	0.79	0.74	0.67
S_{PRESS}	0.21 (0.23)	0.20	0.22	0.21	0.24	0.26
RMSE	0.19 (0.22)	0.18	0.21	0.20	0.22	0.25
Cruzain (6 compounds)						
r^2	0.76 (0.67)	0.81	0.44	0.69	0.69	0.55
S_{PRESS}	0.41 (0.47)	0.37	0.62	0.47	0.46	0.56
RMSE	0.33 (0.39)	0.30	0.51	0.38	0.38	0.46
Thermolysin (15 compounds) [g]						
r^2	0.60[d] (0.63)	0.67[e]/0.79[f]	0.72	0.65	0.70	0.64
S_{PRESS}	1.26[d] (1.20)	1.14[e]/0.92[f]	1.05	1.18	1.09	1.18
RMSE	0.93[d] (0.93)	0.92[e]/0.93[f]	0.93	0.93	0.93	0.93

[a] Data taken from the original works. Data in parenthesis refer to CoMFA analysis performed in this study. [b] Root-mean square error. [c] Parameters obtained after exclusion of compound 34 are given in italics. [d] Statistical results from models obtained with a grid spacing of 1 Å. ComSIA results obtained by using two (electrostatic + steric; left) and three (electrostatic + steric + hydrophobic; right) grids are reported (data taken from ref. 47).

The prospective potential of the QM/MST-based hydrophobic descriptors can be calibrated from the results given in Table 2.

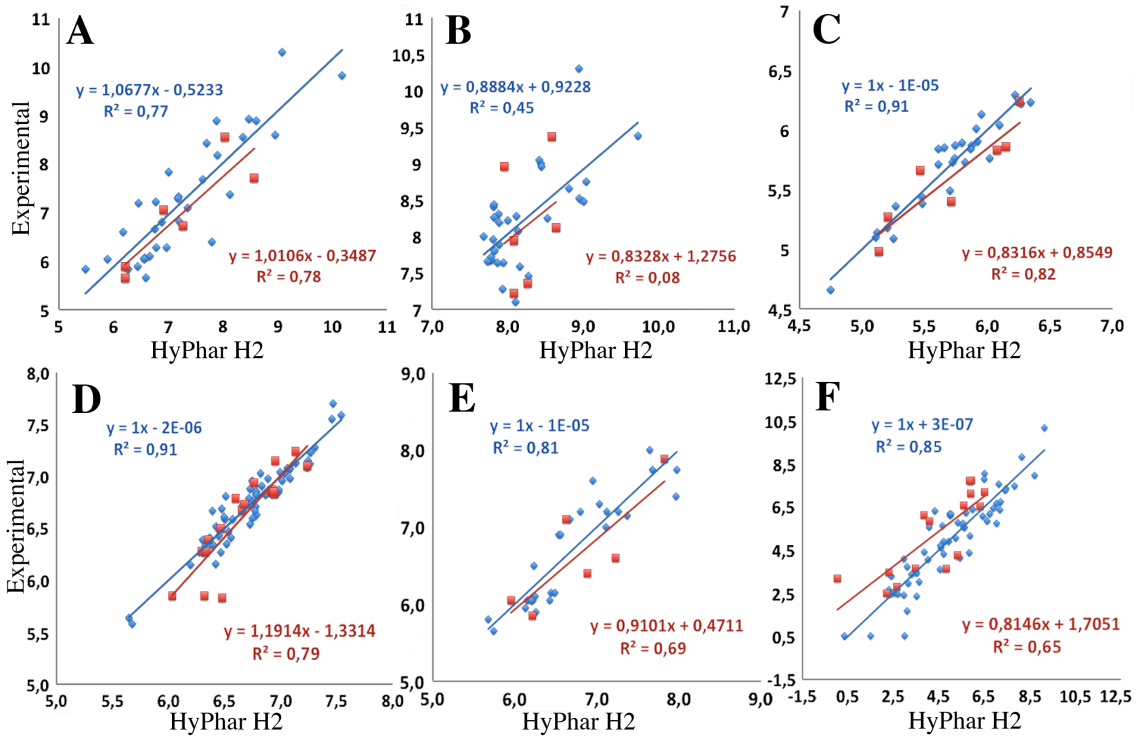


Figure 3. Comparison of the experimental data and the results predicted from the HyPhar H2 model for (A) D2, (B) D4, (C) 2-aryl-4-chromanones, (D) GSK-3α, (E) cruzain and (F) thermolysin systems. Compounds of the training/test set are shown in blue/red, respectively.

There is generally a nice correlation between the experimental activities and the HyPhar results, as noticed for the H2 model in Figure 3, especially for the test compounds of D2 antagonists, 2-aryl-4-chromanones, GSK-3, cruzain and thermolysin inhibitors. From a quantitative point of view, the results point out that H1-H4 models possess a predictive accuracy comparable to the original CoMFA and CoMSIA models.

The root-mean square error (RMSE) is generally lower than 0.5, but for the test sets of D4 antagonists (RMSE around 0.75) and thermolysin inhibitors (RMSE of 0.93), which also proved to be challenging datasets for CoMFA and CoMSIA.[43]

Nevertheless, for D4 test compounds HyPhar predictions improve after exclusion of compound 34, which has the highest D4 inhibitory activity among the tricyclic compounds, leading to RMSE values of 0.54-0.61, a trend also observed in CoMFA calculations. Nevertheless, HyPhar predictions for the test set of D4 compounds improve after exclusion of compound 34, which has the higher D4 inhibitory activity among the tricyclic compounds, leading to RMSE values of 0.54-0.61, a trend also observed in CoMFA calculations. Let us note that although compound 34 cannot be considered in stricto sensu an outlier (i.e., according to Grubb's test), it is highly influential on the regression model, as suggested by Cook's distance estimate,[50] which would thus make convenient to examine the effect of its exclusion in the predictive analysis.

Overall, the comparative analysis supports the reliability of the QM/MST-based hydrophobic descriptors for the calculation of molecular fields and the ability of the electrostatic and non-electrostatic components of $\log P$ to encode the electrostatic and steric components typically utilized in CoMFA studies.

Graphical analysis of the HyPhar pharmacophoric maps

Besides disclosing quantitative relationships between experimental data and molecular fields, a 3D-QSAR model should provide an easily interpretable graphical representation of physico-chemical properties relevant for the biological activity. In this context, even though our aim here is not to exploit the pharmacophoric models for drug design purposes, a graphical comparison of isocontour maps derived from CoMFA and HyPhar models is valuable to gain insight into the predictive potential of the QM/MST-derived hydrophobic descriptors. For the sake of clarity, we limit ourselves to the comparison of isocontour maps obtained from CoMFA and HyPhar H2 models. In all

cases maps were generated using the PLS coefficients corrected by the standard deviation (see Table S6 for the levels shown in Figures 4, 6, 8, 10 and 11). The discussion will first examine the molecular systems where X-ray crystallographic information is available (GSK-3a, cruzain and thermolysin).

*GSK-3*a *inhibitors*. Comparison of CoMFA and HyPhar H2 isocontour maps is shown in Figure 4, which reveals the large resemblance between the most relevant pharmacophoric areas obtained for electrostatic (Coulomb) and $\log P_{ele}$ fields, and for steric (L-J) and $\log P_{cav}$ descriptors.

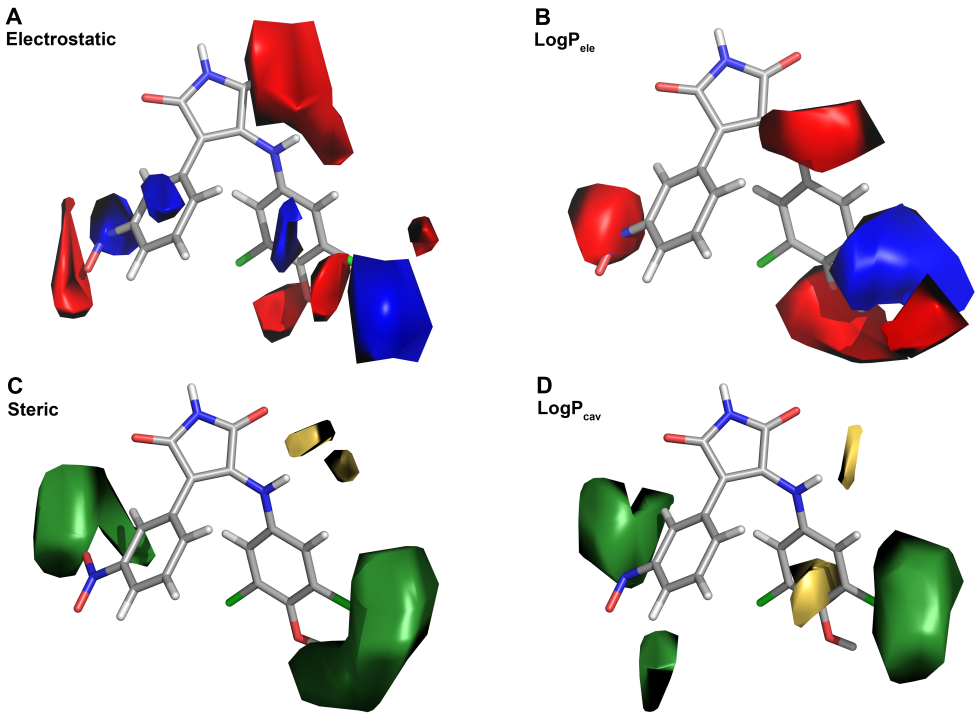


Figure 4. Comparison of CoMFA (A, C) and H2 (B, D) isocontour maps for GSK-3α inhibitors using as template the most potent compound within the series (compound 38; pIC50 of 7.70). Electrostatic fields (Coulomb in CoMFA and in HyPhar) are shown as blue (positive) and red (negative) isocontours, which correspond to areas where polarity decreases/increases the biological activity. Steric parameters (L-J in CoMFA and in HyPhar H2 model) are shown as yellow/green isocontours and denote areas where steric bulk is unfavourable/favourable.

Thus, both CoMFA and HyPhar identify three major regions that involve i) *meta*/*para* substituents on 3-aniline group (ring C in Figure 1C), ii) *ortho/meta* substituents on 4-phenyl ring (ring B in Figure 1C), and iii) hydrogen-bond donor/acceptor features of the pyrrolidin-2,5-dione moiety (ring A in Figure 1C), in agreement with previous results.[45] The H2 model points out that the presence of polar groups around the *meta* position of the phenyl ring, the aniline NH group, and the *para* position of the aniline ring would favor the inhibitory potency (Figure 4B).

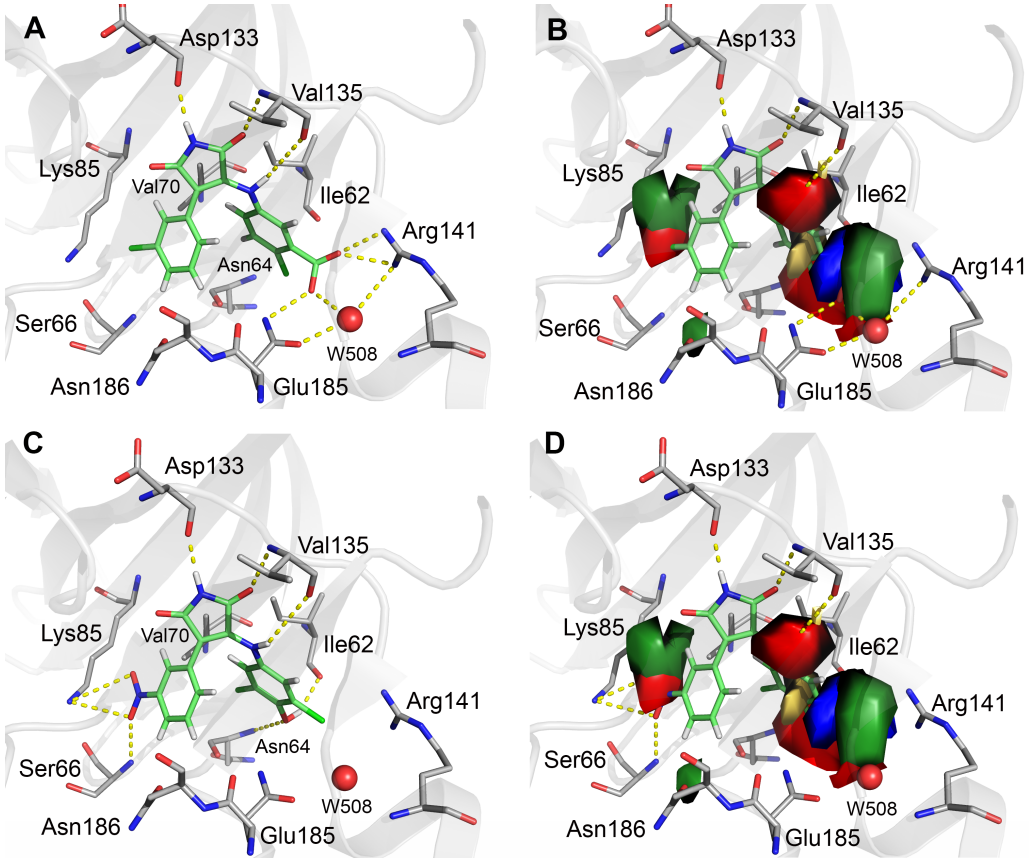


Figure 5. Crystallographic analysis of ligand-receptor complementarity (A, B) using as template the X-ray structure (PDB ID 1Q4L) of GSK-3β bound to compound 52 (pIC50 of 7.12). (C, D) Structure obtained upon replacement of compound 52 by compound 38 (pIC50 of 7.70). The isocontour maps from H2 model are shown in plots B and D. Water molecules are shown as red spheres.

Likewise, steric effects would be favorable around positions *ortho* and *meta* of the phenyl ring, and around position 3 of the aniline moiety (Figure 4D).

The interpretation of the maps is facilitated by the availability of the X-ray crystal structure of GSK-3b with compound 52 (PDB ID 1Q4L; see Figure 5), taking advantage of the high sequence similarity (97%) between the kinase domains of GSK-3a and GSK-3β.[45] Within the series of compounds a common motif is the pyrrolidine moiety, which would form hydrogen-bond interactions with the carbonyl group of Asp133 and the backbone NH unit of Val135. An enhancement in the polarity of the aniline NH group is expected to reinforce the electrostatic stabilization with the carbonyl group of Val135.

Thus, methylation of the NH group leads to a drastic reduction in activity (pIC_{50} values of 5.58-5.85 for compounds 65-67; see Table S3 and ref. 45). The presence of a polar (i.e, hydroxyl) substituent in *para* position of the aniline ring would enable hydrogen-bonds with the backbone NH of Asn64 and the CO unit of Ile62, while the presence of apolar groups (i.e, chlorine) in *meta* position would contribute to shield the hydrogen-bond from bulk water molecules. In fact, high inhibitory potencies are found for compounds with the 3,5-diCl,4-OH motif in the aniline ring (compounds 32-41; Table S4; pIC_{50} values of 6.83-7.70). Similarly, the presence of a polar group (NO_2) in *meta* position of the phenyl ring would lead to favorable interactions with the protonated amino group of Lys85 and the main-chain NH unit of Ser66. Thus, replacement of the NO_2 unit in compound 38 (pIC_{50} of 7.70) by hydrogen reduces the inhibitory potency by near 1 unit (compound 32; pIC_{50} of 6.83). Finally, the presence of hydrophobic groups in *ortho* position of the phenyl ring would fill the void pocket formed by side chains of Val70, Leu132 and Lys85.

Cruzain inhibitors. A significant similarity is found between the isocontour maps determined from CoMFA and H2 models, as shown in Figure 6. Three main pharmacophoric elements can be identified: i) electrostatic contributions around positions 3 and 5 of the aniline moiety (ring C in Figure 1D), ii) steric effects around position 9 of the purine ring (site A in Figure 1D, and iii) electrostatic and steric contributions around positions 2 and 7 of the purine ring. In line with previous results,[46] the H2 model reveals that steric effects around position 3 of the aniline ring and positions 8 and 9 of the purine moiety would be favorable, while substituents around positions 3 and 9 of the purine ring would be unfavorable.

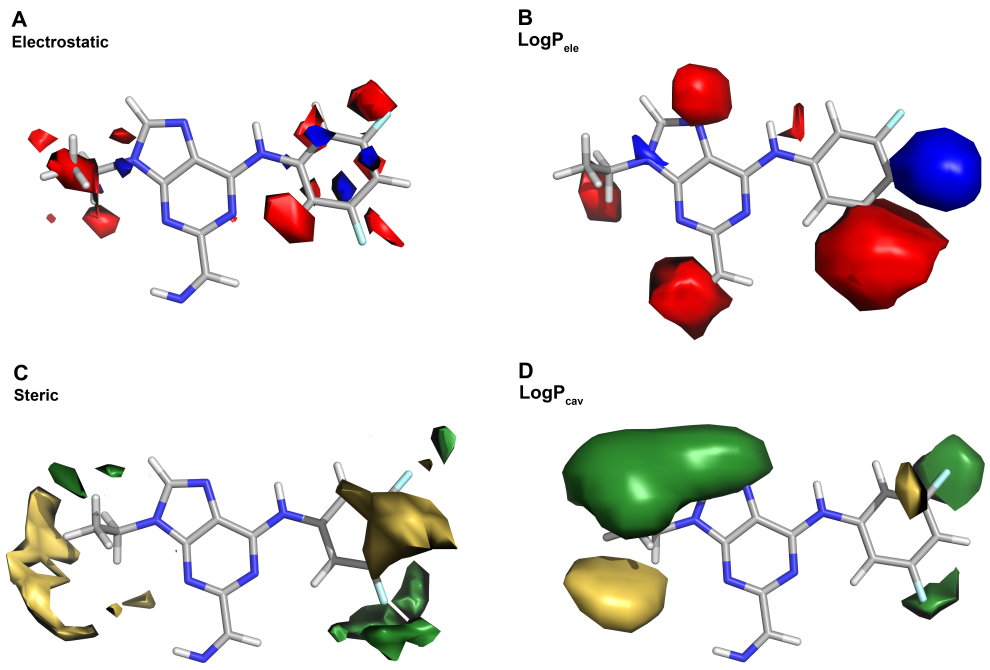


Figure 6. Comparison of CoMFA (A, C) and H2 (B, D) isocontour maps for cruzain inhibitors using as template the most potent compound within the series (compound 23; pIC50 of 8.00). Electrostatic fields (Coulomb in CoMFA and in HyPhar) are shown as blue (positive) and red (negative) isocontours, which correspond to areas where polarity decreases/increases the biological activity. Steric parameters (L-J in CoMFA and in HyPhar H2 model) are shown as yellow/green isocontours and denote areas where steric bulk is unfavourable/favourable.

On the other hand, polar groups around positions 3 and 6 should enhance the biological activity, whereas their attachment to *para* position of the aniline ring would be detrimental. The inhibition mechanism of cruzain involves the nucleophylic attack of Cys25 to the nitrile moiety of the ligand (see Figure 7). To this end, the aniline moiety is placed in a hydrophobic cavity in the binding site, while the five-membered ring of the purine moiety is exposed to the bulk solvent.

The presence of a small group in *meta* position of the aniline ring is favorable for the inhibitory potency because it fills the hydrophobic site shaped by Trp26 and Met68. Thus, triazine 14 and purines 26 and 27, with 3-chloro at the aniline ring, have pIC_{50} values ranging from 7.10 to 7.74. Nevertheless, a certain degree of polarity is also required because of the presence of the carboxylate group of Glu208 at around 6 Å from the *para* position of the aniline ring. In fact, the most potent inhibitors contain a 3,5-difluoroaniline moiety (compounds 20, 22-25; pIC_{50} values of 7.20-8.00), facilitating the formation of water-bridged contacts.

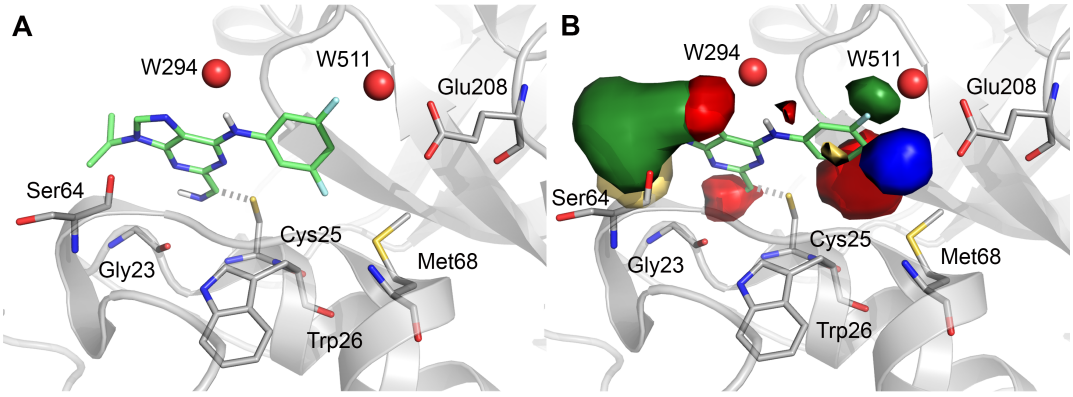


Figure 7. Crystallographic analysis of ligand-receptor complementarity (A) using as template the X-ray structure (PDB ID 3I06) of cruzain bound to compound 23 (pIC50 of 8.00). The isocontour maps from H2 model are shown in plot B. The covalent bond with Cys25 is shown as dashed yellow line. Water molecules are shown as red spheres. The two orientations found for the solvent-exposed ethyl group placed in front of Ser64 are shown.

On the other hand, the attachment of a bulky group at position 9 of the purine ring is detrimental due to potential clashes with Gly23 and Ser64, which would impede a suitable arrangement of the compound for covalent inhibition of the enzyme. In this line, all triazine compounds with a cyclopentyl group in this position have inhibitory potencies in the range $6.40<pIC_{50}<7.20$ (compounds 12-21). Finally, the red and green isocontours found over the nitrogen atom at position 7 of purine reflects the expansion of the central ring (from triazine to purine) and the solvent exposure of this nitrogen to bulk waters.

Thermolysin. Comparison of CoMFA and HyPhar H2 isocontour maps is shown in Figure 8, which also shows the X-ray structure of the most potent inhibitor within the series (ZFPLA; PDB ID: 4TMN; *pKi* of 10.17).[47] Four areas important for activity can be revised (A-D in Figure 1E). The first (A) denotes the interaction of the oxygen atoms with the Zn cation, while the interaction of the ligand to the binding pocket is assisted by a variety of hydrogen bonds (B, D). Finally, the side chain of the leucine moiety in the ligand fills a hydrophobic pocket (denoted as S1' subsite and shaped by residues Leu133, Phe130 and Val 139), which is important for tight binding.

Inspection of the $\log P_{cav}$ isocontours highligths the importance of filling the S1' subsite, for improving the inhibitory potency (see Figure 9). Furthemore, the results also identify the sterically unfavorable region around the group coordinated to the Zn cation the central part of the ligand. While these trends are common to both CoMFA and HyPhar models, distinct trends can be identified for the sterically favorable regions around the methyl side chain of the alanine residue and around the benzyl moiety present in ZFPLA.

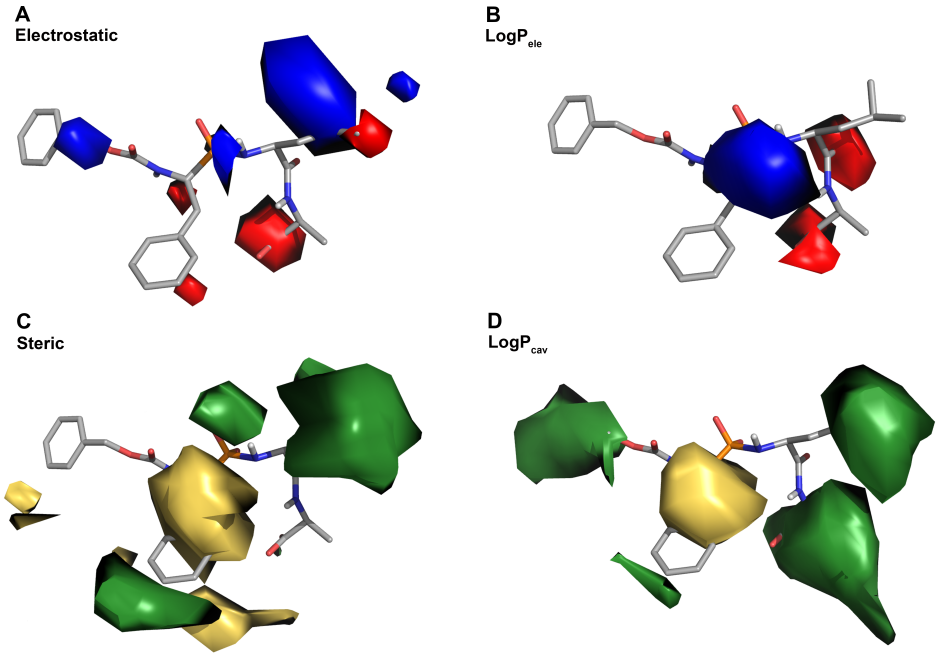


Figure 8. Comparison of CoMFA (A, C) and H2 (B, D) isocontour maps for thermolysin inhibitors using as template the most potent compound within the series (ZFPLA; pKi of 10.17). Electrostatic fields (Coulomb in CoMFA and in HyPhar) are shown as blue (positive) and red (negative) isocontours, which correspond to areas where polarity decreases/increases the biological activity. Steric parameters (L-J in CoMFA and in HyPhar H2 model) are shown as yellow/green isocontours and denote areas where steric bulk is unfavourable/favourable.

The former can be interpreted in terms of filling the S2' subsite, which is able to accommodate larger groups such as the indole ring of tryptophan. This is noted in the activity of compounds DAH50 and DAH53, which differ by the substitution of the benzene moiety present in DAH53 (pK_i of 6.66) by an indole ring in DAH50 (pK_i of 7.96). On the other hand, the green isocontour around the terminal benzyl unit can be related to the stacking interaction formed with the aromatic ring of the Tyr157 residue.

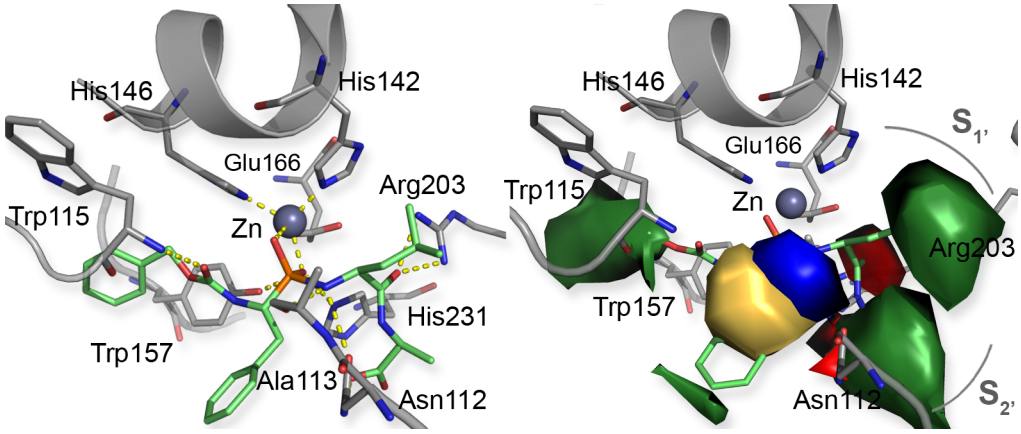


Figure 9. Crystallographic analysis of ligand-receptor complementarity (A) using as template the X-ray structure (PDB ID 4TMN) of endopeptidase thermolysin bound to ZFPLA (pKi of 10.17). The isocontour maps from the Hyphar H2 model are shown in plot B. The two hydrophobic sites, $S_{1'}$ and $S_{2'}$, are also shown. The zinc metal cation is showns as a violet sphere.

These features reproduces similar trends found in CoMSIA analysis for these compounds.[43] With regard to the $\log P_{ele}$ field, the blue isocontour found around the Zn-coordinating group reflects unfavorable electrostatic contacts due to polar/apolar mismatches between ligand and target. This explains the gain in inhibitory potency found for compounds ZGPOLL (pK_i of 5.05), ZGPCLL (pK_i of 6.74) and ZGPLL (pK_i of 8.04), as this series of compounds imply the change from -O- (ZGPOLL) to -CH_2- (ZGPCLL) to -NH_2- (ZGPLL) in the position vicinal to the Zn-coordinating site, which faces the electrostatic interaction with the carbonyl oxygen of Ala113. Moreover, the red isocontours denote the favourable interactions related to hydrogen bonding between the carbonyl units of the ligand and the side chains of Arg203 and Asn112. Again, these features agree with the findings reported by Klebe and co-workers.[43]

Dopamine D2/D4 antagonists. To the best of our knowledge, there is no X-ray structure available for the sets of D2/D4 antagonists and 2-aryl-4-chromanones.

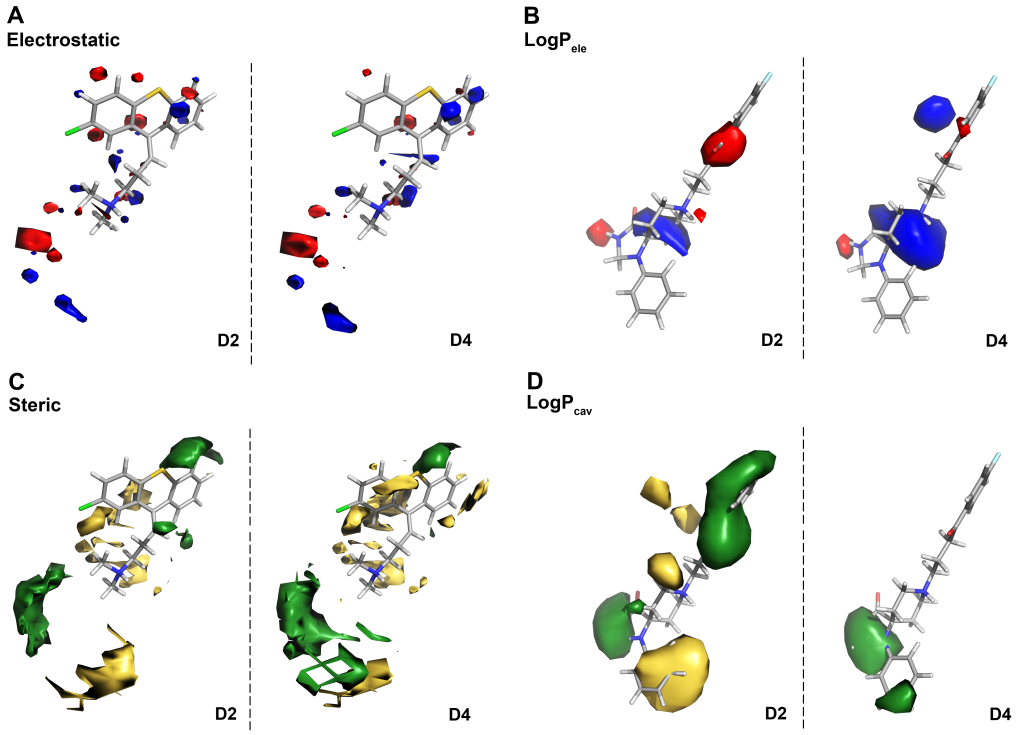


Figure 10. Comparison of CoMFA (A,C) and H2 (B,D) isocontour maps for dopamine D2/D4 inhibitors. The structure of the most potent compounds within the two series of inhibitors (10, pKi of (D2) 8.92 and (D4) 8.41; 23: pKi of (D2) 10.30 and (D4) 10.30) is also shown. Electrostatic fields (Coulomb in CoMFA and in HyPhar) are shown as blue (positive) and red (negative) isocontours, which correspond to areas where polarity decreases/increases the biological activity. Steric parameters (L-J in CoMFA and in HyPhar H2 model) are shown as yellow/green isocontours and denote areas where steric bulk is unfavourable/favourable.

Hence, we limit the discussion to the comparison of the pharmacophores derived from CoMFA and HyPhar H2 models, which are shown in Figures 10 and 11 for the sets of D2/D4 antagonists and 2-aryl-4-chromanones, respectively.

While the CoMFA maps are highly similar for D2 and D4 receptors, relevant differences are found in the maps from HyPhar, which could be relevant to explain the selectivity against the two receptors. The isocontour maps (Figure 10) identify three main pharmacophoric regions corresponding to aromatic rings (denoted A, B and G in Figure 1A) and the ammonium nitrogen (site C in Figure 1A).

Regarding the non-polar contributions, filling the space around rings B is favorable, but this trend is only seen for the binding to D2. In constrast, a more complex profile is seen for ring G, since some areas around ring G would favor/disfavor sterically the binding to D2, whereas steric bulk favors binding to D4. Noteworthy, these findings are in agreement with the steric features reported from CoMSIA (Figure 6 from ref. 43). With regard to electrostatic terms, the H2 pharmacophore also reveals differential trends between D2 and D4, which are not observed in CoMFA maps. In particular, a polar group between rings A and B favors the activity against D2 receptor, whereas this feature is less prominent in D4. Furthermore, the presence of apolar groups around ring A is beneficial for activity, a trend not found in D2. Remarkably, these findings also reflect the features observed in CoMSIA (Figure 7 in ref. 43).

Antifungal 2-aryl-4-chromanones. Two main pharmacophoric regions are found in the H2 model (Figure 11). The first involves the area surrounding positions 6 and 7 on the chromanone scaffold, and the second one is located around the heterocyclic 4-aryl moiety (ring B in Figure 1B). Green isocontour reveals that a certain steric hindrance is tolerated around positions 6 and 7 on the chromanone scaffold, while apolar groups would be beneficial. In thiese positions the most potent inhibitors (compounds 2, 19, 20, 22, 23 and 33) contain substituents such as bromine, methyl and methoxy (pI_{50} values ranging from 6.01 to 6.29), and their replacement by hydrogen leads to a substantial reduction in activity (compounds 3, 8, 17, 18, 20-22, 24-32; $4.66<pI_{50}<5.85$).

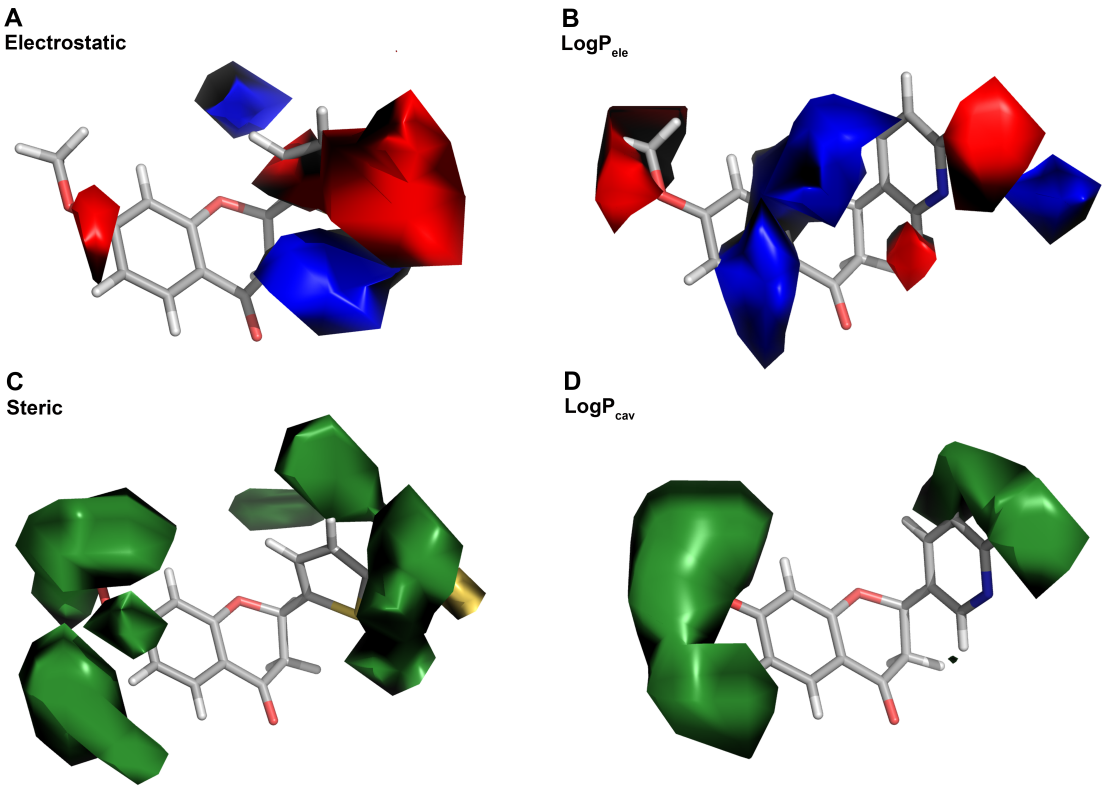


Figure 11. Comparison of CoMFA (A, C) and H2 (B, D) isocontour maps for 2-aryl-4-chromanones. The structure of the most potent compounds (2 and 23; pI50 of 6.29 and 6.24) is also shown. Electrostatic fields (Coulomb in CoMFA and in HyPhar) are shown as blue (positive) and red (negative) isocontours, which correspond to areas where polarity decreases/increases the biological activity. Steric parameters (L-J in CoMFA and in HyPhar H2 model) are shown as yellow/green isocontours and denote areas where steric bulk is unfavourable/favourable.

On the other hand, the green contour around ring B shows the favorable effect of substituents in *meta* and *para* (compounds 24, 25, 28, 33 and 34, with $5.73<pI_{50}<6.23$). While apolar groups (OCH_3 and Cl) are favored in *para* position, a nitro substituent is beneficial in *meta* position.

Final Remarks

Desolvation is recognized to be one of the major forces that modulate the binding of ligands to the target receptors. Hence, the hydrophobic/hydrophilic balance of a

molecule, particularly the hydrophobic/hydrophilic complementarity between ligand and receptor, is a key requirement for the binding affinity. Indeed, the relevance of pocket shape and hydrophobicity in drug binding has been highlighted in previous studies of target druggability.[57-60] Specifically, despite the complexity of factors that modulate the binding affinity of drugs,[61] ligand desolvation is largely responsable of the variation in maximal achievable binding energy for a drug-like molecule.[62] Nevertheless, while druggable binding sites appear to be closed and "greasy" cavities, polar interactions are crucial for binding and selectivity.[63-65] In this context, it is not surprising that previous efforts have attempted to develop QM-based strategies for the calculation of lipophilic descriptors. This is reflected in the heuristic molecular lipophilic potential,[66] which relies on the analysis of the electrostatic potential at the molecular surface to provide a unified lipophilicity and hidrophilicity potential. More recently, Klamt and coworkers have examined the use of s-profiles within the framework of the COSMO solvation model as an alternative to force-field based molecular interaction fields.[67] In particular, a local, grid-based sigma s-profiles are shown to have a linear dependency with the biological activity, and the application to different datasets have demonstrated the robustness of the predictive models.[68]

In this work, following our previous studies on hydrophobic similarity,[69-71] the QM/MST-based hydrophobic contributions have been utilized as physicochemical descriptors suitable for 3D-QSAR studies. By combining the electrostatic and non-electrostatic components of the octanol/water partition coefficient, which can obtained through a suitable partitioning of the solvation free energy,[39,40] the 3D-QSAR models derived for the five molecular systems have a predictive accuracy that compares well with standard CoMFA and CoMSIA techniques. Furthermore, the graphical representation of the pharmacophoric fields closely agrees with the key features derived

from CoMFA models, which suggest that the "electrostatic" and "steric" components typically utilized in 3D-QSAR studies are effectively encoded in the $\log P_{ele}$ and $\log P_{n-ele}$ components of the octanol/water partition coefficient. Overall, the results support the suitability of QM MST-based fractional contributions to the hydrophobicity as descriptors to be used in ligand-based drug discovery.

Even though the QM treatment of the molecules inherently demands a computational cost larger than the use of descriptors taken from classical force-fields, it benefits from a more accurate description of the molecular charge distribution, which takes into account specific effects associated to the ionization, tautomerization and conformational state of molecules. This likely explains the finding that the predictive accuracy of CoMFA and CoMSIA models is generally achieved with a lower number of principal components with HyPhar models, at least for the set of molecular systems examined here. On the other hand, even though present results have been derived by using the QM MST solvation continuum model, extension to other parametrized version of the IEF-PCM model, such as the parametrized version of the Solvation Model with Electron Density (SMD)[72] should be straigthforward, thus expanding the range of application of the HyPhar methodology.

Supporting Information

Representation of the chemical structures of compounds in the five test systems (Figures S1-S6), histograms of molecular properties (Figure S7-S11), analysis of orthogonality for the MST-derived hydrophobic descriptors (Figure S12), graphical comparison of published CoMFA *versus* our CoMFA results (Figure S13) and of reported CoMSIA *versus* the HyPhar H2 model (Figure S14), and tables reporting the

experimental activity and the estimated value from HyPhar H2 model for the five test systems (Tables S1-S5), and the levels used in the isocontour maps (Table S6).

Acknowledgment

We thank the financial support from Ministerio de Economía y Competitividad (SAF2014-57094-R) and the Generalitat de Catalunya (2014-SGR-1189). We are grateful to the Consorci de Serveis Universitaris de Catalunya for computational resources. FJL acknowledges the support from ICREA Academia.

References

[1] J. J. Sutherland., L. A. O'Brien, D. F. Weaver, *J. Med. Chem*. **2004**, *47*, 5541.

[2] W. Sippl, in Pharmacophores and Pharmacophore Searches. Methods and Principles in Medicinal Chemistry Series; T. Langer, R. D. Hoffman, Eds.; Wiley, Weinheim, **2006**; Vol. 32, Chapter 11, pp 223-249.

[3] J. Verma, V. M. Khedkar, E. C. Coutinho, *Curr. Top. Med. Chem*. **2010**, *10*, 95.

[4] A. Artese, S. Cross, G. Costa, S. Distinto, L. Parrotta, S. Alcaro, F. Ortuso, G. Cruciani, *WIRES Comput. Mol. Sci*. **2013**, *3*, 594.

[5] R. D. Cramer III, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959.

[6] G. Klebe, U. Abraham, T. Mietzner, *J. Med. Chem.* **1994**, *37*, 4130.

[7] T. R. Kroemer, P. Hecht, *J. Comput.-Aided Mol. Design* **1995**, *9*, 205.

[8] J. L. Melville, J. D. Hirst, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1294.

[9] T. Sulea, T. I. Oprea, S. Muresan, S. L. Chan, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1162.

[10] T. Kotani, K. Higashiura, *J. Med. Chem*. **2004**, *47*, 2732.

[11] A. N. Jain, K. Koile, D. Chapman, *J. Med. Chem*. **1994**, *37*, 2315.

[12] G. Klebe, U. Abraham, *J. Comput.-Aided Mol. Des.* **1999**, *13*, 1.

[13] M. Böhm, G. Klebe, *J. Med. Chem.* **2002**, *45*, 1585.

[14] P. Gaillard, P.-A. Carrupt, B. Testa, A. Boudon, *J. Comput.-Aided Mol. Des*. **1994**, *8*, 83.

[15] G. E. Kellog. , S. F. Semus, D. J. Abraham, *J. Comput.-Aided Mol. Des*. **1991**, *5,* 545.

[16] G. E. Kellog, D. J. Abraham, *Eur. J. Med. Chem*. **2000**, *35*, 651.

[17] P. J. Goodford, *J. Med. Chem*. **1985**, *28*, 849.

[18] G. Cruciani, P. Crivori, P.-A. Carrupt, B. Testa, *J. Mol. Struct. (THEOCHEM)* **2000**, *503*, 17.

[19] P. Crivori, G. Cruciani, P.-A. Carrupt, B. Testa, *J. Med. Chem*. **2000**, *43*, 2204.

[20] Kotani. T.; Higashiura. K. *J. Med. Chem*. **2004**, *47*, 2732.

[20] B. D. Silverman, D. E. Platt, *J. Med. Chem*. **1996**, *39*, 2129.

[21] A. M. Ferguson, T. Heritage, P. Jonathon, S. E. Pack, L. Phillips, J. Rogan, P. J. Snaith, *J. Comput.-Aided Mol. Des.* **1997**, *11*, 143.

[22] D. B. Turner, P. Willet, A. M. Ferguson, T. W. Heritage, *J. Comput.-Aided Mol. Des.* **1999**, *13*, 271.

[23] R. Bursi, T. Dao, T. van Wijk, M. de Gooyer, E. Kellenbach, P. Verwer, *J. Chem. Inf. Comput. Sci.* **1999**, ***39***, 861.

[24] E. Besalú, X. Gironés, L. Amat, R. Carbó. R. *Acc. Chem. Res*. **2002**, *35*, 289.

[25] T. Zhou, D. Huang, A. Caflisch. A. *Curr. Top. Med. Chem*. **2010**, *10*, 33.

[26] K. Raha, M. B. Peters, B. Wang, N. Yu, A. M. Wollacott, L. M. Westerhoff, K. M. Merz Jr., *Drug. Discov. Today* **2007**, *12*, 725.

[27] S. Dixon, K. M. Merz Jr., G. Lauri, J. C. Ianni, *J. Comput. Chem*. **2005**, *26*, 23.

[28] M. B. Peters, K. M. Merz Jr., *J. Chem. Theory Comput*. **2006**, *2*, 383.

[29] S. Güssregen, H. Matter, M. Henneman, T. Clark, *J. Chem. Inf. Mod*. **2013**, *53*, 1486.

[30] J. Wan, L. Zhang, G. Yang, *Comput. Chem*. **2004**, *25*, 1827.

[31] S. Van Damme, P. Bultinck, *J. Comput. Chem*. **2009**, *30*, 1749.

[32] R. Dolezal, S. Van Damme, P. Bultinck, K. Waisser, *Eur. J. Med. Chem*. **2009**, *44*, 869.

[33] S. Güssregen, H. Matter, G. Hessier, M. Müller, F. Schmidt, T. Clark, *J. Chem. Inf. Model.* **2012**, *52*, 2441.

[34] F. J. Luque, X. Barril, M. Orozco, *J. Comput.-Aided Mol. Des*. **1999**, *13*, 139.

[35] C. Curutchet, M. Orozco, F. J. Luque, *J. Comput. Chem*. **2001**, *22*, 1180.

[36] I. Soteras, C. Curutchet, A. Bidon-Chanal, M. Orozco, F. J. Luque, *J. Mol. Struct. (THEOCHEM)* **2005**, *727*, 29.

[37] B. Mennucci, E. Cancès, J. Tomasi, *J. Phys. Chem. B* **1997**, *101*, 10506.

[38] C. Hansch, A. Leo, Exploring QSAR: Fundamentals and Applications in Chemistry and Biology; S. R. Heller, Ed.; American Chemical Society, Washington, **1995**; Vol. 1.

[39] F. J. Luque, C. Curutchet, J. Muñoz-Muriedas, A. Bidon-Chanal, I. Soteras, A. Morreale, J. L. Gelpí, M. Orozco, *Phys. Chem. Chem. Phys*. **2003**, *5*, 3827.

[40] F. J. Luque, J. M. Bofill, M. Orozco, *J. Chem. Phys*. **1995**, *103*, 10183.

[41] R. A. Pierotti, *Chem. Rev*. **1976**, *76*, 717.

[42] P. Claverie, in Intermolecular Interactions: From Diatomics to Biopolymers; B. Pullman, Ed.; Wiley, New York, **1978**; Vol. 1. pp 69-305.

[43] J. Boström, M. Böhm, K. Gundertofte, G. Klebe, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1020.

[44] D. G. Wei, G. F. Yang, J. Wan, C. G. Zhan, *J. Agric. Food Chem.* **2005**, *53*, 1604.

[45] S. Prasanna, P. R. Daga, A. Xie, R. J. Doerksen, *J. Comput. Aided Mol. Des.* **2009**, *23*, 113.

[46] O. Méndez-Lucio, J. Pérez-Villanueva, A. Romo-Mancillas, R. Castillo, *Med. Chem. Commun.* **2011**, *2*, 1058.

[47] M. Böhm, J. Stürzebecher, J, G. Klebe, *J Med. Chem.* **1999**, 42, 458-477.

[48] J. Boström, K. Gundertofte, T. Liljefors, *J. Comput.-Aided Mol. Des*. **2000**, 14, 769-786.

[49] T. Sander, J. Freyss, M. von Korff, C. Rufener, *J. Chem. Inf. Model*. **2015**, *55*, 460.

[50] H. Aguinis, R. K. Gottfredson, H. Joo, *Organ. Res. Methods.* **2013**, *16*, 270.

[51] Sybyl 8.1; Tripos Inc.; St. Louis, MO; **2008**.

[52] B. T. Mott, R. S. Ferreira, A. Simeonov, A. Jadhav, K. K. Ang, W. Leister, M. Shen, J. T. Silveira, P. S. Doyle, M. R. Arkin, J. H. McKerrow, J. Inglese, C. P. Austin, C. J. Thomas, B. K. Shoichet, D. J. Maloney, *J Med Chem.* **2010**, *53*, 52.

[53] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, Gaussian 09, Revision D.01; Gaussian. Inc.; Wallingford, CT, **2009**.

[54] PharmQSAR; Pharmacelera; Barcelona; **2015**.

[55] S. Wold, M. Sjöström, L. Eriksson, *Chemometr. Intell. Lab.* **2001**, *58*, 109.

[56] D. L. Alexander, A. Tropsha, D. A. Winkler, *J. Chem. Inf. Model.* **2015**, *55*, 1316.

[57] M. R. Arkin, J. A. Wells, *Nat. Rev. Drug Disc.* **2004**, *3*, 301.

[58] P. J. Hajduk, J. R. Huth, S. W. Fesik, *J. Med. Chem.* **2005**, *48*, 2518.

[59] M. Nayal, B. Honig, *Proteins* **2006**, *63*, 892.

[60] U. Egner, R. C. Hillig, *Expert. Opin. Drug Discov*. **2008**, *3*, 391.

[61] G. Klebe, *Nat. Rev. Drug. Discov*. **2015**, *14*, 95.

[62] A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg, E. S. Huang, *Nat. Biotechnol*. **2007**, *25*, 71.

[63] P. Schmidtke, X. Barril, X. *J. Med. Chem*. **2010**, *53*, 5858.

[64] P. Schmidtke, F. J. Luque, J. B. Murray, X. Barril, *J. Am. Chem. Soc*. **2011**, *133*, 18903.

[65] D. Alvarez-Garcia, X. Barril, *J. Med. Chem*. **2014**, *57*, 8530.

[66] Q. Du, P.-J. Liu, P. G. Mezey, *J. Chem. Inf. Model*. **2005**, *45*, 347.

[67] M. Thormann, A. Klamt, K. Wichmann, *J. Chem. Inf. Model*. **2012**, 52, 2149.

[68] A. Klamt, M. Thormann, K. Wichmann, P. Tosco, *J. Chem. Inf. Model*. **2012**, *52*, 2157.

[69] J. Muñoz, X. Barril, B. Hernández, M. Orozco, F. J. Luque, *J. Comput. Chem*. **2002**, *23*, 554.

[70] J. Muñoz-Muriedas, S. Perspicace, N. Bech, S. Guccione, M. Orozco, F. J. Luque, *J. Comput.-Aided Mol. Des*. **2005**, *19*, 401.

[71] J. Muñoz-Muriedas, X. Barril, J. M. López, M. Orozco, F. J. Luque, *J. Mol. Mod*. **2007**, *13*, 357.

[72] A. V. Marenich, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. B* **2009**, *113*, 6378.

Supporting Information

Development and Validation of Hydrophobic Molecular Fields Derived from the Quantum Mechanical IEF/PCM-MST Solvation Model in 3D-QSAR

Tiziana Ginex,[1] Jordi Muñoz-Muriedas,[2] Enric Herrero,[3] Enric Gibert,[3] Pietro Cozzini,[1*] and F. Javier Luque[4*]

1 Dipartimento di Scienze degli Alimenti, University of Parma, Parco Area delle Scienze 59/A, 43121 Parma, Italy

2 GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, United Kingdom

3 Pharmacelera, Jordi Girona 1-3, Campus Nord Universitat Politècnica de Catalunya, Edific K2M, 08034 Barcelona, Spain

4 Department of Chemical Physics and Institut de Biomedicina (IBUB), Faculty of Pharmacy, University of Barcelona, Av. Prat de la Riba 171, 08921 Santa Coloma de Gramenet, Spain

Figure S1. Chemical structures for Dopamine D2/D4 inhibitors (D2/D4 training set).

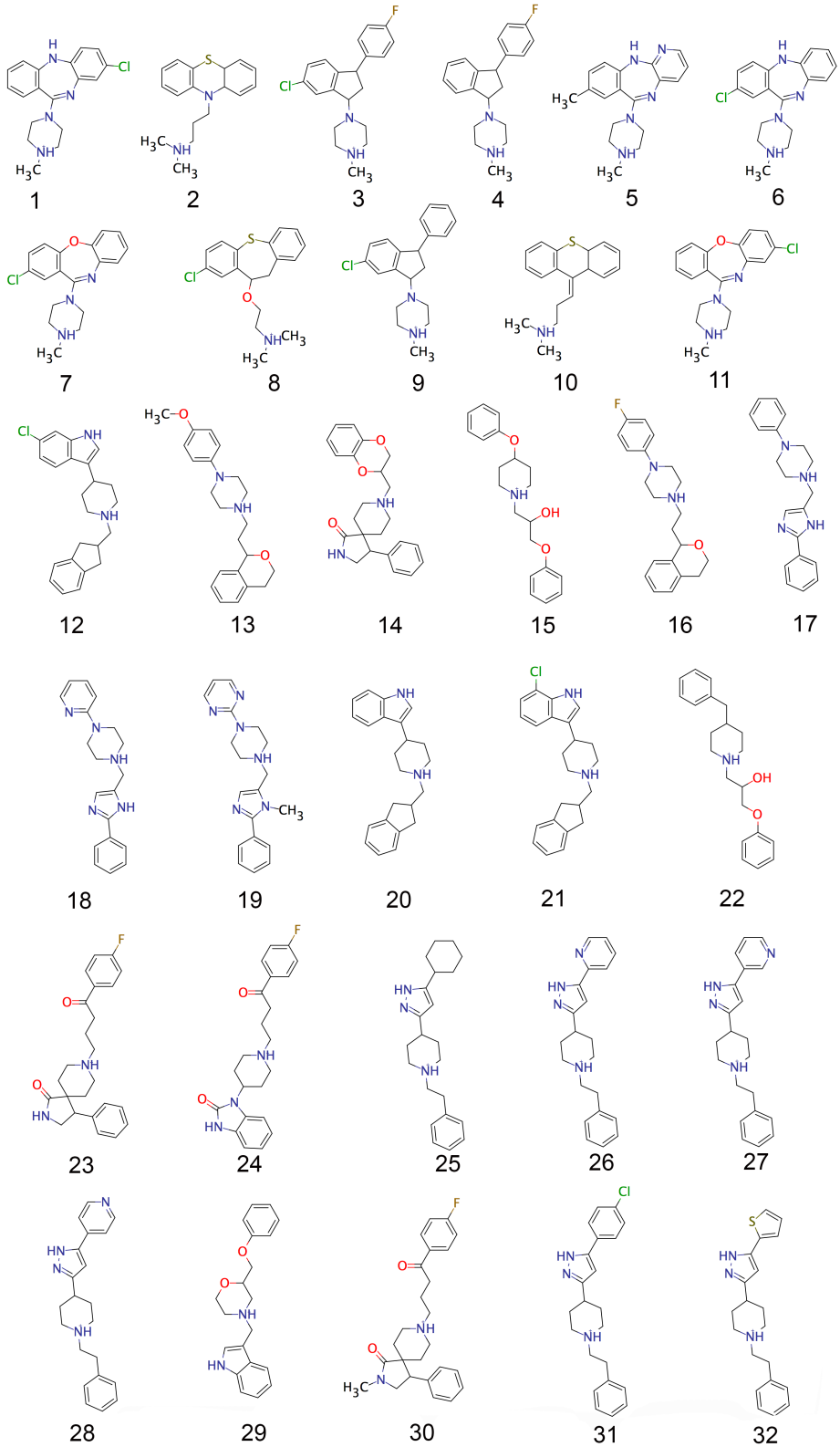


Figure S2. Chemical structures for Dopamine D2/D4 test set inhibitors (D2: 33-38; D4: 34, 36, 37, 39-41).

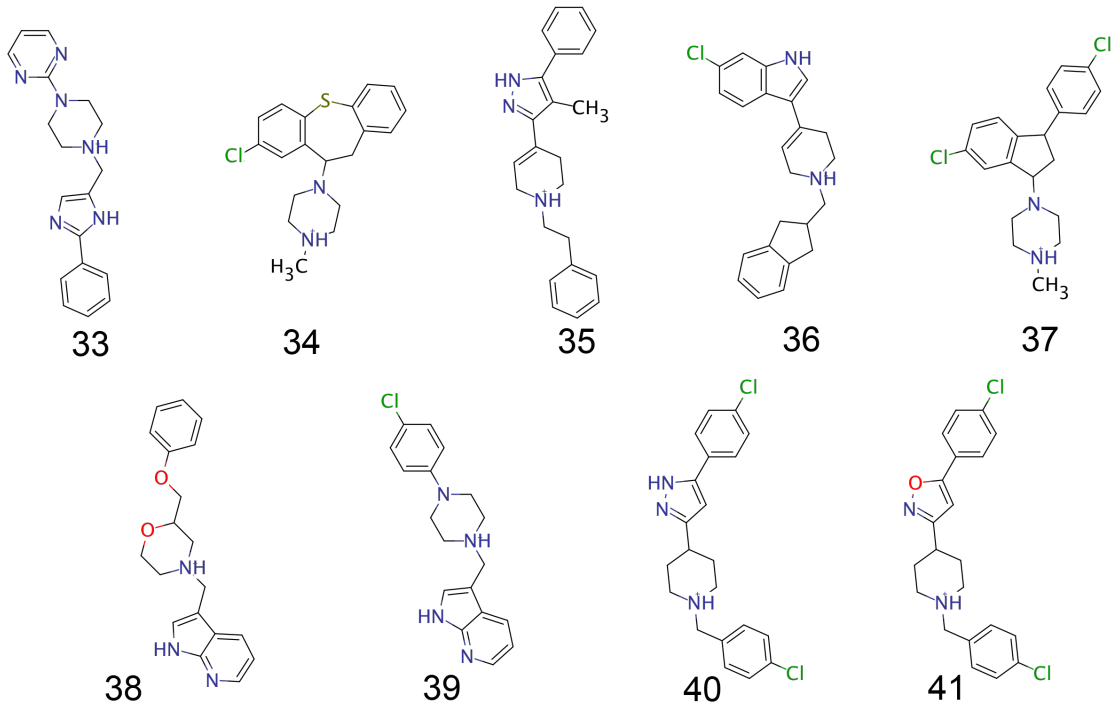
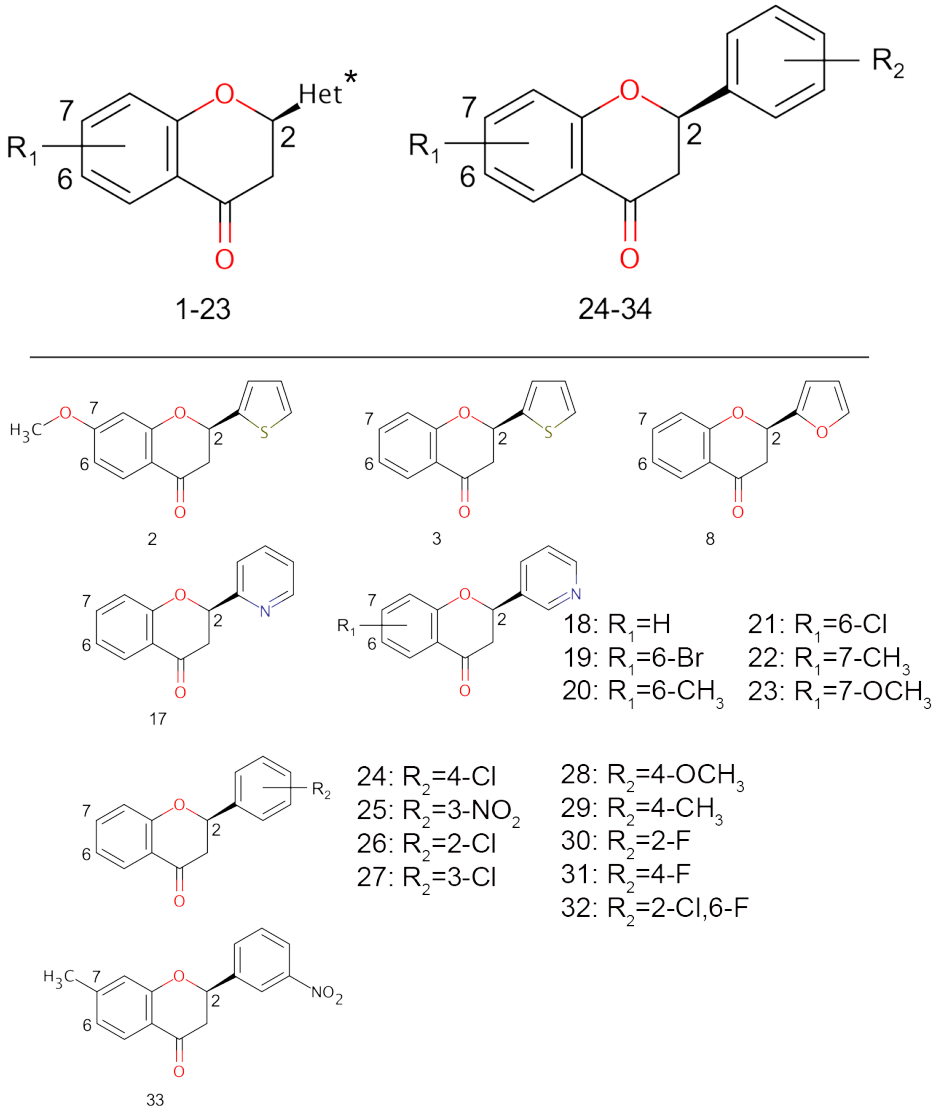
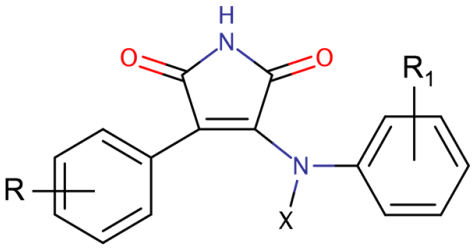


Figure S3. Chemical structures of the two main scaffolds present in the set of antifungal 4-aryl-2-chromanones. Chemical structures of compounds (2, 3, 8, 17-19, 20-33) discussed in the text are shown.

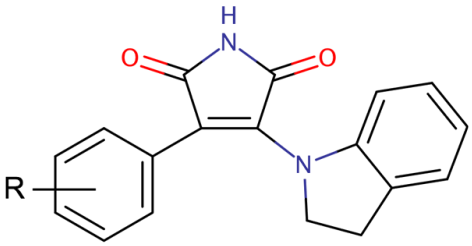


* Het: 2-thienyl, 2-furanyl, 2-pyridinyl or 3-pyridinyl.

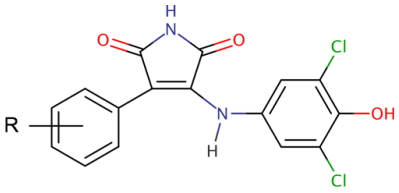
Figure S4. Chemical structures of the two main scaffolds for the series of GSK-3α inhibitors. Chemical structures of compounds (32-41, 52, 65 and 67) discussed in the text are shown.



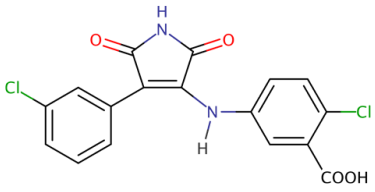
1–67



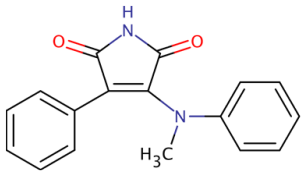
68–74



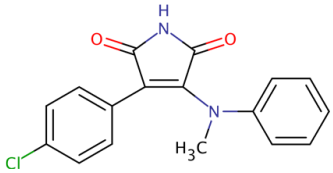
32: H
33: 2-Cl
34: 2-OCH_3
35: 2-NO_2
36: 3-Cl
37: 3-OCH_3
38: 3-NO_2
39: 4-Cl
40: 4-OCH_3
41: 4-NO_2



52

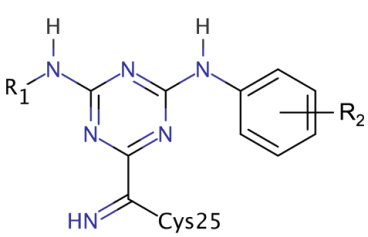


65

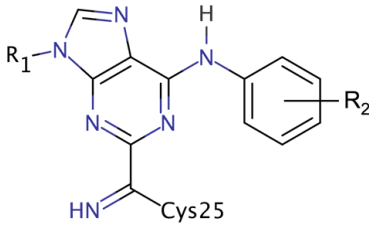


67

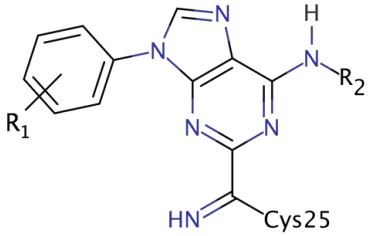
Figure S5. Chemical structures of the three main scaffolds in the set of cruzain inhibitors. Chemical structures of compounds (12-25) discussed in the text are shown.



1-22



23-27



28-32

12: R_1=Cyclopentyl; R_2=3-NO_2
13: R_1=Cyclopentyl; R_2=3-F
14: R_1=Cyclopentyl; R_2=3-Cl
15: R_1=Cyclopentyl; R_2=3-Br
16: R_1=Cyclopentyl; R_2=4-F
17: R_1=Cyclopentyl; R_2=3-CH_3
18: R_1=Cyclopentyl; R_2=4-Br
19: R_1=Cyclopentyl; R_2=3-Phenyl
20: R_1=Cyclopentyl; R_2=3,5-diF
21: R_1=Cyclopentyl; R_2=3,5-diCl
22: R_1=2,2-difluoroethyl; R_2=3,5-diF
23: R_1=Ethyl; R_2=3,5-diF
24: R_1=2,2-difluoroethyl; R_2=3,5-diF
25: R_1=Cyclopentyl; R_2=3,5-diF

Figure S6a. Chemical structures of thermolysin inhibitors (1-20).

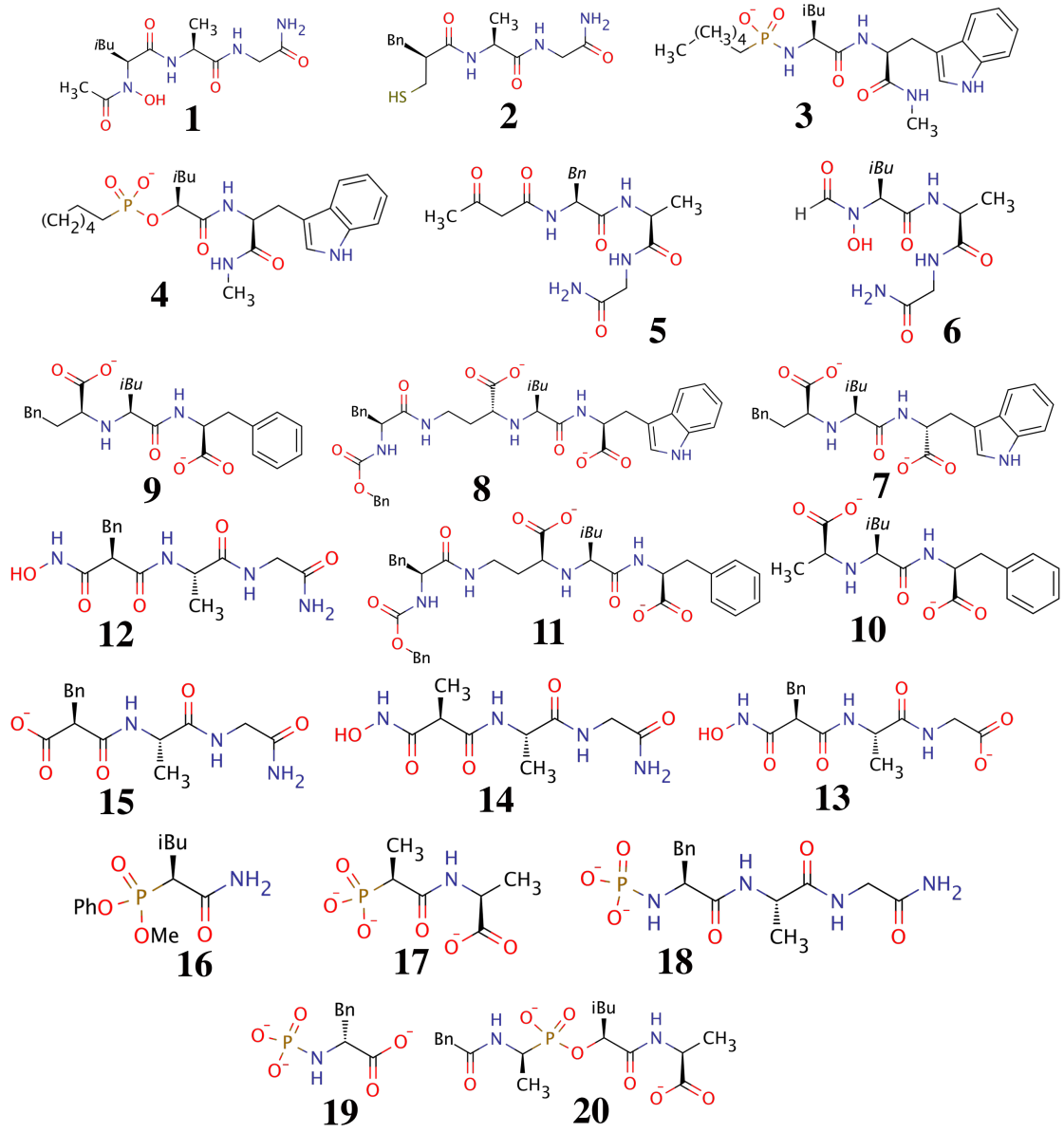


Figure S6b. Chemical structures of thermolysin inhibitors (21-48).

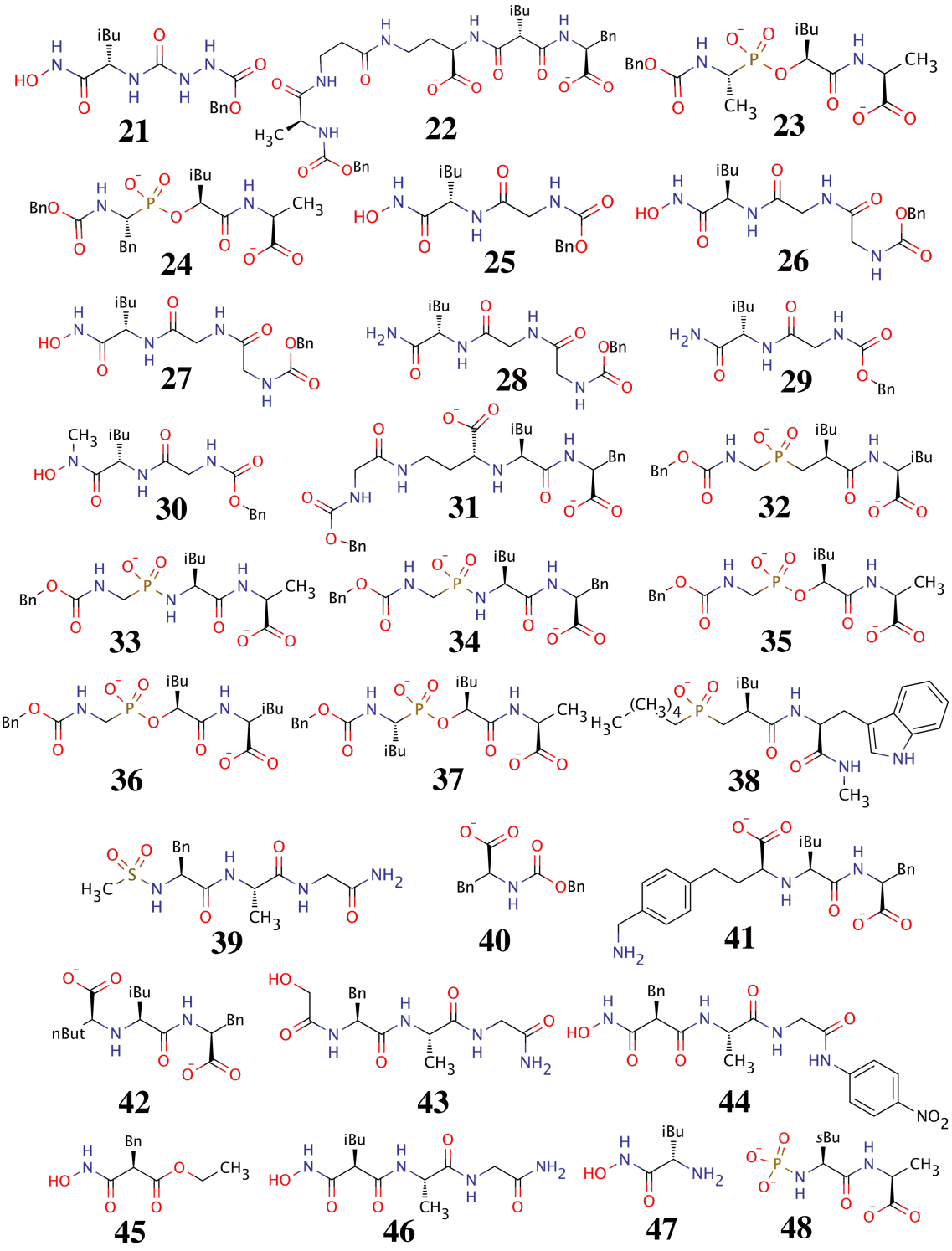
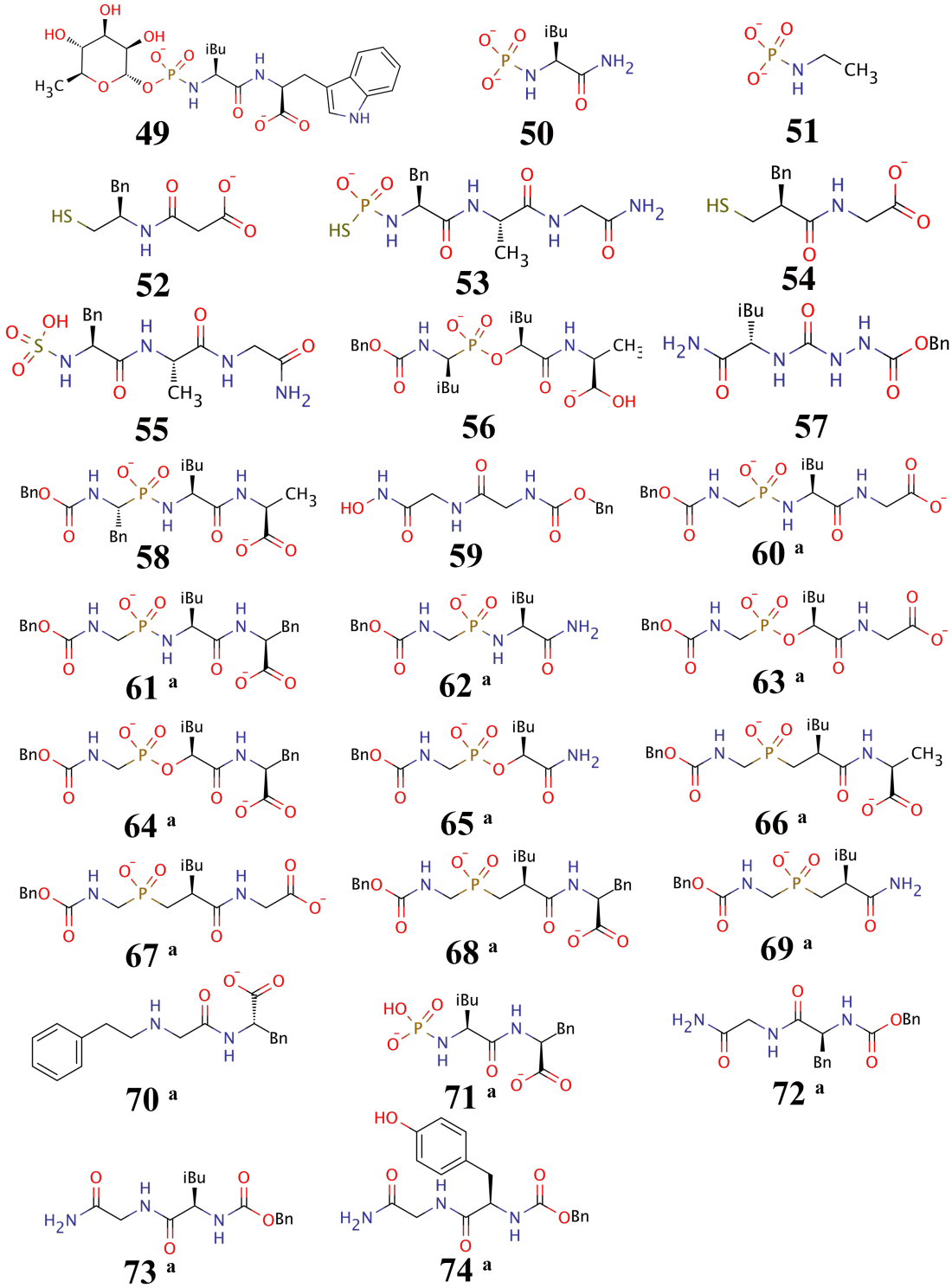


Figure S6c. Chemical structures of thermolysin inhibitors (49-74).



[a] Thermolysin test set compounds.

Figure S7. Histograms of molecular properties (molecular weight, MW; number of hydrogen-bond donors/acceptors, HBD/HBA; clogP; and number of rotatable bonds) for D2/D4 inhibitors.

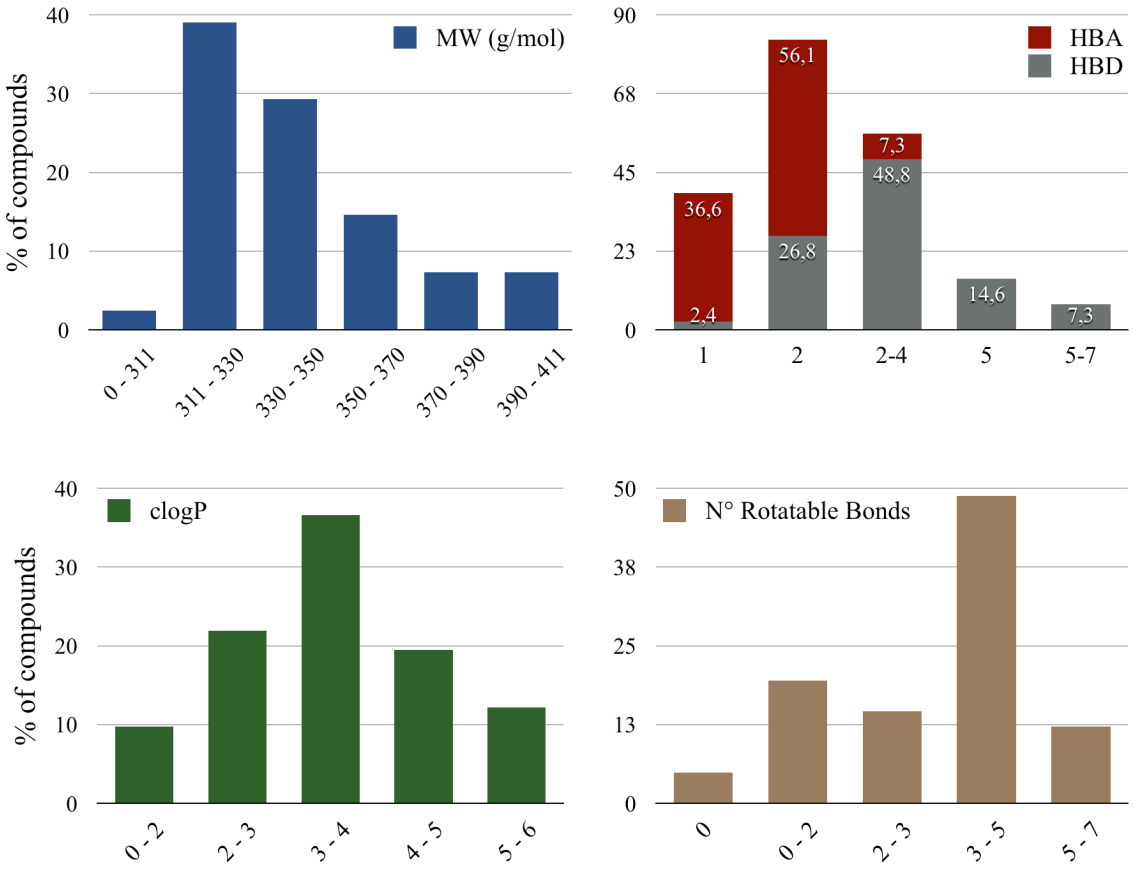


Figure S8. Histograms of molecular properties (molecular weight, MW; number of hydrogen-bond donors/acceptors, HBD/HBA; clogP; and number of rotatable bonds) for 4-aryl-2-chromanones.

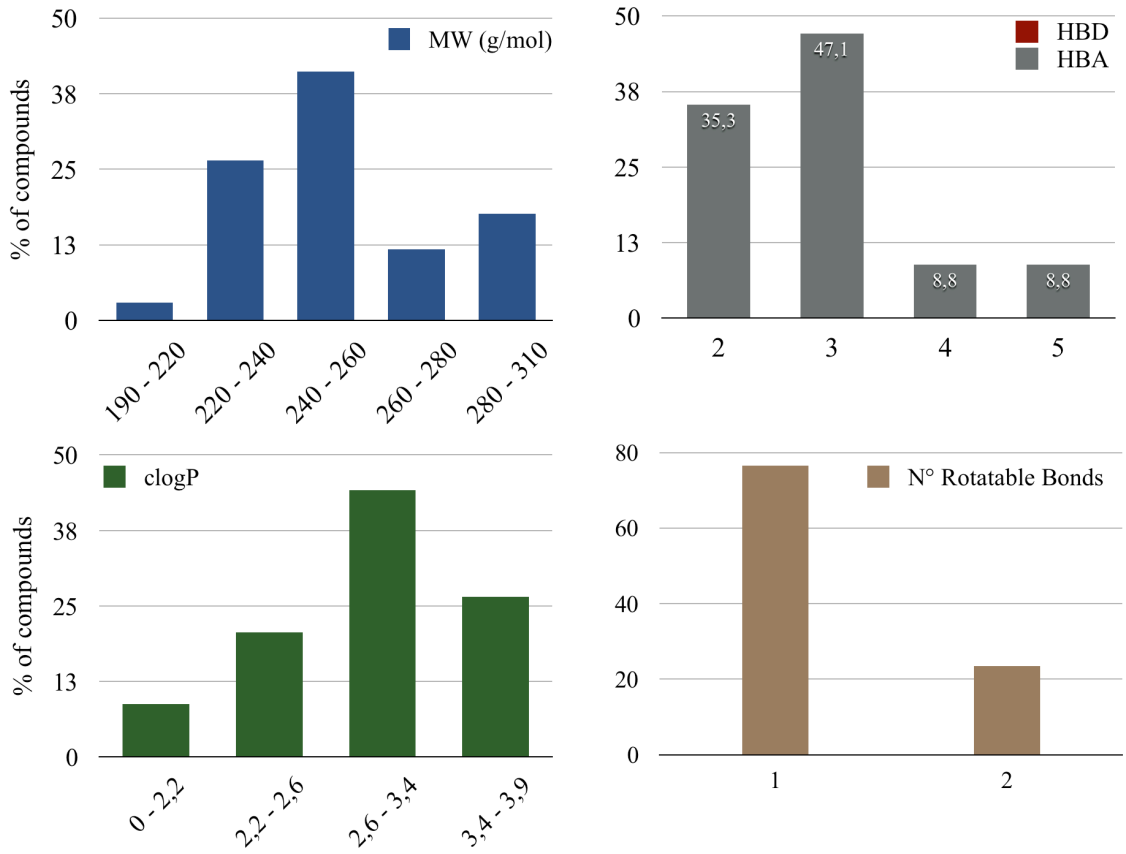


Figure S9. Histograms of molecular properties (molecular weight, MW; number of hydrogen-bond donors/acceptors, HBD/HBA; clogP; and number of rotatable bonds) for GSK-3α inhibitors.

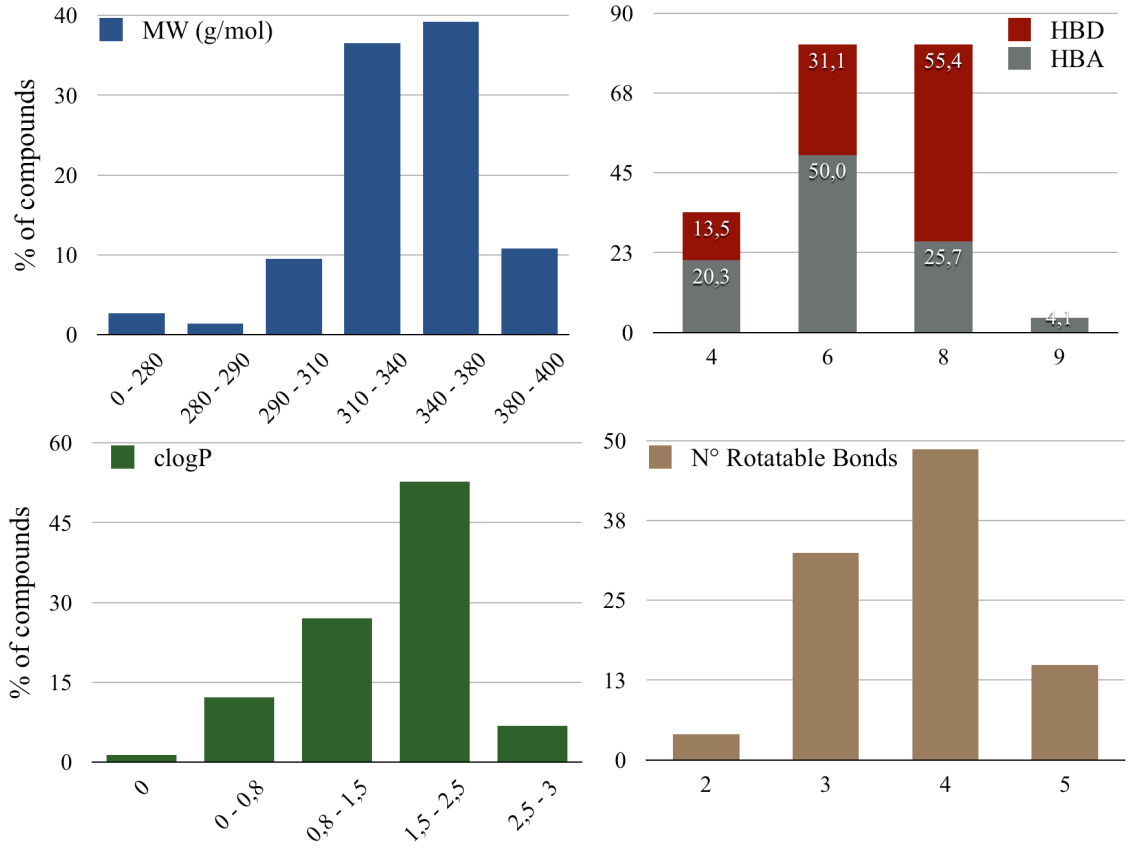


Figure S10. Histograms of molecular properties (molecular weight, MW; number of hydrogen-bond donors/acceptors, HBD/HBA; clogP; and number of rotatable bonds) for cruzain inhibitors.

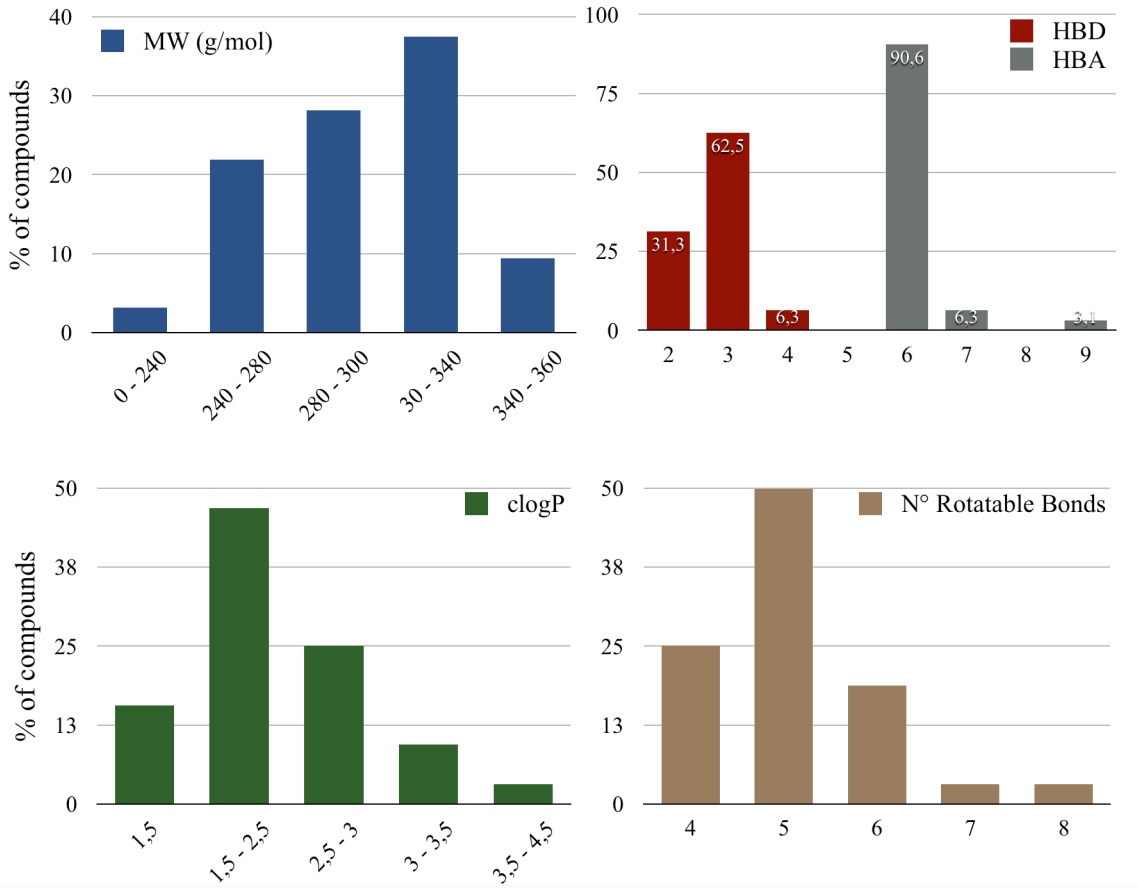


Figure S11. Histograms of molecular properties (molecular weight, MW; number of hydrogen-bond donor/acceptors, HBD/HBA; clogP; and number of rotatable bonds) for thermolysin inhibitors.

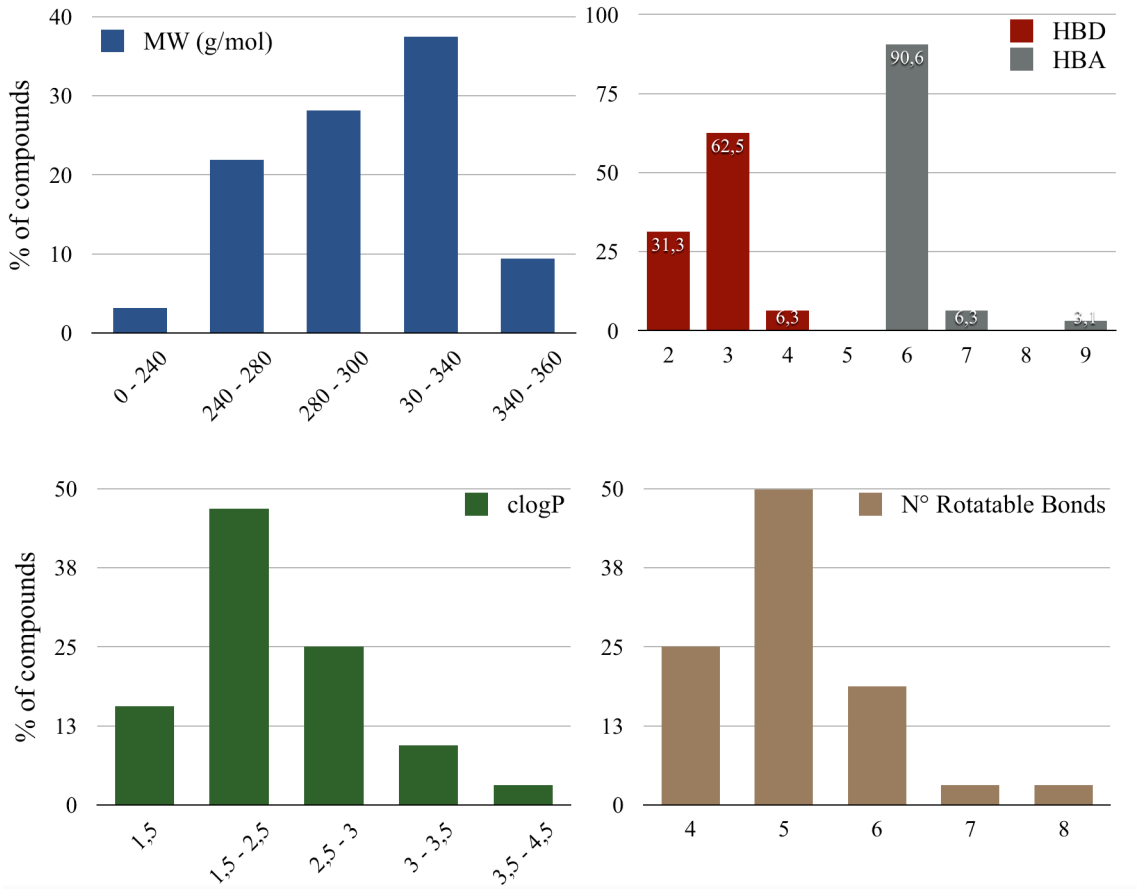


Figure S12. Analysis for the MST-derived hydrophobic descriptors ($logP_{ele}$, $logP_{cav}$ and $logP_{vW}$). For the sake of brevity, comparison is limited to sets of positively charged (D2: A, D), neutral (GSK-3α: B, E) and negatively charged (thermolysin: C, F) systems.

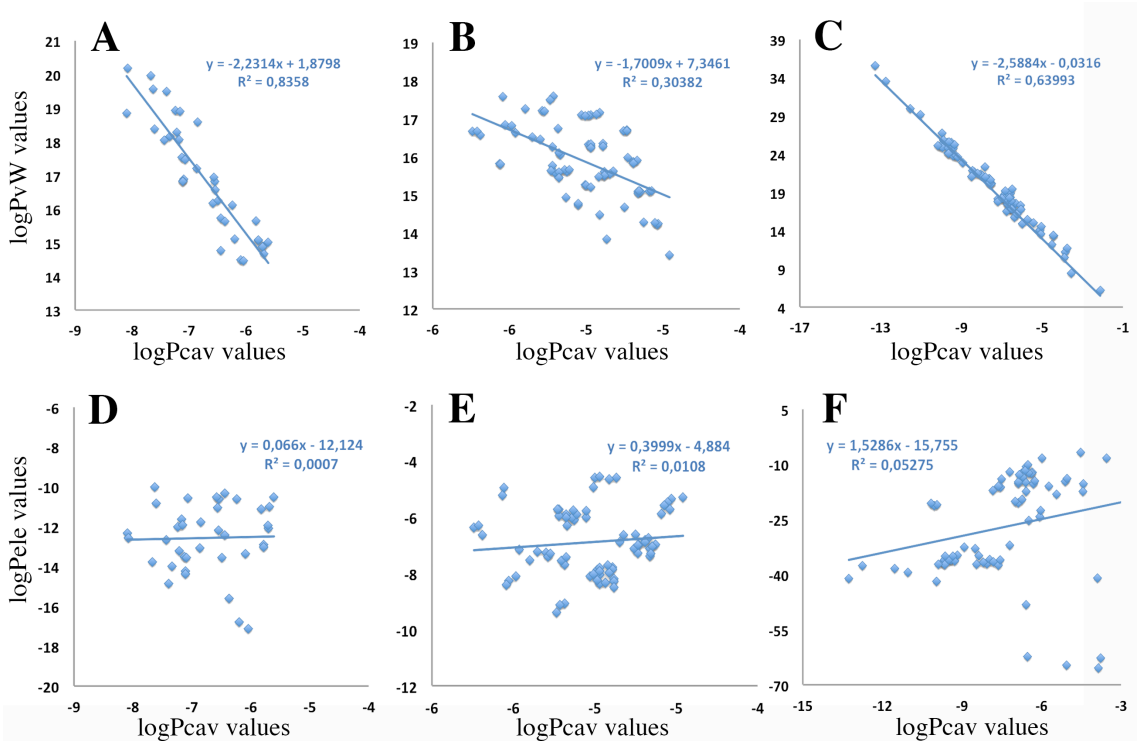


Figure S13. Comparison of the results obtained from standard CoMFA models (data taken from refs. 43-46) and the CoMFA analysis in this study for (A) D2, (B) D4, (C) 2-aryl-4-chromanones, (D) GSK-3a, (E) cruzain and (F) thermolysin systems. Compounds of the training/test set are shown in blue/red, respectively.

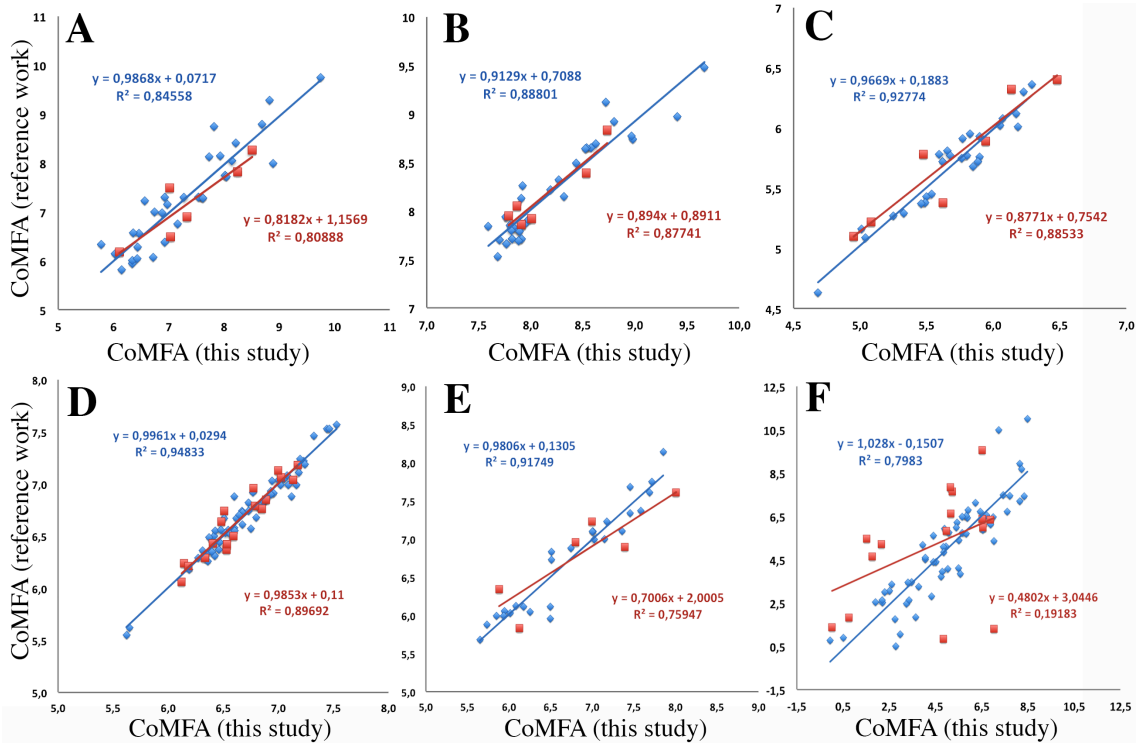


Figure S14. Comparison of the results obtained from standard CoMSIA models (data taken from refs. 43-46) and the MST-based H2 model for (A) D2, (B) D4, (C) 2-aryl-4-chromanones, (D) GSK-3a, (E) cruzain and (F) thermolysin systems. Compounds of the training/test set are shown in blue/red, respectively.

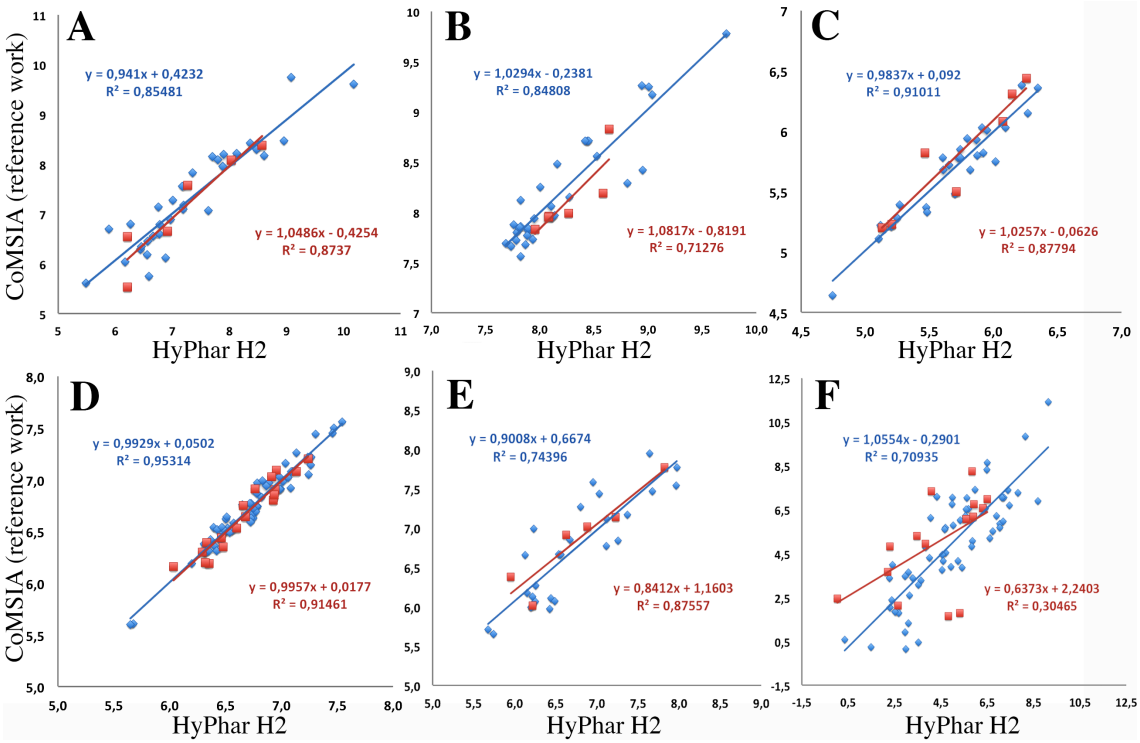


Table S1. Experimental and HyPhar H2 predicted binding constants for Dopamine D2 and D4 training/test set compounds. Values are expressed as pK_i.

	Dopamine D2		Dopamine D4	
Number	Exp	HyPhar H2	Exp	HyPhar H2
1	6.66	6.76	7.68	7.76
2	8.60	8.95	7.46	8.28
3	8.89	7.88	8.19	7.89
4	7.37	8.13	7.10	8.10
5	6.28	6.97	7.68	7.79
6	7.33	7.19	7.80	7.82
7	7.68	7.63	8.31	7.89
8	8.43	7.70	7.89	7.89
9	8.18	7.90	7.96	7.81
10	8.92	8.47	8.41	7.82
11	6.82	7.19	7.64	7.95
12	7.29	7.17	9.00	8.44
13	6.10	6.66	8.66	8.81
14	8.55	8.36	8.48	9.01
15	6.04	5.89	8.44	7.82
16	6.80	6.88	8.52	8.95
17	6.60	6.17	8.28	8.12
18	5.66	6.59	8.07	8.14
19	5.84	5.49	8.22	8.01
20	7.83	7.01	8.96	8.45
21	7.10	7.35	9.05	8.42
22	5.84	6.26	7.59	8.16
23[a]	10.3	9.08	10.3	8.94
24	8.89	8.61	8.75	9.04
25	7.22	6.77	8.00	7.69
26	6.07	6.57	7.66	7.74
27	6.05	6.56	7.64	7.87
28	5.89	6.44	7.28	7.94
29	6.40	7.80	8.25	8.53
30	9.82	10.18	9.38	9.72
31	6.28	6.78	8.26	7.82
32	7.19	6.45	7.72	7.78
33 [a]	5.65	6.22	-	-
34 [a,b]	8.55	8.03	8.96	7.96
35 [a]	7.05	6.91	-	-
36 [a,b]	6.72	7.27	8.12	8.64
37 [a.b]	7.71	8.57	7.94	8.08
38 [a]	5.89	6.21	-	-
39 [b]	-	-	9.37	8.58
40 [b]	-	-	7.22	8.08
41 [b]	-	-	7.36	8.27

[a] Dopamine D2 test set compounds.

[b] Dopamine D4 test set compounds.

Table S2. Experimental and HyPhar H2 predicted inhibitory activities for 4-aryl-2-chromanones training/test set compounds.

	pIC_{50}	
Number	Exp	HyPhar H2
1	5.90	5.92
2	6.29	6.22
3	5.14	5.12
4	5.89	5.80
5	5.73	5.72
8	4.66	4.75
9	5.87	5.87
10	5.44	5.48
12	5.18	5.20
13	6.04	6.10
14	5.84	5.61
15	5.84	5.87
16	5.76	6.02
18	5.09	5.25
19	6.22	6.27
20	6.13	5.95
21	5.87	5.74
22	6.01	5.91
24	5.73	5.82
25	5.85	5.65
26	5.36	5.27
27	5.38	5.48
28	5.76	5.74
29	5.49	5.70
31	5.71	5.61
32	5.10	5.11
33	6.23	6.35
6 [a]	5.83	6.08
7 [a]	5.40	5.71
11 [a]	5.66	5.46
17 [a]	4.98	5.13
23 [a]	6.24	6.26
30 [a]	5.27	5.21
34 [a]	5.86	6.15

[a] Molecules used as external test set.

Table S3. Experimental and HyPhar H2 predicted inhibitory activities for GSK-3α training/test set compounds.

	pIC_{50}			pIC_{50}	
Number	Exp	HyPhar H2	Number	Exp	HyPhar H2
2	6.67	6.39	51	7.55	7.46
4	6.85	6.78	52	7.12	7.26
5	6.29	6.31	53	7.07	7.09
6	6.41	6.36	54	7.59	7.54
7	6.52	6.42	55	6.96	7.01
8	6.71	6.65	56	6.39	6.31
10	6.98	6.89	58	6.96	6.75
11	6.59	6.50	59	6.27	6.47
13	6.35	6.52	60	6.69	6.47
14	6.81	6.51	61	6.82	6.86
15	6.15	6.20	63	6.61	6.50
16	6.43	6.45	64	6.41	6.56
17	6.59	6.57	65	5.58	5.68
18	6.60	6.75	67	5.64	5.65
20	6.33	6.33	68	6.47	6.54
21	6.63	6.79	70	6.88	6.73
23	6.32	6.35	71	6.34	6.40
24	6.91	6.83	72	6.79	6.72
26	6.82	6.78	74	6.16	6.42
28	6.98	7.08	1 [a]	6.28	6.29
29	7.03	6.82	3 [a]	6.67	6.66
30	7.23	7.27	9 [a]	6.94	6.76
31	6.76	6.74	12 [a]	7.15	6.95
33	7.03	7.04	19 [a]	5.83	6.48
34	7.09	7.24	22 [a]	6.39	6.35
35	7.28	7.31	25 [a]	6.50	6.46
37	6.85	6.98	27 [a]	6.86	6.91
38	7.70	7.47	32 [a]	6.83	6.94
39	7.04	6.99	36 [a]	7.24	7.13
40	7.08	7.06	43 [a]	6.87	6.93
41	7.15	7.24	46 [a]	7.10	7.24
42	6.54	6.73	49 [a]	6.85	6.94
44	6.87	6.99	57 [a]	6.79	6.60
45	6.71	6.78	62 [a]	6.28	6.33
47	6.73	6.72	66 [a]	5.85	6.03
48	6.67	6.77	69 [a]	6.73	6.68
50	7.13	7.13	73 [a]	5.85	6.32

[a] Molecules used as external test set.

Table S4. Experimental and HyPhar H2 predicted inhibitory activities for cruzain training/test set compounds.

	pIC_{50}	
Number	Exp	HyPhar H2
1	6.15	6.44
2	6.15	6.49
3	6.10	6.25
4	6.05	6.42
6	6.05	6.22
7	6.05	6.15
8	6.05	6.20
9	5.90	6.25
11	5.80	5.68
12	7.20	7.12
14	7.10	6.67
15	7.00	7.11
16	6.90	6.54
17	6.90	6.56
18	5.95	6.13
19	5.65	5.74
20	7.20	6.80
22	7.60	6.95
23	8.00	7.64
25	7.74	7.97
26	7.74	7.67
27	7.40	7.96
28	7.30	7.03
29	7.20	7.26
30	7.15	7.37
32	6.50	6.23
5 [a]	6.05	5.95
10 [a]	5.85	6.21
13 [a]	7.10	6.62
21 [a]	6.40	6.88
24 [a]	7.89	7.82
31 [a]	6.60	7.23

[a] Molecules used as external test set.

Table S5. Experimental and HyPhar H2 predicted inhibitory activities for Thermolysis inhibitors training/test set compounds.

		pKi				*pKi*	
N	Name	Exp	HyPhar H2	N	Name	Exp	HyPhar H2
1	ace_ohleu_agnh2	2.47	2.38	38	c6pcltnme	7.28	7.39
2	bzsag	6.12	5.00	39	ch3o2s_fagnh2	0.52	2.99
3	c6pltnme	8.82	8.14	40	cbzphe	3.29	2.41
4	c6poltnme	5.84	6.61	41	dah54	5.77	5.33
5	ch3coch2co_fagnh2	2.51	2.29	42	dah55	2.42	2.96
6	cho_ohleu_agnh2	2.47	2.68	43	hoch2co_fagnh2	2.54	2.52
7	cltzncrys	7.47	7.82	44	nhohbzmagna	6.37	7.18
8	dah50	7.96	8.69	45	nhohbzmoet	4.70	4.56
9	dah51	6.22	5.69	46	nhohibmagnh2	6.32	4.32
10	dah52	5.55	5.62	47	nhohleu	3.72	3.10
11	dah53	6.66	7.04	48	p_ile_aoh	6.44	6.89
12	nhohbzmagnh2	6.18	4.97	49	phosphoramidon	7.55	7.13
13	nhohbzmagoh	6.18	6.73	50	pleunh2	4.10	2.96
14	nhohmalagnh2	2.96	3.15	51	pnhet	0.52	0.35
15	ohbzmagnh2	3.38	3.30	52	r_thiorphan	5.64	4.66
16	p_ophe_ome_leunh2	0.52	1.49	53	s02p_fagnh2	5.16	5.85
17	paaoh	4.06	4.01	54	s_thiorphan	5.74	5.56
18	po3_fagnh2	5.59	5.58	55	so3_fagnh2	2.37	3.52
19	ppheoh	4.14	5.42	56	z_d_lpola	4.38	5.83
20	z_d_apola	4.62	4.60	57	z_nh_glnh2	3.42	3.52
21	z_nh_glnhoh	5.57	4.06	58	zfplazncrys	10.17	9.13
22	zala	6.07	6.42	59	zggnhoh	3.03	3.64
23	zapola	5.74	7.03	60	zgplg [a]	6.57	5.58
24	zfpola	7.35	7.45	61	zgplf [a]	7.12	5.89
25	zg_d_lnhoh	4.32	4.71	62	zgplnh2 [a]	6.12	3.83
26	zgg_d_lnhoh	3.60	4.54	63	zgpolg [a]	3.64	4.82
27	zgglnhoh	4.41	3.85	64	zgpolf [a]	4.27	5.32
28	zglnh2	1.68	3.09	65	zgpolnh2 [a]	3.18	0.03
29	zglnhoh	4.89	4.70	66	zgpcla [a]	7.73	5.93
30	zglnmeoh	2.65	2.30	67	zgpclg [a]	6.52	6.31
31	zgly	6.39	5.98	68	zgpclf [a]	7.18	6.50
32	zgpcllzncrys	6.74	7.21	69	zgpclnh2 [a]	5.85	4.08
33	zgpla	7.78	6.48	70	ppphe [a]	2.79	2.65
34	zgpllzncrys	8.04	6.50	71	plfoh [a]	7.72	5.84
35	zgpola	4.89	4.93	72	zfgnh2 [a]	3.46	2.30
36	zgpollzncrys	5.05	5.23	73	zlgnh2 [a]	2.51	2.20
37	zlpola	6.17	5.03	74	zygnh2 [a]	3.66	3.47

[a] Molecules used as external test set.

Table S6. Levels of isocontours (expressed as the PLS coeficients corrected by the standard deviation) used in the pharmacophoric maps shown in Figures 4, 6, 8, 10 and 11. All values were scaled by a factor of 10^5.

System	Field	Level
GSK-3a (Figure 4)		
CoMFA	Electrostatic	+3 / -3
	Steric	+225 / -225
HyPhar H2	logPele	+17 / -17
	logPcav	+6 / -6
Cruzain (Figure 6)		
CoMFA	Electrostatic	+8 / -8
	Steric	+165 / -165
HyPhar H2	logPele	+1.3 / -1.3
	logPcav	+8 / -8
Thermolysin (Figure 8)		
CoMFA	Electrostatic	+30 / -70
	Steric	+2000 / -1500
HyPhar H2	logPele	+300 / -300
	logPcav	+30 / -18
D2 (Figure 10)		
CoMFA	Electrostatic	+2.5 / -2.5
	Steric	+53 / -53
HyPhar H2	logPele	+1.8 / -1.8
	logPcav	+2.5 / -2.5
D4 (Figure 10)		
CoMFA	Electrostatic	+2.5 / -2.5
	Steric	+65 / -65
HyPhar H2	logPele	+6 / -6
	logPcav	+0.7 / -0.7
2-aryl-4-chromanones (Figure 11)		
CoMFA	Electrostatic	+1.5 / -1.5

	Steric	+450 / -450
HyPhar H2	logPele	+4 / -4
	logPcav	+7 / -7

Paper 2

Application of the Quantum Mechanical IEF/PCM-MST hydrophobic descriptors to selectivity in ligand binding

Application of the Quantum Mechanical IEF/PCM-MST Hydrophobic Descriptors to Selectivity in Ligand Binding

Tiziana Ginex,[a] Jordi Muñoz-Muriedas,[b] Enric Herrero,[c] Enric Gibert,[c] Pietro Cozzini,[a]* and F. Javier Luque[d]*

[a] Dipartimento di Scienze degli Alimenti, University of Parma, Parco Area delle Scienze 59/A, 43121 Parma, Italy

[b] GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, United Kingdom

[c] Pharmacelera, Jordi Girona 1-3, Campus Nord Universitat Politècnica de Catalunya, Edifici K2M, 08034 Barcelona, Spain

[d] Department of Chemical Physics and Institut de Biomedicina (IBUB), Faculty of Pharmacy, University of Barcelona, Av. Prat de la Riba 171, 08921 Santa Coloma de Gramenet, Spain

* E-mail: pietro.cozzini@unipr.it (PC) or fjluque@ub.edu (FJL)

Abstract

We have recently reported the development and validation of quantum mechanical (QM)-based hydrophobic descriptors derived from the parametrized IEF/PCM-MST continuum solvation model for 3D-QSAR studies within the framework of the Hydrophobic Pharmacophore (HyPhar) method. In this study we explore the applicability of these descriptors to the analysis of selectivity fields. To this end, we have examined a series of 88 compounds with inhibitory activities against thrombin, trypsin and factor Xa, and the Hyphar results have been compared with 3D-QSAR models reported in the literature. The quantitative models obtained by combining the electrostatic and non-electrostatic components of the octanol/water partition coefficient yield results that compare well with the predictive potential of standard CoMFA and CoMSIA techniques. The results also highlight the potential of Hyphar descriptors to discriminate the selectivity of the compounds against thrombin, trypsin, and factor Xa. Moreover, the graphical representation of the hydrophobic maps provides a direct linkage with the pattern of interactions found in crystallographic structures. Overall, the results support the usefulness of the QM/MST-based hydrophobic descriptors as a complementary approach for disclosing structure-activity relationships in drug design and for gaining insight into the molecular determinants of ligand selectivity.

Keywords: target selectivity · hydrophobic molecular field · continuum solvation model · 3D-QSAR

Introduction

The binding affinity between a therapeutic drug and its macromolecular target is the most relevant property in drug design [1]. Depending on the pharmacological assay, it is typically estimated from the half-maximal inhibitory concentration (IC_{50}) or the inhibition (K_i) or dissociation (K_D) constants through *in vitro* assays where the concentration of both target and drug are precisely controlled. However, the binding affinity must be carefully balanced with physico-chemical features that influence properties such as bioavailability, toxicity and efficacy, which are generically encoded under the term 'drug-likeness' [2-5]. Therapeutic drugs must also fulfill selectivity criteria in order to minimize off-target interactions, which can trigger undesired side effects, to limit the biological effect to a given target within a family of homologue proteins (i.e., a receptor subtype or an enzyme isoform), or to confer narrow coverage against several targets of interest (i.e., multiple pathways in a signalling cascade) [6-8]. Examples of the relevance of selectivity in drug design are the development of ATP-competitive inhibitors in kinases [9-11], the induction of the 'cheese reaction' by non-selective monoamine oxidase inhibitors [12], the role of non steroidal anti-inflammatory drugs specific for COX-2 on the control of pain and inflammation with limited gastrointestinal toxicity [13], and the avoidance of side-effects in the design of inhibitors against the large family of phosphodiesterases [14].

Attaining the goal of target selectivity is more challenging than improving the binding affinity. Thus, medicinal chemists must identify the unique structural features among the targets that determine the differences in pharmacological responses, and then translate these differences into suitable modifications of the chemical scaffold of ligands. To this end, several strategies have been proposed [7, 8] and implemented in distinct computational frameworks [15-21]. In the context of ligand-based 3D-QSAR

methods [22-24], models of drug selectivity have been examined using Comparative Molecular Field Analysis (CoMFA [25]) and Comparative Molecular Similarity Indices Analysis (CoMSIA [26]) techniques. In CoMFA Lennard-Jones (L-J) and Coulomb potentials are mapped onto regularly spaced grid points surrounding the mutually aligned molecules. Then, Partial Least Squares (PLS) is used to identify regions in steric and electrostatic fields that explain the differences in biological activity of ligands. The predictive model can be visualized through appropriate isocontour maps, thus facilitating the interpretation of the hidden relationships between chemical structure and biological activity. In turn, CoMSIA relies on similarity indices determined using a gaussian-type functional form, which leads to smoother distance dependence, avoid singularities and provides improved contour maps. Furthermore, CoMSIA was formulated extending the number of fields, including electrostatic, steric, hydrophobic, and hydrogen-bond donor and acceptor properties [27, 28].

CoMFA and CoMSIA have been successfully used to study ligand selectivity. For instance, Böhm *et al*. utilized CoMSIA to examine the binding affinity of inhibitors of thrombin, trypsin and factor Xa, and to discuss the origin of ligand selectivity between thrombin and trypsin from the differences in binding affinity, which were used as the dependent variable in the analysis [29]. Baskin *et al*. addressed the problem of selectivity between NMDA and AMPA receptors by using the difference in biological activity of the receptor subtypes for building CoMFA models of ligand selectivity [30]. Walline *et al*. utilized CoMFA to examine the differences in antagonist potency of antidepressants against the wild-type and mutated forms of the human serotonin transporter (hSERT) [31]. To this end, the antagonist activity against the wild-type hSERT was substracted from the activity against the mutated species in order to disclose the role of point mutations in the transporter. In a distinct approach, Sharma *et*

al. utilized the concept of pairwise binding affinity to enhance the potency of triazolo-[1,5-a]-quinoxaline compounds for both AMPA and KA receptors [32].

Very recently we have examined the suitability of descriptors of molecular hydrophobicity derived from quantum mechanical (QM) solvation calculations for their use in 3D-QSAR studies [33]. This work was motivated by several reasons. Studies of target druggability highlighted the relevance of shape and hydrophobicity in drug binding [34-37]. In particular, ligand desolvation was recognized to be largely responsable of the variation in maximal achievable binding free energy for a drug-like molecule [38]. On the other hand, polar interactions in druggable binding sites are crucial for both binding and selectivity [39-41]. Hence, it is reasonable to expect that the analysis of the 3D pattern of hydrophobicity/hydrophilicity of ligands could be valuable to identify key features of ligand recognition at druggable pockets. In fact, this assumption was already considered in previous works focused on the implementation of the Molecular Lipophilicity Potential [42] and Hydropathic Interactions (HINT) [43, 44] methods, which used empirical contributions to the molecular hydrophobicity. The development of refined versions of QM-Self Consistent Reaction Field (QM-SCRF) codes provide an alternative approach to evaluate the solvation free energy in a variety of solvents, and specifically the partition coefficient between water and octanol [45-48]. Thus, QM-SCRF methods can be useful to provide novel descriptors relevant for understanding ligand recognition and binding. Finally, a novel set of hydrophobicity-related descriptors can be directly compared with experimental measurements of octanol/water partition coefficients, which are widely adopted in drug design studies.

This work pursues to explore the suitability of the QM-SCRF hydrophobic descriptors to study selectivity fields among multiple targets. The computational procedure relies on the Hydrophobic Pharmacophore (HyPhar), which was formulated using a rigorous

partitioning scheme of the solvation/transfer free energy into fragment contributions in the IEF/PCM-MST continuum model (see Methods [33]). As a test case, we use these descriptors to examine the selectivity of a series of serine protease inhibitors against thrombin, trypsin and factor Xa. The suitability of the QM-SCRF hydrophobic fields is calibrated through comparison with the results reported in the literature. The results support the use of the QM-SCRF hydrophobic descriptors to assist the design of selective ligands.

Methods

Molecular system

The suitability of the IEF/PCM-MST-based atomic hydrophobicity (HyPhar) descriptors has been tested for a series of 88 derivatives acting as inhibitors of trypsin, thrombin, and factor Xa [29] (Figure 1). Although the compounds share the 3-amidinophenylalanine unit, the substituents (R^1, R^2) present in this congeneric series confer a significant chemical diversity. Thus, most of the compounds have a formal positive charge, but few are neutral or have two positive charges. Moreover, the analysis of selected properties (molecular weight, clogP, number of hydrogen-bond donors/acceptors, and number of rotatable bonds; calculated using the DataWarrior program [49]) reveals that the compounds cover a wide range of values (see Figure S1 in Supporting Material), as noted for instance in the range of clogP values (from -0.5 to 6) and the number of rotatabe bonds (generally from 4 to 11).

The inhibitory activity, expressed as pK_i, differs by 4.1, 4.7, and 3.0 units for thrombin, trypsin, and factor Xa, respectively (Tables S1-S3 in Supporting Material). For the sake of comparison, we have adopted the same partition between training and test compounds adopted in the reference work [29]. For the set of training compounds (**1**-

72), the activity varied from 4.357 (**72**) to 8.377 (**1**) for thrombin, from 3.000 (**72**) to 7.699 (**11**) for trypsin, and from 3.000 (**66**) to 6.046 (**54**) for factor Xa.

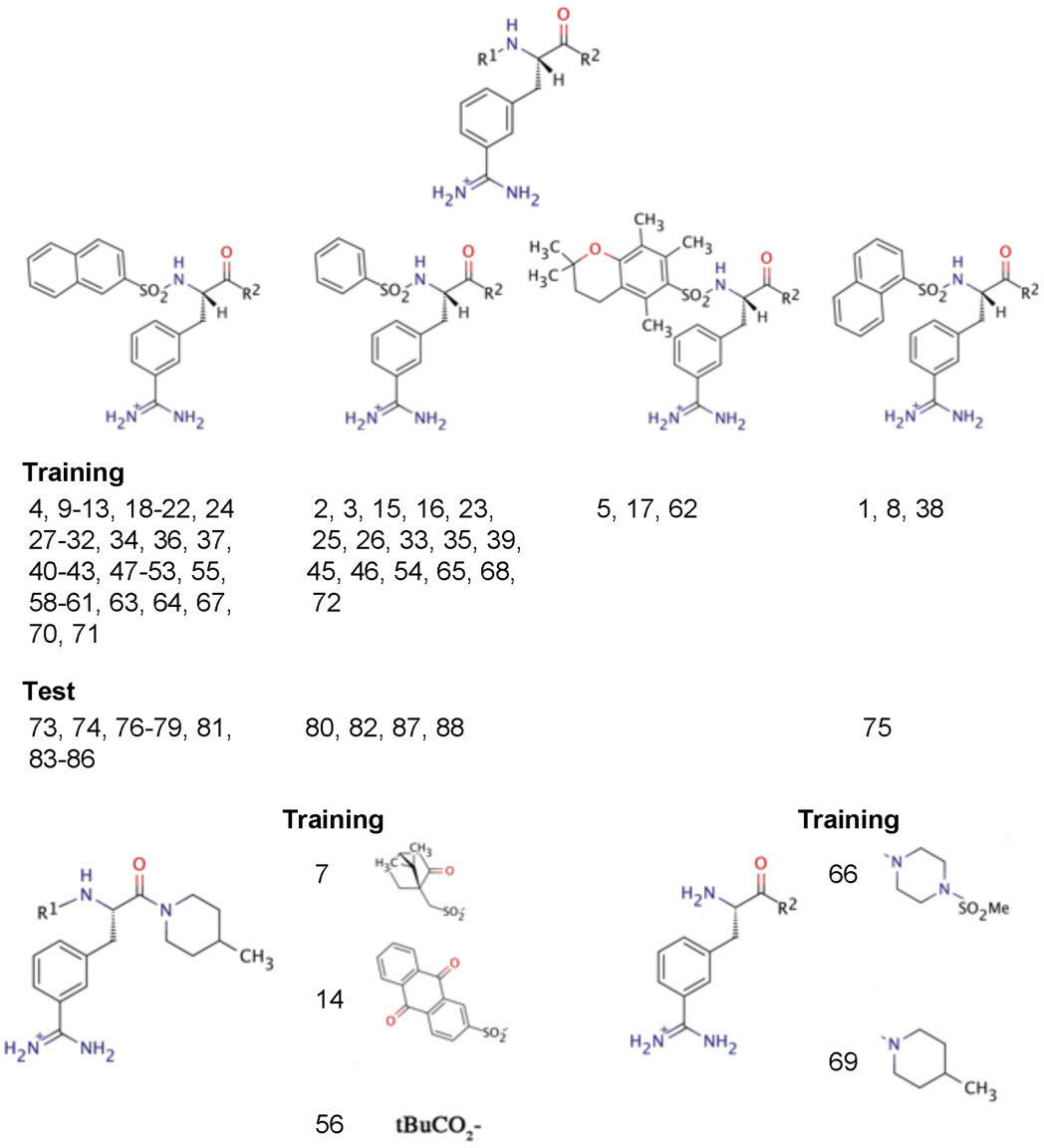


Figure 1. Schematic representation of the chemical scaffold in the series of 88 inhibitors (training: 1-72; test: 73-88) and distribution of the compounds into the six major structural families (for a complete representation see ref. 29).

With regard to the 16 compounds included in the test set (**73-88**), the range of pK_i values was comparable for thrombin (from 4.745 (**88**) to 8.481 (**73**)), and slightly

smaller for trypsin (from 4.337 (**88**) to 7.638 (**86**)) and factor Xa (from 4.284 (**80**) to 5.509 (**74**)).

The availability of structural data for the three targets, in conjunction with previous target-based studies on these systems [50-53], have identified the differences in the protein environment of the binding sites, as illustrated from the superposition of the binding pockets for thrombin, trypsin and factor Xa in PDB structures 1ETS, 1PPH, and 1HCG (Figure 2; see also Figure S2 in Supporting Material).

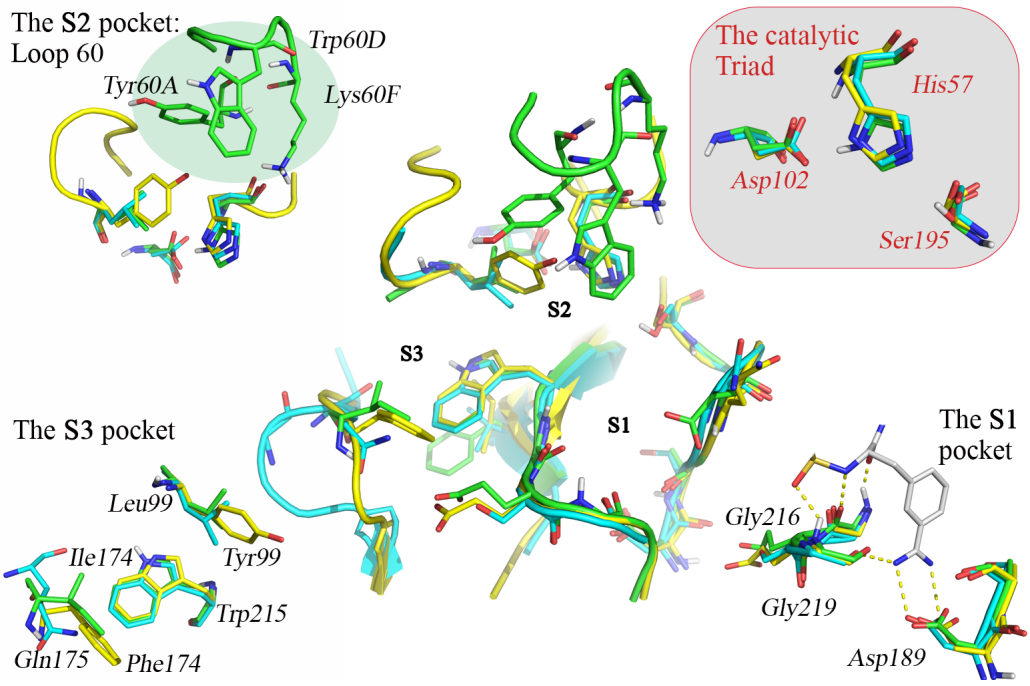


Figure 2. Structural superposition of the binding sites in thrombin (PDB ID 1ETS, green), trypsin (PDB ID 1PPH, light blue) and factor Xa (PDB ID 1HCG, yellow). The central cavity filled by the inhibitors contains three pockets (S1-S3). The benzamidine moiety binds at pocket S1 (shown as grey sticks).

Three main pockets can be identified. The first one (S1) allows the accommodation of the benzamidine moiety of the ligands through polar interactions with Asp189 and Gly219 at the bottom of the site, and with Gly216 at the rim of the S1 site. In the upper part of the cavity, the S2 pocket (also denoted P site) is delimited by aromatic and charged residues pertaining to the Loop60 (Tyr60A, Trp60D and Lys60F) and by the

catalytic triad (His57, Asp102 and Ser195). Among the three receptors, only thrombin presents the insertion of the Loop60, thus making the binding site more sterically constrained and less solvent-exposed with regard to the other two proteins, but more prone to accommodate hydrophobic groups. Compared to thrombin and trypsin, the S3 pocket (also known as D site) of factor Xa is delimited by aromatic aminoacids (Phe174, Trp215 and Tyr99), whereas in the other enzymes only two residues (Trp215, Leu99) are of apolar nature. Comparison with other human structures shows that the major structural features of the binding pocket are well preserved, and that only small changes are induced upon binding by other inhibitors, mainly found at the outer rim of the S3 site.

Hyphar descriptors

The hydrophobicity of a molecule is typically determined from the partitioning between octanol and water ($P_{o/w}$), which in turn is related to the free energy of transfer ($\Delta G_{o/w}$) of a given solute between these two solvents (Eq. 1).

$$\log P_{o/w} = -\frac{\Delta G_{o/w}}{2.303RT} = -\frac{\Delta G_w - \Delta G_o}{2.303RT} \quad (1)$$

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

Though the $\log P_{o/w}$ provides a direct measure of the overall hydrophobicity of a molecule, a fractional decomposition into atomic contributions is necessary in order to evaluate the hydrophobic complementarity between a given molecule and its biological

target. In this regard, the 3D hydrophobicity pattern of a molecule can be defined from the atomic contributions to $\log P_{o/w}$ taking advantage of the decomposition scheme [54, 55] formulated for the solvation free energy within the MST version of the IEF-PCM solvation model [46, 56, 57].

In the MST method the solvation free energy (ΔG_{sol}) is calculated by adding three contributions (Eq. 2). The first one is the cavitation term (ΔG_{cav}), which is the work required for creating a cavity shaped to accommodate the solute in the solvent. The second component is the van der Waals term (ΔG_{vW}), which accounts for dispersion-repulsion between solute and solvent. Finally, the third component is the electrostatic term (ΔG_{ele}), which measures the work needed to build up the solute charge distribution in the solvent.

$$\Delta G_{sol} = \Delta G_{ele} + \Delta G_{cav} + \Delta G_{vW} \tag{2}$$

With regard to ΔG_{cav} and ΔG_{vW}, their linear dependence on the solvent-exposed surface of each atom provides a straightforward way to estimate the atomic contribution ($\Delta G_{cav,i}$), as noted in Eqs. 3 and 4.

$$\Delta G_{cav,i} = \frac{S_i}{S_T} \Delta G_{P,i} \tag{3}$$

where $\Delta G_{P,i}$ is the cavitation free energy of atom *i* determined using Pierotti's formalism,[41] whose contribution is weighted by the contribution of the solvent-exposed surface (S_i) of atom *i* to the total surface (S_T).

$$\Delta G_{vW,i} = \xi_i S_i \tag{4}$$

where ξ_i denotes the atomic surface tension of atom *i*, which is determined by fitting the experimental free energy of solvation.

Finally, the electrostatic term is derived from a perturbative treatment of the interaction between the whole wavefunction in the gas phase (Ψ^o) and the set of imaginary charges (q_j^{sol}), spread on the solute cavity, that represent the solvent reaction field generated by the presence of the solute (Eq. 5) [55].

$$\Delta G_{ele,i} = \sum_{\substack{j=1 \\ j \in i}}^{M} \left\langle \Psi^o \left| \frac{1}{2} \frac{q_j^{sol}}{|r_j - r_i|} \right| \Psi^o \right\rangle \tag{5}$$

where M is the total number of reaction field charges distributed onto the solvent-exposed surface of atom *i*.

From Eqs 3-5, the atomic contribution to the molecular hydrophobicity can be determined as follows.

$$\log P_{o/w} = \sum_{i=1}^{N} \log P_i = \sum_{i=1}^{N} \left(\log P_{ele,i} + \log P_{cav,i} + \log P_{vW,i} \right) \quad (7)$$

The partitioning of $\log P_{o/w}$ into electrostatic and non-electrostatic terms allows us to explore the relationships with the biological activity using different combinations of descriptors. Since both $\log P_{cav}$ and $\log P_{vW}$ depend on the solute-exposed surface of atoms, they encode the size and shape of the molecule, and can be treated as mimics of the steric field. On the other hand, since $\log P_{ele}$ depends on the solvent reaction field induced from the solute charge distribution, it should encode information related to electrostatic features of molecules. Accordingly, for our purposes we have utilized four combinations of descriptors: i) $\log P_{ele}$ and the cube of the atom radii, which was taken as a simple measure of the atomic size, ii) $\log P_{ele}$ and $\log P_{cav}$, iii) $\log P_{ele}$ and $\log P_{vW}$, and, finally iv) $\log P_{ele}$ and the total non-electrostatic term, $\log P_{n-ele}$, which was obtained as the addition of $\log P_{cav}$ and $\log P_{vW}$. In the following Hyphar models obtained from the preceding combinations of descriptors will be denoted H1-H4, respectively.

Model generation and statistical analysis

HyPhar models were derived using the in-house PharmQSAR software [58]. To this end, solvation calculation in water and octanol were performed using the B3LYP/6-31G(d) version of the IEF/PCM-MST solvation model [57]. The molecular geometries of the inhibitors were retrieved from the literature [59] and used in calculations without further reoptimization of the molecular aligment. This was intended to calibrate the performance of the Hyphar descriptors by direct comparison with the CoMFA and

CoMSIA results reported in the original work [29]. All calculations were performed with Gaussian 09 [60].

The set of aligned molecules were enclosed in a lattice (2.0 Å grid spacing) with boundaries chosen to allow a minimun of 3 Å extension from the molecules. The total number of points in the grid was 2184. The atomic hydrophobicities were projected into the grid using the similarity index function (A^q) implemented in CoMSIA (Eq. 8).

$$A^q = \sum_{i=1}^{N} w_{probe} w_i e^{-\alpha r_{iq}^2} \quad (8)$$

where w_i is the actual value of the atomic hydrophobicity of atom *i*, w_{probe} is the hydrophobicity of the probe atom, which is taken as +1, α is the attenuation factor, which was set to 0.3, and r_{iq} is the distance between the probe atom at grid point *q* and atom *i* of the test molecule.

PharmQSAR uses PLS to extract the hidden relationships between biological data and the hydrophobic field using an algorithm based on NIPALS [61]. To this end, the projected field was stored as an *M*x*Ng* matrix (*M* is the number of molecules and *Ng* denotes the number of grid points). The field values were then corrected by the mean value and normalized to unit variance, and columns with a standard deviation lower than a certain threshold (typically 0.1) were removed. The optimum number of components in the PLS analysis was selected on the basis of the leave-one-out cross-validation, and checked with the lowest standard deviation error in prediction of the actual experimental values corrected by the number of degrees of freedom of the model,

S_{PRESS}, and the ability of the model to predict the biological activity of the test compounds.

For the sake of comparison, a CoMFA model was also derived by combining the partial charges obtained from B3LYP calculations in the gas phase and the cube of the atom radii (taken from TRIPOS force field [62]) to account for electrostatic and steric fields. Compared with the CoMFA analysis reported in [29], the only difference concerns the partial charges, which were derived at the semiempirical AM1 level.

Results and Discussion

Analysis of binding affinities. Before exploring the relationships between hydrophobic descriptors and selectivity, we have first examined their ability to predict the binding affinity for thrombin, trypsin, and factor Xa. To this end, Table 1 reports the statistical results obtained from Hyphar (H1-H4) models, and the CoMFA results obtained in this work. For the sake of comparison, CoMFA and CoMSIA results reported in [29] are also provided. On the other hand, the HyPhar H2 predicted activities for the three enzymes are reported in Supporting Information (Tables S1-S3).

The HyPhar models generally lead to results statistically comparable to CoMFA, especially when one considers models H2 ($\log P_{ele}$ and $\log P_{cav}$) and H3 ($\log P_{ele}$ and $\log P_{vW}$), which show similar performances, as expected from the dependence of $\log P_{cav}$ and $\log P_{vW}$ on the solvent-exposed surface of atoms. Indeed, there is a high correlation between the values of these properties among the set of compounds (see Figure S3 in Supporting Information). Hyphar H2/H3 and CoMFA models lead to q^2 values of 0.60-0.65 and 0.61 for thrombin and trypsin, respectively, including a similar number of components in the final model (4 for thrombin and 5-6 for trypsin). With regard to factor Xa, models H2/H3 render a larger q^2 (0.49-0.50) compared to CoMFA

(0.41), though including a larger number of components in the model (6 in H2 and H3 *versus* 2 in CoMFA). Moreover, the weigth of $\log P_{ele}$ in Hyphar H2/H3 models is larger than the contribution of the electrostatic term in CoMFA, especially for factor Xa.

Table 1. Summary of statistical parameters obtained for HyPhar (H1-H4) models derived from the QM MST-based hydrophobic descriptors for the compounds in the training set.

	Ref. 29 [a]	CoMFA	H1	H2	H3	H4
Thrombin						
r^2	0.88/0.95	0.88	0.88	0.88	0.86	0.75
S	0.37/0.24	0.35	0.35	0.35	0.37	0.50
q^2	0.69/0.76	0.60	0.62	0.65	0.62	0.55
S_{press}	0.59/0.53	0.67	0.65	0.63	0.66	0.71
Nc [b]	4/6	4	4	4	4	5
Fields (%) [c]						
	E 38/15	E 20	$\log P_{ele}$ 27	$\log P_{ele}$ 36	$\log P_{ele}$ 29	$\log P_{ele}$ 46
	S 62/21	S 80	R^3 73	$\log P_{cav}$ 64	$\log P_{vW}$ 71	$\log P_{n\text{-}ele}$ 54
	H --/30					
	D --/9					
	A --/25					
Trypsin						
r^2	0.92/0.97	0.96	0.85	0.94	0.92	0.85
S	0.26/0.16	0.20	0.40	0.25	0.30	0.39
q^2	0.63/0.75	0.61	0.49	0.61	0.61	0.61
S_{press}	0.56/0.45	0.57	0.65	0.57	0.57	0.58
Nc	5/9	6	5	6	5	8
Fields (%)						
	E 34/16	E 17	$\log P_{ele}$ 11	$\log P_{ele}$ 36	$\log P_{ele}$ 33	$\log P_{ele}$ 52
	S 66/17	S 83	R^3 89	$\log P_{cav}$ 64	$\log P_{vW}$ 67	$\log P_{n\text{-}ele}$ 48
	H --/29					
	D --/10					
	A --/28					
Factor Xa						
r^2	0.68/0.91	0.63	0.88	0.91	0.90	0.85
S	0.37/0.19	0.62	0.35	0.30	0.32	0.39
q^2	0.37/0.59	0.41	0.41	0.49	0.50	0.48
S_{press}	0.52/0.42	0.49	0.51	0.47	0.47	0.47
Nc	3/6	2	6	6	6	6
Field (%)						
	E 30/16	E 24	$\log P_{ele}$ 27	$\log P_{ele}$ 41	$\log P_{ele}$ 42	$\log P_{ele}$ 36
	S 70/17	S 76	R^3 73	$\log P_{cav}$ 59	$\log P_{vW}$ 58	$\log P_{n\text{-}ele}$ 64
	H --/35					
	D --/9					
	A --/23					

[a] CoMFA (left) and CoMSIA (right) results taken from [29]. [b] Number of principal components. [c] Fraction of the field (in percentage). E: electrostatic; S: steric; H: hydrophobic; D: H-bond donor; A: H-bond acceptor; R^3: cube of the atomic radius.

Compared to models H2/H3, the performance of model H1 ($\log P_{ele}$ and R^3) is worst for trypsin (q^2 of 0.49), but compares well for thrombin and factor Xa, whereas model H4 ($\log P_{ele}$ and $\log P_{n-ele}$) provides a slightly worst description for thrombin (q^2 of 0.55).

The models H2/H3 are also in overall agreement with the CoMFA results reported by Böhm and coworkers [29] for thrombin (r^2 and q^2 of 0.88 and 0.69 in [29] *versus* 0.86-0.88 and 0.62-0.65 for HyPhar) and trypsin (r^2 and q^2 of 0.92 and 0.63 in [29] *versus* 0.92-0.94 and 0.61 for Hyphar).

Table 2. Summary of statistical parameters obtained for HyPhar (H1-H4) models derived from the QM MST-based hydrophobic descriptors for the compounds in the test set.

	Ref. 29 [a]	**CoMFA**	**H1**	**H2**	**H3**	**H4**
Thrombin [b]						
r^2	0.48/0.51	0.39	0.21	0.37	0.38	0.33
	0.58/0.69	*0.56*	*0.63*	*0.61*	*0.59*	*0.45*
S_{PRESS}	0.70/0.68	0.69	0.79	0.70	0.69	0.72
	0.65/0.55	*0.61*	*0.55*	*0.57*	*0.58*	*0.67*
RMSE [c]	0.66/0.63	0.64	0.73	0.65	0.64	0.67
	0.60/0.51	*0.56*	*0.51*	*0.53*	*0.54*	*0.60*
Trypsin						
r^2	0.65/0.84	0.68	0.72	0.72	0.73	0.85
S_{PRESS}	0.54/0.37	0.53	0.49	0.49	0.49	0.36
RMSE	0.51/0.34	0.49	0.46	0.46	0.45	0.34
Factor Xa						
r^2	0.42/0.46	0.30	0.45	0.40	0.40	0.43
S_{PRESS}	0.28/0.27	0.32	0.28	0.29	0.30	0.29
RMSE	0.26/0.25	0.29	0.26	0.27	0.28	0.27

[a] Calculated from the data reported in [29].[b] Parameters obtained after exclusion of compound **84** are given in italics.[c] Root-mean square error.

With regard to factor Xa, models H2/ H3 lead to improved statistics (r^2 and q^2 of 0.68 and 0.37 for CoMFA *versus* 0.90-0.91 and 0.49-0.50 for HyPhar models).

The predictive ability of the Hyphar models can be examined from the data reported in Table 2 and Figure 3. As pointed out in [29], we found that compound **84** deviates significantly from the general predictive behaviour in case of thrombin. The residual for

compound **84** reaches 2 logarithmic units only for thrombin (Figure 3D), whereas in the other two cases it deviates less than 1.0 logarithmic unit (Figure 3E,F). Indeed, this compound can be considered to be an outlier according to Grubb's test ($P < 0.05$) and a highly influential point according to Cook's distance estimate [58]. Accordingly, Table 2 also reports the statistical data after its exclusion, which improves significantly the predictive ability for the activity against thrombin. For models H2/H3, values of r^2 close to 0.60, 0.73 and 0.40 are obtained for thrombin, trypsin, and factor Xa, which compare well with the values obtained from CoMFA calculations for thrombin and trypsin, and improve the CoMFA prediction for factor Xa. They also compare well with the CoMFA statistical analysis reported in [29].

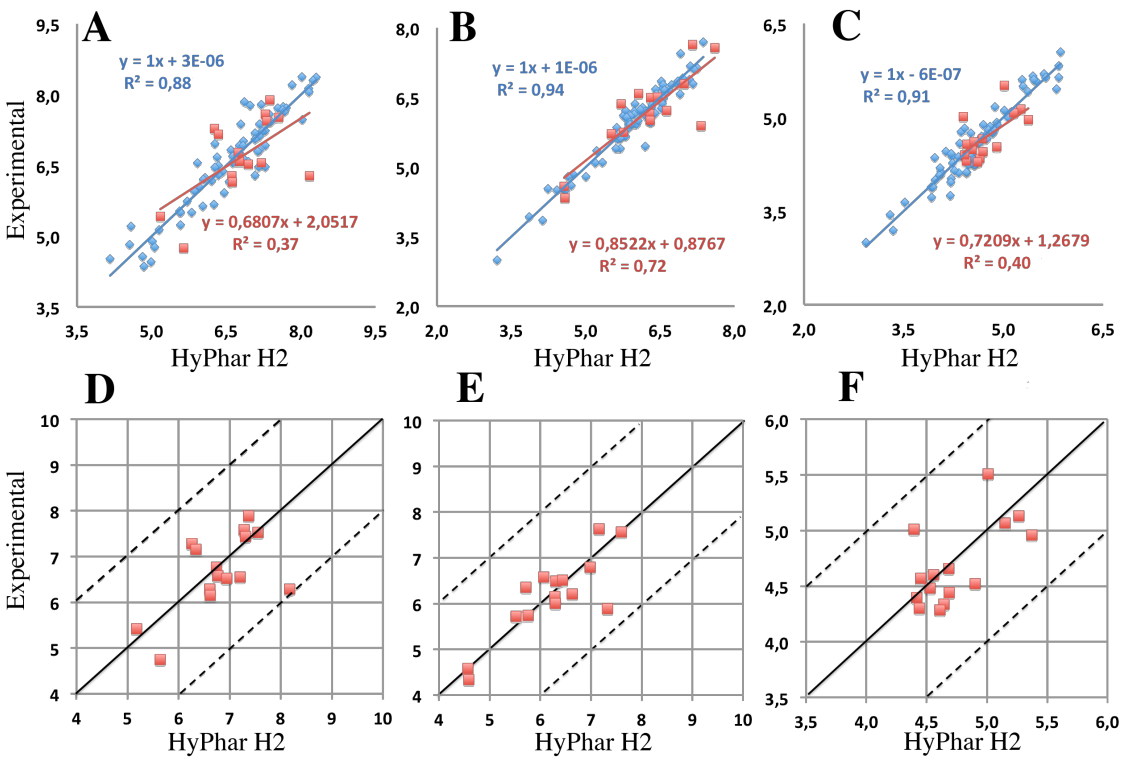


Figure 3. Comparison of the experimental data and the results predicted from the HyPhar H2 model for thrombin (**A, D**), trypsin (**B, E**), and factor Xa (**C, F**). The black dashes lines mark deviations of 2.0 (Factor Xa: 1.0) logarithmic units from the optimal prediction. Compounds of the training/test sets are shown in blue/red, respectively.

Pharmacophoric maps of targets. A 3D-QSAR model should provide an easily interpretable graphical representation of physico-chemical properties relevant for the biological activity. For the sake of brevity, we limit our discussion to the isocontour maps obtained from the HyPhar H2 model (Figure 4).

For thrombin the $\log P_{ele}$ isocontour maps reveal two distinct areas. The negative (red) isocontour would favor the presence of polar ($\log P_{ele}$ <0) groups, filling an area delimited by the catalytic residues (Ser195, His57) and the Lys60F residue inserted in Loop60 (see Figure S2 in Supporting Information). In fact, this latter residue would stabilize the presence of groups with negative charge, which was suggested to be a determinant for thrombin recognition [53]. The positive contour (blue), which would favour a reduction in polarity, is associated to the polar groups of inhibitors implicated in contacts with the main chain groups of Gly216 and Gly219. This can be interpreted from the need to keep the balance between the proper positioning for formation of favorable interactions with the target and the dehydration cost of the highly polar groups (i.e., amido, sulfonyl) present in the inhibitor. With regard to the $\log P_{cav}$ component, bulkier groups, corresponding to the area filled by the piperidine group (green isocontour), are favoured within the pocket defined by Loop60 (S2 site). This finding agrees with previous studies based on the structural analysis of the three targets [50]. The absence of bulky substituents leads to weaker inhibitory potencies (i.e., compounds **62-64**, pK_i of 5.509-5.208; **70**, pK_i of 4.456). The additional green isocontour around the anthraquinone ring points out that extension toward the S3 pocket is also favoured. On the other hand, the areas in yellow could be associated to steric clashes with His57, and with Tyr60A and Trp60D of Loop60.

The $\log P_{ele}$ and $\log P_{cav}$ maps for trypsin possess features that resemble the isocontours obtained for thrombin, as expected from the structural similarity between these two systems (Figure S2). Nevertheless, subtle differences can be found, such as the absence of the yellow area (when plotted at the same isocontour level), which can be related to the lack of the steric constraint imposed by the Loop60. Furthermore, a red isocontour around one of the NH_2 groups in the amidino moiety would highlight the relevance of this interaction for the affinity, suggesting that an enhancement in polarity would reinforce the hydrogen-bond formed with the carbonyl unit of Gly219. Since this interaction may be formed in the three targets, this trend might reflect the effect arising from Ser190 in assisting the proper positioning of the amidino group of the inhibitors, whereas such residue is replaced by Ala in thrombin and factor Xa, a feature also highlighted in previous studies [53]. Compared to thrombin, the isocontour maps of factor Xa exhibit large differences. Thus, unlike thrombin and trypsin, the role of steric factors on the S2 pocket is notably diminished, though small yellow and green areas (plotted at the same isocontour level in Figure 4) can be identified. The replacement of Leu99 in thrombin and trypsin by Tyr in factor Xa can explain the notable change in the pattern of $\log P_{cav}$ maps (Figure S2). On the other hand, a small green isocontour near the anthraquinone ring reflects the favorable extension toward the S3 site, as was also identified in previous works [50, 53].

Overall, these results highlight the existence of distinct signatures in the pharmacophore maps of the three enzymes, revealing the existence of few structural differences valuable to confer selectivity to the inhibitors.

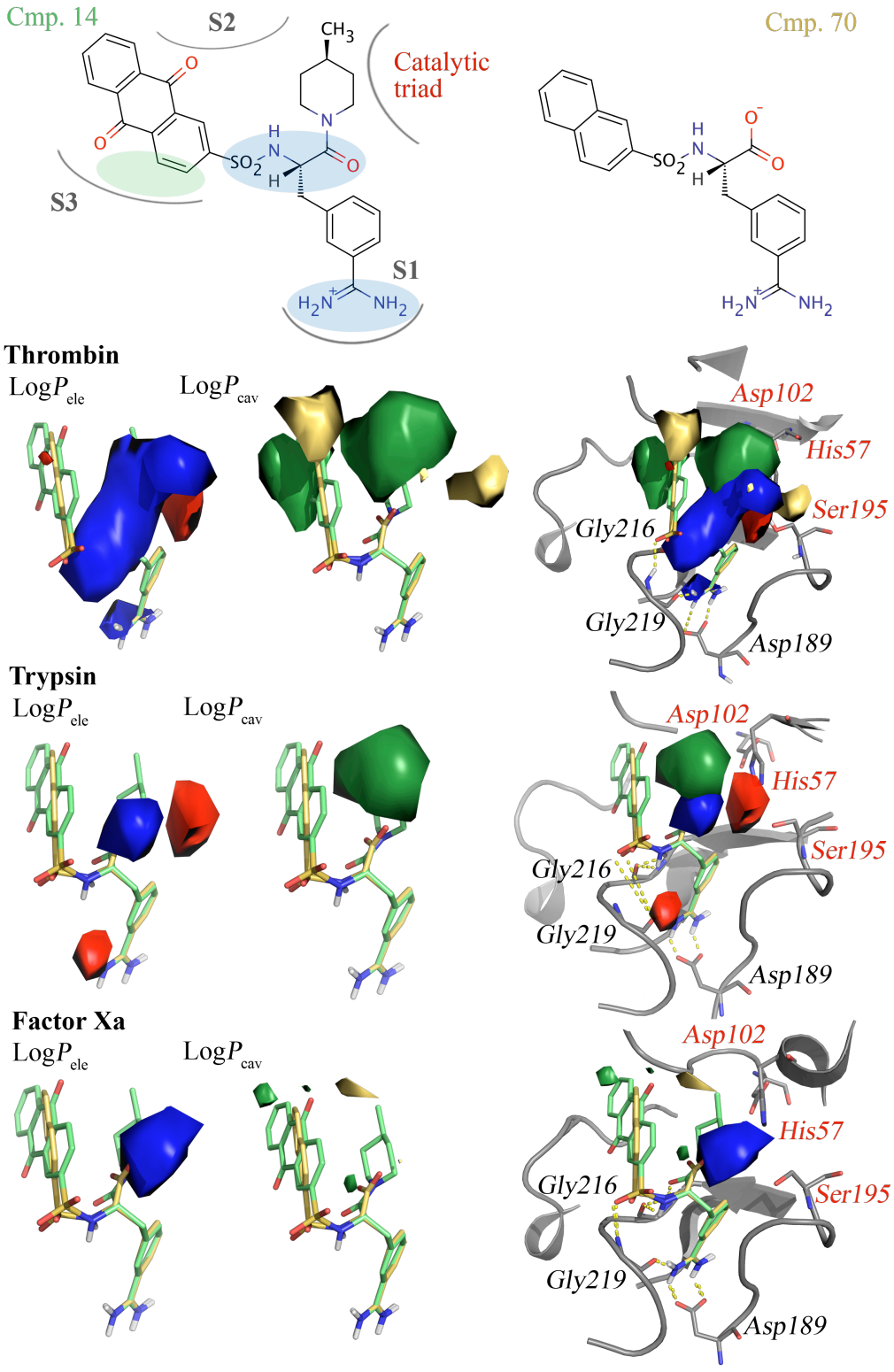


Figure 4. Representation of HyPhar H2 isocontour maps for thrombin, trypsin, and factor Xa. Isocontours are expressed as the PLS coefficients corrected by the standard deviation and scaled by a factor of 10^5. The electrostatic field ($\log P_{ele}$) is shown as blue (positive) and red (negative) isocontours (values of 150 and -50, respectively), which correspond to areas where polarity decreases/increases the biological activity. The non-electrostatic field ($\log P_{cav}$) is shown as yellow/green isocontours (values of

+35 and -35, respectively) and denote areas where steric bulk is unfavourable/favourable. The ligand-receptor complementary is shown for compounds **14** (green) and **70** (yellow), which are in the range of potent and weak inhibitors. For compound **14**, the areas in blue/green highlight relevant topological (polar/apolar) pharmacophore elements.

Analysis of selectivity models. As pointed out by Böhm *et al*. in their original work [29], the activity data for thrombin and trypsin exhibit the highest correlation among the three pairwise systems (0.72 *versus* 0.28 and 0.46). Thus, the analysis of selectivity between thrombin and trypsin seems more challenging than the comparison against factor Xa. In order to build up the selectivity models, 3D-QSAR models were derived using the differences in affinity of the ligands for the three pairwise systems (thrombin/trypsin, thrombin/factor Xa, and trypsin/factor Xa).

Table 3. Summary of statistical parameters obtained for pairwise selectivity models derived from HyPhar descriptors for the compounds in the training set.

	H1	**H2**	**H3**	**H4**
Thrombin/Trypsin				
r^2	0.82	0.80	0.83	0.83
S	0.43	0.45	0.42	0.42
q^2	0.50	0.51	0.53	0.54
S_{press}	0.52	0.51	0.51	0.51
Nc	4	4	4	5
Field (%)				
	$logP_{ele}$ 31	$logP_{ele}$ 28	$logP_{ele}$ 28	$logP_{ele}$ 32
	R^3 69	$logP_{cav}$ 72	$logP_{vW}$ 72	$logP_{n\text{-}ele}$ 68
Thrombin/Factor Xa				
r^2	0.90	0.91	0.89	0.85
S	0.32	0.31	0.34	0.39
q^2	0.61	0.60	0.56	0.49
S_{press}	0.67	0.67	0.71	0.77
Nc	5	5	5	5
Field (%)				
	$logP_{ele}$ 19	$logP_{ele}$ 22	$logP_{ele}$ 22	$logP_{ele}$ 30
	R^3 81	$logP_{cav}$ 78	$logP_{vW}$ 78	$logP_{n\text{-}ele}$ 70
Trypsin/Factor Xa				
r^2	0.89	0.87	0.86	0.90
S	0.34	0.36	0.37	0.33
q^2	0.54	0.59	0.54	0.50
S_{press}	0.57	0.54	0.57	0.61
Nc	5	4	4	7
Field (%)				
	$logP_{ele}$ 20	$logP_{ele}$ 23	$logP_{ele}$ 22	$logP_{ele}$ 32
	R^3 80	$logP_{cav}$ 77	$logP_{vW}$ 78	$logP_{n\text{-}ele}$ 68

The results are reported in Table 3, and the HyPhar H2 predicted differences in activity are reported in Tables S4-S6 in Supporting Information.

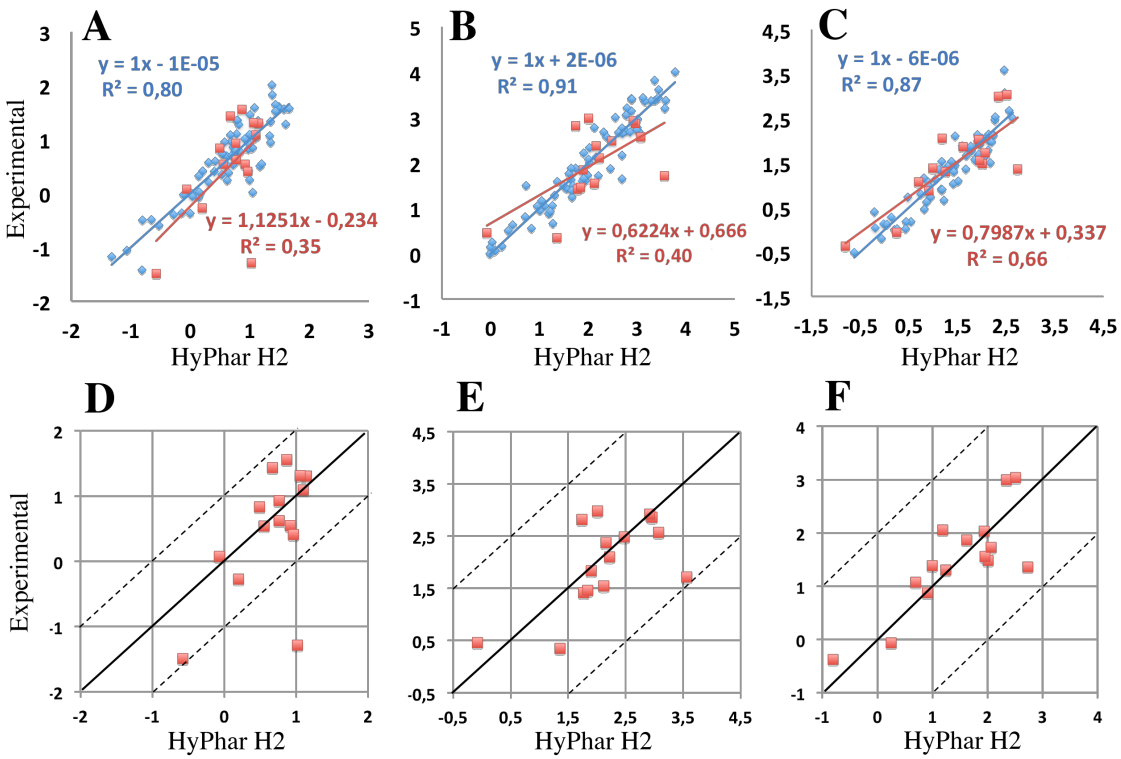


Figure 5. Comparison of the experimental data and the results predicted from the HyPhar H2 model for pairwise selectivity models of thrombin/trypsin (A, D), thrombin/factor Xa (B, E), and trypsin/factor Xa (C, F) selectivity models. The dashes lines in black mark deviations of 1.0 (factor Xa: 2.0) logarithmic unit from the optimal prediction. Compounds of the training/test sets are shown in blue/red respectively.

The selectivity models toward factor Xa perform slightly better that for the thrombin/trypsin pair, as expected from the larger correlation between the activity data for these two systems. Thus, q^2 values for models H2/H3 range from 0.56-0.60 for thombin/factor Xa to 0.54-0.59 for trypsin/factor Xa, and to 0.51-0.53 for thrombin/trypsin. For this latter system, the results compare with those reported from CoMSIA analysis in ref. 29, which yielded a q^2 value of 0.57 for a model with four principal components and five molecular fields. In models H2/H3 the non-electrostatic

component is consistently more relevant than the electrostatic term, especially for the comparison with factor Xa, as noted in a ratio of $\log P_{ele}$ and $\log P_{cav}$ fields close to 1:4. The prospective potential of the selectivity models can be calibrated from the results given in Table 4 and Figure 5, which compares the experimental and predicted differences in activity obtained from Hyphar H2 model.

Table 4. Summary of statistical parameters obtained for pairwise selectivity models derived from HyPhar descriptors for the compounds in the training set.

	H1	**H2**	**H3**	**H4**
Thrombin/Trypsin [a]				
r^2	0.22	0.35	0.32	0.39
	0.74	*0.73*	*0.66*	*0.60*
S_{PRESS}	0.84	0.77	0.79	0.74
	0.43	*0.44*	*0.49*	*0.53*
RMSE	0.79	0.72	0.73	0.69
	0.40	*0.41*	*0.45*	*0.49*
Thrombin/Factor Xa				
r^2	0.24	0.40	0.30	0.33
	0.50	*0.58*	*0.53*	*0.47*
S_{PRESS}	0.77	0.68	0.73	0.72
	0.64	*0.59*	*0.54*	*0.66*
RMSE	0.71	0.63	0.68	0.67
	0.59	*0.54*	*0.48*	0.61
Trypsin/Factor Xa				
r^2	0.74	0.66	0.73	0.65
S_{PRESS}	0.49	0.56	0.50	0.57
RMSE	0.46	0.52	0.46	0.53

[a] Parameters obtained after exclusion of compound 84 are given in italics.

As noted above, compound **84** behaves as an outlier according to Grubb's test ($P < 0.05$), and its exclusion improves significantly the statistical results for the thrombin/trypsin and thrombin/factor Xa systems (note also that the residual of this compound exceeds the 1.0 logarithmic unit only in the case of thrombin; see Figure 5). The best results are found for the thrombin/trypsin and trypsin/factor Xa models, with r^2 values in the range 0.66-0.73 and for models H2 and H3, while the results obtained for the thrombin/factor Xa pair are mildly poorer (r^2 of 0.53-0.58).

Selectivity maps. The HyPhar H2 isocontour maps for pairwise selectivity features are reported in Figures 6-8. For the thrombin/trypsin pair (Figure 6), the isocontours for $\log P_{ele}$ are shown in light blue and orange, which denote regions where an increase/decrease in polarity would favour/disfavour the selectivity against thrombin. In particular, enhancing the polarity in the light blue isocontour would increase selectivity toward thrombin, a trend that may be ascribed to the electrostatic stabilization with Glu192, whereas a reduction in polarity is better tolerated in trypsin due to the replacement of Glu192 by Gln (see Figure S2 in Supporting Information). On the other hand, a reduction in polarity for the orange areas would be detrimental for thrombin, hence more beneficial for trypsin inhibition, since this would destabilize the interaction with Lys60F, and would facilitate the formation of contacts with the rim of the S1 site less dependent on the correct positioning of the inhibitor subject to the constraints imposed by Loop60.

With regard to the non-electrostatic terms, the isocontours in light pink denote areas that favour selectivity toward thrombin. Thus, they primarily correspond to areas that would favour van der Waals interactions with residues that shape the S2 site, specifically the space delimited by the side chains of His57, Tyr60A and Lys60F, favorable contacts with the side chain of Trp60D, and with residues in the S3 site.

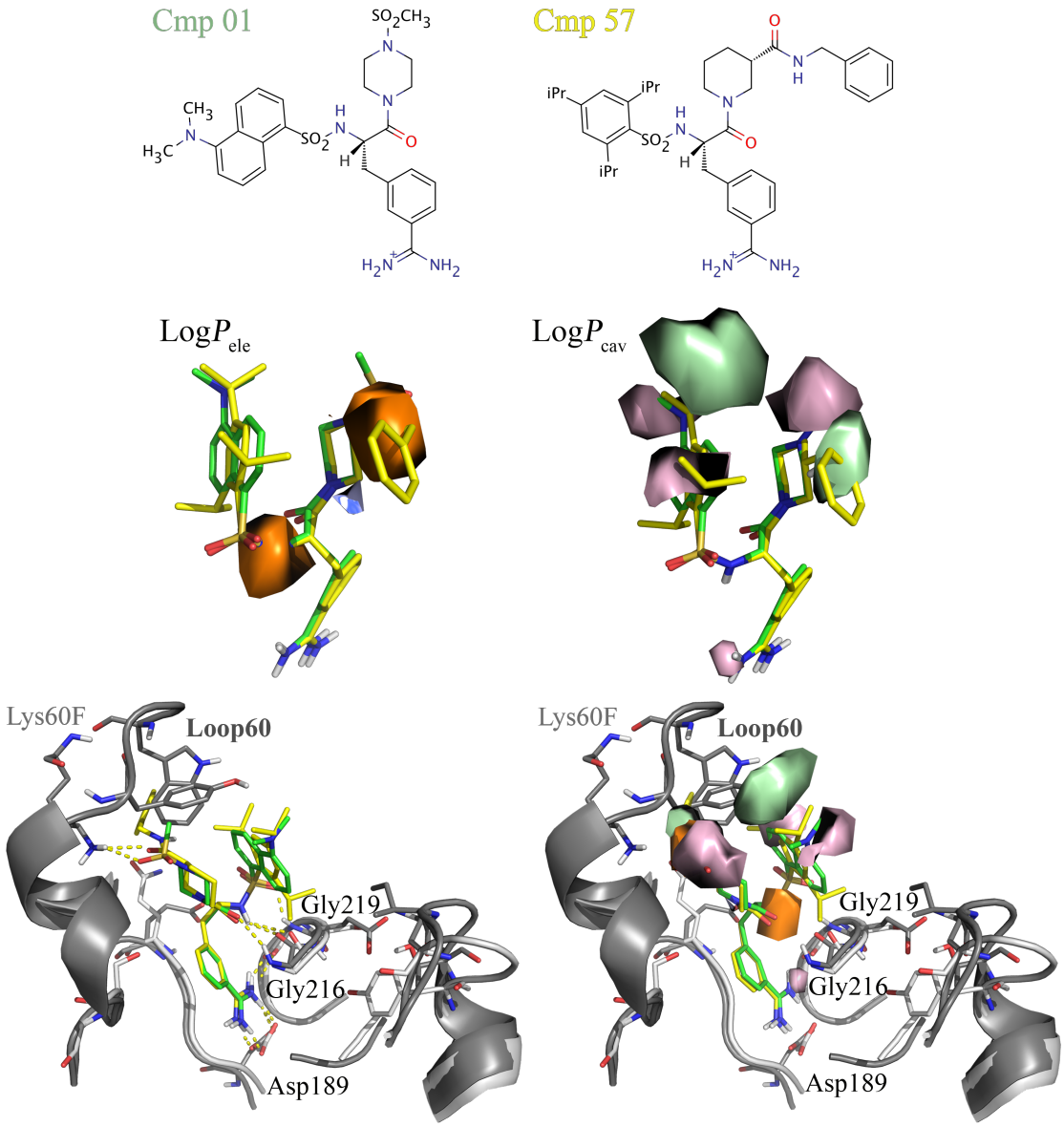


Figure 6. Representation of HyPhar H2 isocontour maps for the thrombin/trypsin selectivity model. Isocontours are expressed as the PLS coefficients corrected by the standard deviation and scaled by a factor of 10^5. The electrostatic field ($\log P_{ele}$) is shown as light blue (negative) and orange (positive) isocontours (values of -150 and 150, respectively), which correspond to areas where polarity increases/decrease the selectivity against thrombin. The non-electrostatic field ($\log P_{cav}$) is shown as light pink/pale green isocontours (values of -50 and 50, respectively) and denote areas where steric bulk is favourable/unfavourable for the selectivity against thrombin. The ligand-receptor complementary is shown for compounds **1** (green) and **57** (yellow), which are selective against thrombin and trypsin, respectively.

These findings also agree with previous results obtained from GRID and FLOGTV calculations for different structures of the targets [50, 51], which highlighted the relevance of Tyr60A and Trp60D for conferring selectivity against thrombin. A small pink isocontour is also found around the amidino unit, which would account for the proper positioning of the benzamidino moiety for interaction with Asp189 subject to the steric constraint imposed by the Loop60 on the substituents filling the S2 and S3 pockets. In contrast, the pale green areas are detrimental for the selectivity against thrombin due to steric clashes with residues in the inserted Loop60 of thrombin.

The selectivity maps between thrombin and factor Xa are shown in Figure 7. The topology of the map for the electrostatic component is similar to the representation found for the thrombin/trypsin pair, as expected from the changes of Lys60F and Glu192 by Gln61 and Gln192 in factor Xa. In agreement with these findings, it is worth noting that removal of the carboxylate group from DX-9065A, an inibitor selective for factor Xa where the acid group is located close to Gln192, reduces the selectivity over thrombin by a factor of more than 100 [64]. Likewise, as discussed in [50], incorporating an acid group in 1,2-dibenzamidobenzene inhibitors increased selectivity for factor Xa relative to thrombin, but not to trypsin [65]. On the other hand, with regard to the non-electrostatic term, the main feature is the light pink isocontour, which reflects the favourable effect of filling the area delimited by the Loop60 for gaining selectivity against thrombin, and to the steric clash originated from the mutation of Leu99 in thrombin to Tyr in factor Xa (Figure S2). Finally, the pale green area would be disfavorable for thrombin selectivity due to steric hindrance with Tyr60A.

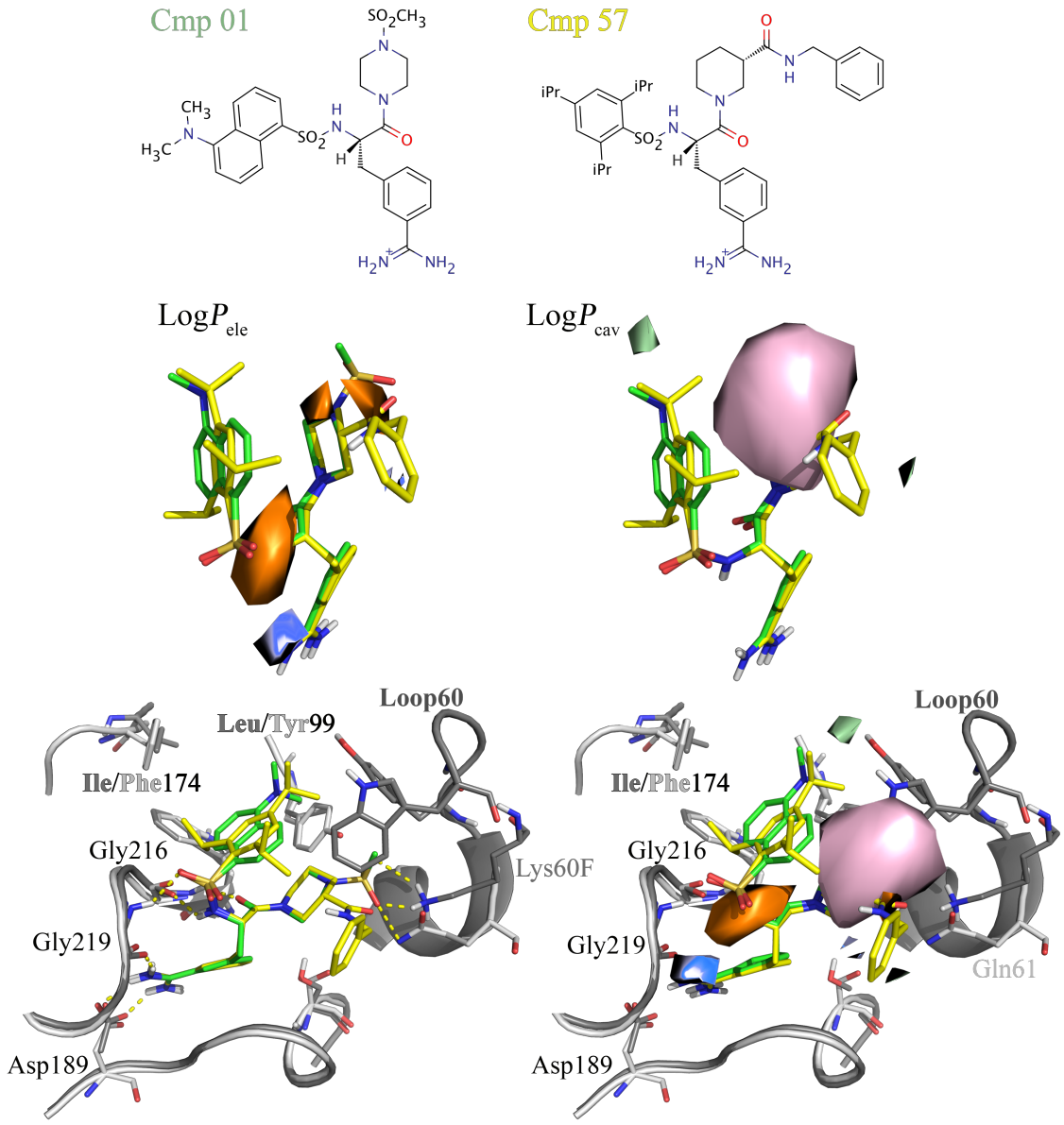


Figure 7. Representation of HyPhar H2 isocontour maps for the thrombin/factor Xa selectivity model. Isocontours are expressed as the PLS coefficients corrected by the standard deviation and scaled by a factor of 10^5. The electrostatic field ($\log P_{ele}$) is shown as light blue (negative) and orange (positive) isocontours (values of -100 and 100, respectively), which correspond to areas where polarity increases/decrease the selectivity against thrombin. The non-electrostatic field ($\log P_{cav}$) is shown as light pink/pale green isocontours (values of -35 and 25, respectively) and denote areas where steric bulk is favourable/unfavourable for the selectivity against thrombin. The ligand-receptor complementary is shown for compounds **1** (green) and **57** (yellow).

Finally, the selectivity maps for the trypsin/factor Xa pair are shown in Figure 8.

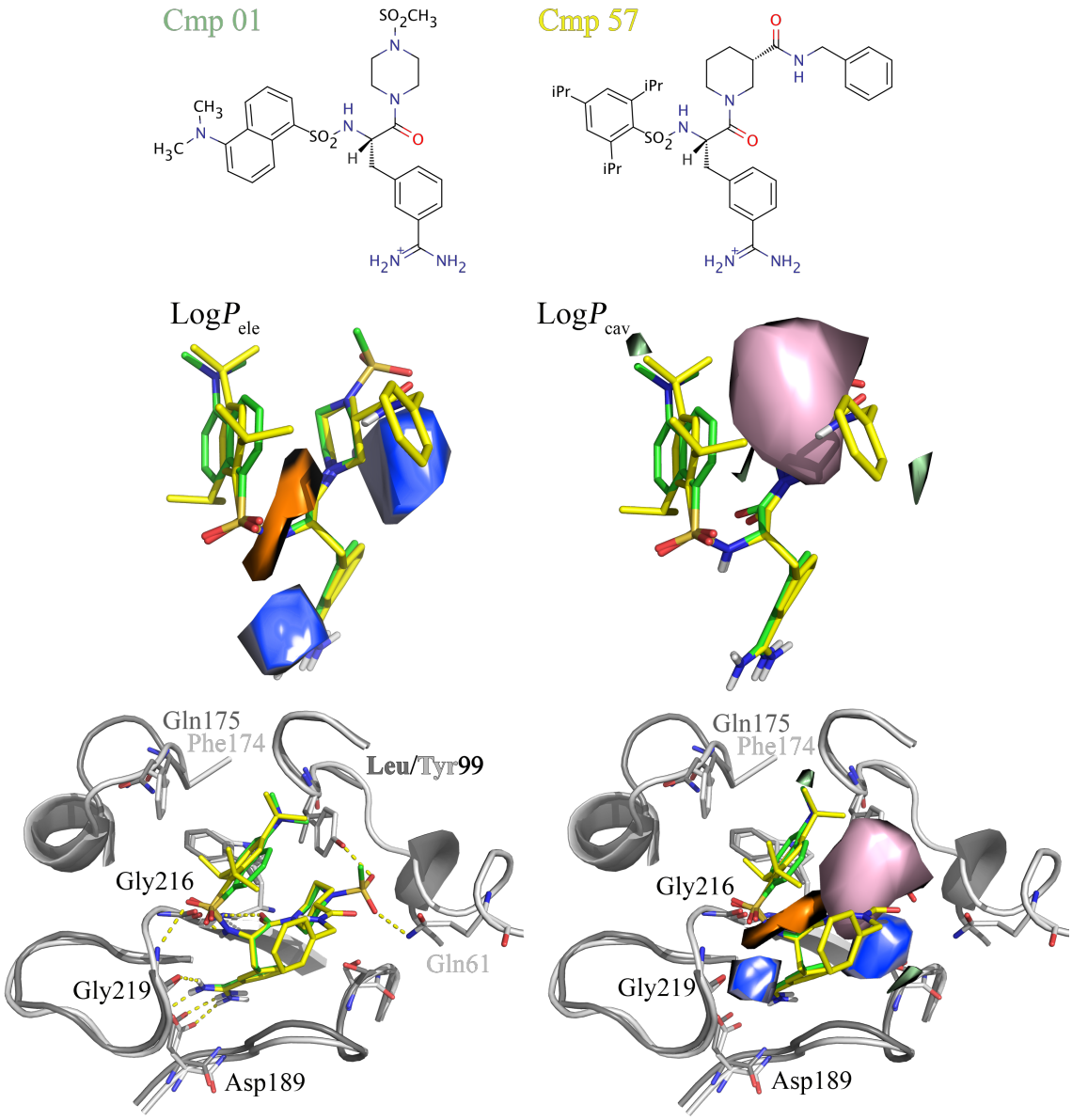


Figure 8. Representation of HyPhar H2 isocontour maps for the trypsin/factor Xa selectivity model. Isocontours are expressed as the PLS coefficients corrected by the standard deviation and scaled by a factor of 10^5. The electrostatic field ($\log P_{ele}$) is shown as light blue (negative) and orange (positive) isocontours (values of -150 and 100, respectively), which correspond to areas where polarity increases/decrease the selectivity against trypsin. The non-electrostatic field ($\log P_{cav}$) is shown as light pink/pale green isocontours (values of -50 and 40, respectively) and denote areas where steric bulk is favourable/unfavourable for the selectivity against thrombin. The ligand-receptor complementary is shown for compounds **1** (green) and **57** (yellow).

The light blue isocontour in the $\log P_{ele}$ field favour the selectivity against trypsin, presumably reflecting the net effect of Ser190 and Gln61 on the proper positioning of the inhibitor to fill the S1 (Asp189) and catalytic (His57, Ser195) sites, and the backbone differences in the Gly216-Gly219 loop [51]. Inspection of the non-electrostatic term shows that filling the pink area would favour the activity against trypsin. As noted above, this reflects the steric clashes that would arise in factor Xa with the side chain of residue Tyr99. On the other hand, the pale green areas denote selectivity against factor Xa, reflecting the extension toward the S3 site and the formation of van der Waals contacts with aromatic residues (Tyr99, Phe174, Trp215), and the occupancy of the catalytic pocket close to Gln61 due to the larger protrusion of the loop 59-64 toward the catalytic site in factor Xa.

Conclusions

This work constitutes an extension of the recently reported QM-SCRF-based hydrophobic descriptors [33] toward the definition of structure-activity relationships of congeneric series and the identification of selectivity determinants between targets, exemplified with the binding of benzamidine derivatives to thrombin, trypsin, and factor Xa. Quantitative models of affinity prediction compare well with 3D-QSAR studies for this set of compounds [29]. This supports the notion that the "electrostatic" and "non-electrostatic" components of the atomic hydrophobicity are effective in encoding the electrostatic and steric contributions typically used in 3D-QSAR. Furthermore, models of pairwise selectivity have also been derived and the selectivity maps can be interpreted in light of the moleculars determinants discussed in the literature using both ligand-based [29] and target-based techniques [50-53]. Overall, we believe that Hyphar parameters can be used to provide a complementary view to, rather

than a replacement for, standard descriptors in 3D-QSAR, thus leading to a more comprehensive understanding of the key factors involved in ligand-target recognition and binding.

Other 3D-QSAR models have been proposed to take into account parameters related to molecular hydrophobicity, such as Molecular Lipophilicity Potential [42] and Hydropathic Interactions [43,44] methods, which rely on empirical contributions to $\log P_{o/w}$. QM approaches led to the heuristic molecular lipophilic potential [66], which was based on the electrostatic potential at the molecular surface, and the use of the s-profiles derived from COSMO calculations as an alternative to molecular interaction fields [67,68]. Our strategy consists of deriving hydrophobic parameters from the partitioning of $\log P_{o/w}$ into atomic contributions from QM-SCRF IEF-PCM/MST continuum calculations in water and octanol, which were already used in molecular similarity studies [69-71]. However, this strategy might be easily extended to other refined versions of QM-SCRF methods, which are available in a variety of QM codes, thus facilitating the implementation of these parameters for 3D-QSAR studies. Compared to methods based on empirical contributions, the major disadvantage is the computational cost required for QM-SCRF calculations. However, this is compensated by the more accurate description of the molecular charge distribution, while it accounts for the specific ionization, tautomerization and conformational features of molecules. Moreover, the approach also benefits from the usage of a descriptor widely adopted in drug discovery studies and that can be compared with experimental data.

Supporting Information

Histograms of molecular properties (Figure S1), structural comparison of targets (Figure S2), comparison of electrostatic and non-electrostatic components of the

octanol/water partition coefficient (Figure S3), and tables reporting the experimental and predicted (H2 model) activities (Table S1-S3) and selectivities (Tables S4-S6) models.

Acknowledgment

We thank the financial support from Ministerio de Economía y Competitividad (SAF2014-57094-R) and the Generalitat de Catalunya (2014-SGR-1189). We are grateful to the Consorci de Serveis Universitaris de Catalunya for computational resources. FJL acknowledges the support from ICREA Academia.

References

1. Gohlke H, Klebe G (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew. Chem. Int. Ed. 41: 2644-2676.
2. Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. Nature 432: 855-861.
3. Walters WP, Murcko MA (2002) Prediction of 'drug-likeness'. Adv. Drug Deliv. Rev. 54: 255-271.
4. Vistoli G, Pedretti A, Testa B (2008) Assessing drug-likeness – what are we missing? Drug Discov Today 13: 285-294.
5. Urus O, Rayan A, Goldblum A, Oprea TI (2011) Understanding drug-likeness. WIREs Comput Mol Sci 1: 760-781.
6. Rognan D (2007) Chemogenomic approaches to rational drug design. Br J Pharmacol 1-15.
7. Kawasaki Y, Freire E (2011) Finding a better path to drug selectivity. Drug Discov. Today 16: 985-990.
8. Huggins DJ, Sherman W, Tidor B (2012) Rational approaches to improving selectivity in drug design. J Med Chem 55: 1424-1444.
9. Knight ZA, Shokat KM (2005) Features of selective kinase inhibitors. Chem Biol 12: 621-637.
10. Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR (2011) Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nature Biotech 29: 1039-1045.

11. Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP (2011) Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotech 29: 1046-1051.
12. Youdim MB, Weinstock M (2004) Therapeutic applications of selective and non-selective inhibitors of monoamine oxidase A and B taht do not cause significant tyramine potentiation. Neurotoxicology 25: 243-250.
13. Bertolini A, Ottani A, Sandrini M (2009) Selective COX-2 inhibitors and dual acting anti-inflammatory drugs: Critical remarks. Curr Med Chem 9: 1033-1043.
14. Card GL, England BP, Suzuki Y, Fong D, Powell B, Lee B, Luu C, Tabrizizad M, Gillete S, Ibrahim PN, Artis DR, Bollag G, Milburn MV, Kim S-H, Schlessinger J, Zhang KYJ (2004) Structural basis for the activity of drugs that inhibit phosphodiesterase. Structure 12: 2233-2247.
15. Kastenholz MA, Pastor M, Cruciani G, Haaksma EEJ, Fox T (2000) GRID/CPCA: A new computational toll to design selective ligands. J Med Chem 43: 3033-3044.
16. Ortiz AR, Gomez-Puertas P, Leo-Macias A, Lopez-Romero P, Lopez-Viñas E, Morreale A, Murcia M, Wang K (2006) Computational approaches to model ligand selectivity in drug design. Curr Top Med Chem 6: 41-55.
17. Kold P, Phan K, Gao Z-G, Marko AC, Sali A, Jaconson KA (2012) Limits of ligand selectiviy from docking to models: In silico screening for A1 adenosine receptor antagonists. PLoS ONE 7: e49910.
18. Rodrigues T, Kudoh T, Roudnicky F, Lim YF, Lin Y-C, Koch CP, Seno M, Detmar M, Schneider G (2013) Steering target selectivity and potency by fragment-based de novo drug design. Angew Chem Int Ed 52: 10006-10009.

19. Rath SL, Senapati S (2013) Molecular basis of differential selectivity of cyclobutyl-substituted imidazole inhibitors against CDKs: insights for rational drug design PLoS ONE 8: e73836.
20. Tarcsay A, Keserú GM (2015) Is there a link between selectivity and binding thermodynamic profiles? Drug Discov Today 20: 86-94.
21. Freyhult E, Gustafsson MG, Strömbergsson H (2015) A machine learning approach to explain drug selectivity to solublre and membrane protein targets. Mol Inform 34: 44-52.
22. Cramer RD, Wendt B (2007) Pushing the boundaries of 3D-QSAR. J Comput Aided Mol Des 21: 23-32.
23. Verma J, Khedar VM, Coutinho EC (2010) 3D-QSAR in drug design – A review. Curr Top Med Chem 10: 95-115.
24. Artese A, Cross S, Costa G, Distinto S, Parrotta L, Alcaro S, Ortuso F, Cruciani G (2013) Molecular interaction fields in drug discovery: Recent advances and future perspectives. WIREs Comput Mol Sci 3: 594-613.
25. Cramer RD, III, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110: 5959-5967.
26. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37: 4130-4146.
27. Klebe G, Abraham U (1999) Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. J Comput-Aided Mol Des 13: 1-10.

28. Böhm M, Klebe G (2002) Development of new hydrogen-bond descriptors and their application to comparative molecular field analyses. J Med Chem 45: 1585-1597.
29. Böhm M, Sturzebecher J, Klebe G (1999) Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. J Med Chem 42: 458-477.
30. Baskin II, Tikhonova IG, Palyulin VA, Zefirov NS (2003) Selectivity fields: Comparative molecular field analysis (CoMFA) of the glycine NMDA and AMPA receptors. J Med Chem 46: 4063-4069.
31. Walline CC, Nichols DE, Carroll FI, Barker EL (2008) Comparative molecular field analysis using selectivity fields reveals residues in the third transmembrane helix of the serotonin transporter associated with substrate and antagonist recognition. J Pharmacol Exp Ther 325: 791-800.
32. Sharma RN, Thakar HM, Vasu KK, Chaturvedi SC, Pancholi SS (2009) Pair wise binding affinity: 3D QSAR studies on a set of triazolo [1,5-a] quinoxalines as antagonists of AMPA and KA receptors. J. Enzyme Inhib Med Chem 24: 1008-1014
33. Ginex T, Muñoz-Muriedas J, Herrero E, Gibert E, Cozzini P, Luque FJ (2016) Development and validation of hydrophobic molecular fields from the quantum mechanical IEF/PCM-MST solvation models in 3D-QSAR. J Comput Chem in press.
34. Arkin MR, Wells JA (2004) Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream. Nat Rev Drug Discov 3: 301-317.

35. Hajduk PJ, Huth JR, Fesik SW (2005) Druggability índices for protein targets derived from NMR-based screening data. J Med Chem 45: 2615-2623.
36. Nayal M, Honig B (2006) On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. Proteins 63: 892-906.
37. Egner U, Hillig RC (2008) A structural biology view of target druggability. Expert Opin Drug Discov 3: 391-401.
38. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES (2007) Structure-based maximal affinity model predicts small-molecule druggability. Nat Biotechnol 25: 71-75.
39. Schmidtke P, Barril X (2010) Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. J Med Chem 53: 5858-5867.
40. Schmidtke P, Luque FJ, Murray JB, Barril X (2011) Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. J Am Chem Soc 133: 18903-18910.
41. Alvarez-Garcia D, Barril X (2014) J Med Chem 57: 8530-8539.
42. Gaillard P, Carrupt P.-A., Testa B, Boudon A (1994) Molecular lipophilicity potential, a tool in 3D QSAR: Method and applications. J Comput-Aided Mol Des 8: 83-96.
43. Kellog GE, Semus SF, Abraham DJ (2000) HINT: A new method of empirical hydrophobic field calculation for CoMFA. J Comput-Aided Mol Des 5: 545-552.
44. Kellog GE, Abraham DJ (2000) Hydrophobicity: is LogPo/w more than the sum of its parts? Eur J Med Chem 35: 651-661.
45. Mennucci B (2012) Polaizable continuum model. WIREs Mol Comput Sci 2: 386-404.

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

47. Cramer CJ, Truhlar DG (2008) A universal approach to solvation modeling. Acc Chem Res 41: 760-768.

48. Klamt A, Mennucci B, Tomasi J, Barone V, Curutchet C, Orozco M, Luque FJ (2009) On the performance of continuum solvation methods. Acc Chem Res 42: 489-492.

49. Sander T, Freyss J, von Korff M, Rufener C (2015) DataWarrior: An open-source program for chemistry aware data visualization and analysis. J Chem Inf Model 55: 460-473.

50. Kastenholz MA, Pastor M, Cruciani G, Haaksma EEJ, Fox T (2000) GRID/CPCA: A new computational tool to design selective ligands. J Med Chem 43: 3033-3044.

51. Sheridan RP, Holloway MK, McGaughey G, Mosley RT, Sing SB (2002) A simple method for visualizing the differences between related receptor sites. J Mol Graphics Model 2002: 71-79.

52. Murcia M, Ortiz AR (2004) Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. J Med Chem 47: 805-820.

53. Murcia M, Morreale A, Ortiz AR (2006) Comparative binding energy analysis considering multiple receptors: A step toward 3D-QSAR models for multiple targets. 49: 6241-6253.

54. Luque FJ, Barril X, Orozco M (1999) Fractional description of free energies of solvation. J Comput-Aided Mol Des 13: 139-152.

55. Luque FJ, Bofill JM, Orozco M (1995) Novel strategies to incorporate the solvent polarization in self-consistent reaction field and free-energy perturbation simulations. J Chem Phys 103: 10183-10191.
56. Curutchet C, Orozco M, Luque FJ (2001) Solvation in octanol: Parametrization of the continuum MST model. J Comput Chem 22: 1180- 1193.
57. Soteras I, Curutchet C, Bidon-Chanal A, Orozco M, Luque FJ (2005) Extension of the MST model to the IEF formalism: HF and B3LYP parametrizations. J Mol Struct (THEOCHEM) 727: 29-40.
58. PharmQSAR; Pharmacelera; Barcelona; 2015.
59. Clark M, Cramer RD III, van Opdenbosch (1989) Validation of the general purpse Tripos 5.2. J Comput Chem 10: 982-1012.
60. Sutherland JJ, O'Brien LA, Weaver DF (2004) A comparison of methods for modeling quantitative structure-activity relationships. 47: 5541-5554.
61. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA Jr., J. E. Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R,Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ, Gaussian 09, Revision D.01; Gaussian. Inc.; Wallingford, CT, 2009.

62. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: A basic tool of chemometrics. Chemometr Intell Lab Syst 58: 109-130.

63. Aguinis H, Gottfredson R K, Joo H (2013) Best-practice recommendations for defining, identifying, and handling outliers. Organ Res Methods16: 270-301.

64. Katakura S, Nagahara T, Hara T, Iwamoto M (1993) A novel factor Xa inhibitor: Structure-activity relationships and selectivity between factor Xa and thrombin. Biochem Biophys Res Commun 197: 965-972.

65. Guilford WJ, Shaw KJ, Dallas JL, Koovakkat S, Lee W, Liang A, Light DR, McCarrick MA, Whitlow M, Ye B, Morrisey MM (1999) Synthesis, characterization, and structure-activity relationships of amidine-substituted (bis)benzylidene-cycloketone olefin isomers as potent and selective factor Xa inhibitors. J Med Chem 42: 5415-5425.

66. Du Q, Liu P-J, Mezey PG (2005) Theoretical derivation of heuristic molecular lipophilic potential. A quantum chemical description for molecular solvation. J Chem Inf Model 45: 347-353.

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

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

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

70. Muñoz-Muriedas J, Perspicace S, Bech N, Guccione S, Orozco M, Luque FJ (2005) Hydrophobic molecular similarity from MST fractional contributions to the octanol/water partition coefficient. J Comput-Aided Mol Des 19: 401-419.
71. Muñoz-Muriedas J, Barril X, López JM, Orozco M, Luque FJ (2007) A hydrophobic similarity analysis of solvation effects on nucleic acid bases. J Mol Mod 13: 357-365.

Supporting Information

Application of the Quantum Mechanical IEF/PCM-MST Hydrophobic Descriptors to Selectivity in Ligand Binding

Tiziana Ginex.[a] Jordi Muñoz-Muriedas.[b] Enric Herrero.[c] Enric Gibert.[c] Pietro Cozzini.[a]* and F. Javier Luque[d]*

[a] Dipartimento di Scienze degli Alimenti. University of Parma. Parco Area delle Scienze 59/A. 43121 Parma. Italy

[b] GlaxoSmithKline. Medicines Research Centre. Gunnels Wood Road. Stevenage SG1 2NY. United Kingdom

[c] Pharmacelera. Jordi Girona 1-3. Campus Nord Universitat Politècnica de Catalunya. Edifici K2M. 08034 Barcelona. Spain

[d] Department of Chemical Physics and Institut de Biomedicina (IBUB). Faculty of Pharmacy. University of Barcelona. Av. Prat de la Riba 171. 08921 Santa Coloma de Gramenet. Spain

* E-mail: pietro.cozzini@unipr.it (PC) or fjluque@ub.edu (FJL)

Figure S1. Histograms of molecular properties (molecular weight (MW); number of hydrogen-bond donors/acceptors (HBD/HBA); clogP. and number of rotatable bonds) for the series of 88 inhibitors.

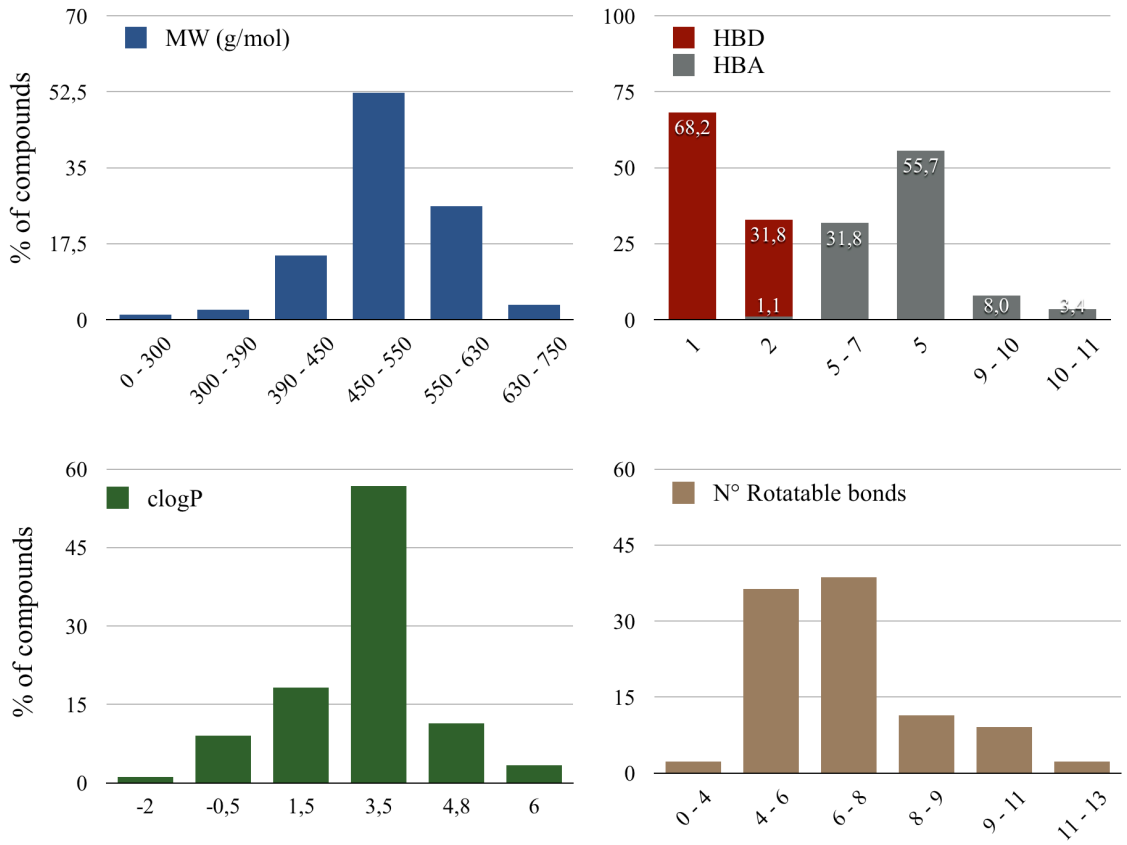


Figure S2. Representation of the superposed X-ray structures of thrombin (1ETS; green), trypsin (1PPH; blue) and factor Xa (1HCG; orange). Selected residues shared for binding to a given pair of targets are shown as white sticks, and for selectivity are shown as coloured sticks.

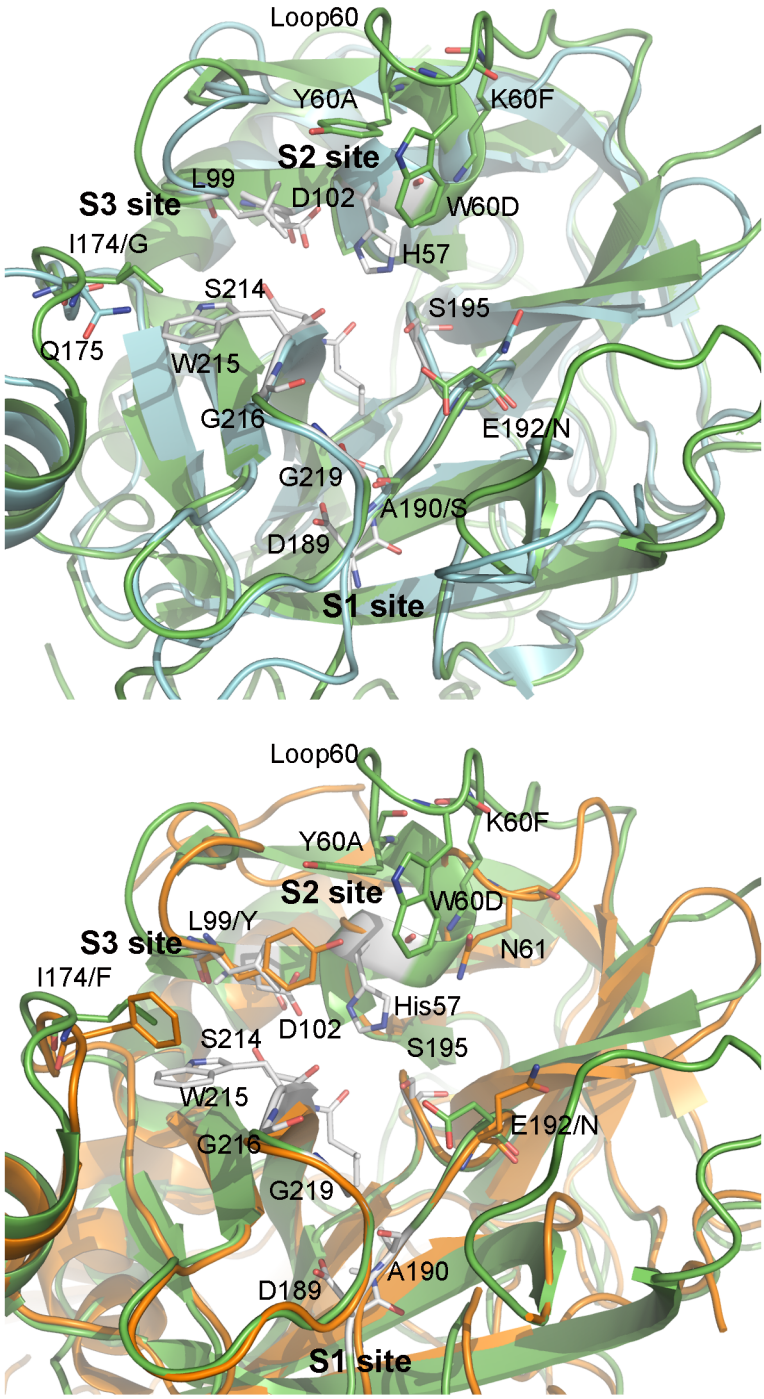


Figure S3. Comparison of the electrostatic and non-electrostatic (cavitation, van der Waals) contributions obtained for the whole set of 88 inhibitors.

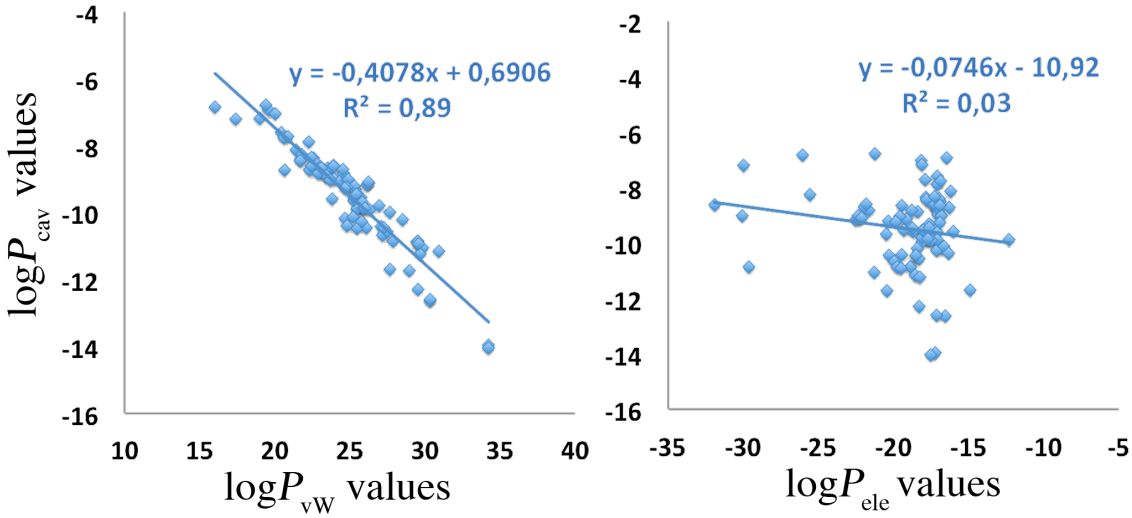


Table S1. Experimental and HyPhar H2 predicted inhibitory activities against thrombin for the set of compounds.

	pK_i			pK_i	
N [a]	Exp	HyPhar H2	N [a]	Exp	HyPhar H2
1	8.377	8.02	45	6.469	7.15
2	8.367	8.31	46	6.456	6.27
3	8.301	8.24	47	6.377	6.43
4	8.208	7.83	48	6.301	6.34
5	8.131	8.16	49	6.292	7.07
6	8.056	8.16	50	6.244	6.08
7	7.854	6.86	51	6.201	6.03
8	7.796	7.20	52	6.180	6.21
9	7.770	6.98	53	6.161	6.56
10	7.745	7.64	54	6.046	5.96
11	7.721	7.70	55	5.959	5.86
12	7.721	7.54	56	5.921	6.46
13	7.678	7.64	57	5.745	5.72
14	7.638	7.67	58	5.678	6.25
15	7.585	7.36	59	5.638	6.03
17	7.469	8.03	60	5.538	5.56
18	7.432	7.40	61	5.509	5.80
19	7.432	7.35	62	5.509	5.58
20	7.377	7.17	63	5.244	5.58
21	7.377	6.79	64	5.208	4.59
22	7.237	7.28	65	5.137	5.15
23	7.229	7.23	66	4.886	5.01
24	7.187	7.08	67	4.824	4.55
25	7.125	7.16	68	4.770	5.06
26	7.051	7.07	69	4.569	4.82
27	7.018	6.84	70	4.523	4.16
28	6.959	6.27	71	4.456	4.99
29	6.921	6.77	72	4.357	4.85
30	6.921	6.84	74	7.886	7.37
31	6.921	7.28	75	7.585	7.28
32	6.824	6.60	76	7.523	7.55
33	6.824	7.17	77	7.444	7.31
34	6.796	7.13	78	7.284	6.26
35	6.745	6.78	79	7.155	6.34
36	6.699	6.62	80	6.770	6.74
37	6.678	6.32	81	6.585	6.77
38	6.638	6.85	82	6.553	7.21
39	6.638	6.72	83	6.523	6.94
40	6.585	6.94	84	6.284	8.17
41	6.553	5.92	85	6.284	6.61

42	6.553	6.63	86	6.149	6.62
43	6.495	6.34	87	5.420	5.17
44	6.469	7.29	88	4.745	5.64

[a] Test set: compounds 74-88.

Table S2. Experimental and HyPhar H2 predicted inhibitory activities against trypsin for the set of compounds.

	pK_i			pK_i	
N [a]	Exp	HyPhar H2	N [a]	Exp	HyPhar H2
1	6.770	6.73	45	5.921	5.87
2	6.796	6.86	46	5.658	5.80
3	6.699	6.55	47	5.678	5.60
4	6.854	6.53	48	6.658	6.59
5	6.119	6.19	49	6.367	6.54
6	6.770	6.56	50	6.237	6.10
7	6.201	6.10	51	6.000	5.86
8	6.201	5.98	52	5.092	5.19
9	7.444	6.91	53	5.921	5.90
10	6.886	7.09	54	6.071	5.82
11	7.699	7.37	55	6.357	6.00
12	6.26	6.53	56	4.854	4.75
13	6.854	6.86	57	6.337	6.33
14	7.131	7.25	58	7.097	7.19
15	6.284	6.15	59	5.102	5.33
17	6.137	5.97	60	4.796	5.00
18	6.585	6.59	61	6.569	6.90
19	6.658	6.68	62	4.495	4.59
20	6.284	6.32	63	4.602	4.70
21	6.678	6.41	64	4.796	4.64
22	5.959	6.31	65	5.62	5.70
23	5.398	5.70	66	4.538	4.24
24	6.481	6.57	67	6	6.06
25	6.161	6.40	68	3.854	4.13
26	6.108	5.76	69	4.538	4.56
27	5.658	5.72	70	3.928	3.86
28	5.854	5.63	71	4.509	4.42
29	5.347	5.22	72	3.000	3.21
30	5.824	5.84	74	6.585	6.07
31	5.398	5.60	75	6.495	6.30
32	6.409	6.53	76	6.215	6.63
33	6.569	6.42	77	5.886	7.32
34	6.796	7.16	78	6.357	5.72
35	6.004	6.28	79	5.721	5.52
36	5.921	6.02	80	6.149	6.29
37	7.174	7.11	81	6.509	6.44
38	6.215	5.81	82	6.009	6.29
39	6.102	5.90	83	6.796	6.98
40	6.201	6.19	84	7.569	7.60
41	5.602	5.64	85	5.745	5.76

42	6.921	6.95	86	7.638	7.15
43	5.921	5.93	87	4.585	4.56
44	5.444	6.20	88	4.337	4.58

[a] Test set: compounds 74-88.

Table S3. Experimental and HyPhar H2 predicted inhibitory activities against factor Xa for the set of compounds.

	pK_i			pK_i	
N [a]	Exp	HyPhar H2	N [a]	Exp	HyPhar H2
1	5.409	5.37	45	4.824	4.41
2	5.167	4.87	46	4.119	4.44
3	4.921	4.90	47	5.495	5.30
4	4.387	4.39	48	4.678	4.66
5	4.125	4.06	49	4.796	4.87
6	4.620	4.63	50	4.363	4.51
7	4.854	4.80	51	4.699	4.67
8	4.377	4.23	52	3.959	4.20
9	4.367	4.32	53	5.114	5.01
10	4.382	4.41	54	6.046	5.86
11	4.114	4.33	55	4.357	3.97
12	4.585	4.63	56	4.194	4.04
13	4.319	4.56	57	5.602	5.45
14	5.638	5.83	58	5.119	5.23
15	4.149	4.26	59	4.244	4.48
17	4.097	4.06	60	3.886	3.91
18	4.745	4.51	61	4.456	4.70
19	4.721	4.73	62	5.000	5.07
20	5.658	5.44	63	4.585	4.34
21	4.796	4.77	64	3.444	3.30
22	4.456	4.14	65	4.886	4.90
23	4.022	4.21	66	3.000	2.93
24	4.420	4.52	67	4.658	4.75
25	5.699	5.61	68	3.721	3.92
26	4.000	3.97	69	3.638	3.52
27	4.237	4.55	70	3.886	3.94
28	5.456	5.80	71	4.387	4.49
29	4.268	4.23	72	3.194	3.34
30	5.638	5.64	74	5.509	5.01
31	4.238	4.57	75	5.013	4.39
32	4.745	4.84	76	4.658	4.68
33	5.585	5.49	77	4.523	4.90
34	4.119	4.45	78	4.301	4.44
35	4.076	4.04	79	4.337	4.64
36	4.569	4.53	80	4.284	4.61
37	5.092	5.12	81	4.481	4.53
38	4.770	4.36	82	5.131	5.26
39	5.824	5.83	83	5.066	5.15
40	4.420	4.21	84	4.569	4.45
41	5.602	5.28	85	4.444	4.69

42	4.770	4.83	86	4.602	4.56
43	4.886	4.69	87	4.959	5.37
44	3.745	4.20	88	4.398	4.42

[a] Test set: compounds 74-88.

Table S4. Experimental differences in binding affinity for the pair thrombin/trypsin and predicted values from HyPhar H2 model for the set of compounds.

	pK_i			pK_i	
N [a]	Exp	HyPhar H2	N [a]	Exp	HyPhar H2
1	1.607	1.49	45	0.548	1.20
2	1.571	1.57	46	0.798	0.76
3	1.602	1.47	47	0.699	0.73
4	1.354	1.13	48	-0.357	-0.14
5	2.012	1.37	49	-0.075	0.14
6	1.286	1.60	50	0.007	0.02
7	1.653	1.43	51	0.201	0.31
8	1.595	1.07	52	1.088	0.67
9	0.326	0.14	53	0.240	0.62
10	0.859	0.71	54	-0.025	-0.12
11	0.022	1.05	55	-0.398	-0.28
12	1.461	1.00	56	1.067	1.33
13	0.824	0.97	57	-0.592	-0.53
14	0.507	1.17	58	-1.419	-0.81
15	1.301	1.15	59	0.536	0.61
17	1.332	1.34	60	0.742	0.54
18	0.847	1.05	61	-1.060	-1.07
19	0.774	0.84	62	1.014	0.91
20	1.093	1.08	63	0.642	0.76
21	0.699	0.61	64	0.412	0.40
22	1.278	0.82	65	-0.483	-0.66
23	1.831	1.37	66	0.348	0.53
24	0.706	0.69	67	-1.176	-1.32
25	0.964	0.69	68	0.916	0.99
26	0.943	1.34	69	0.031	0.14
27	1.360	0.79	70	0.595	0.42
28	1.105	0.78	71	-0.053	0.04
29	1.574	1.66	72	1.357	0.78
30	1.097	0.99	74	1.301	1.15
31	1.523	1.43	75	1.090	1.10
32	0.415	0.21	76	1.308	1.06
33	0.255	0.99	77	1.558	0.87
34	0.000	0.17	78	0.927	0.76
35	0.741	0.54	79	1.434	0.67
36	0.778	0.75	80	0.621	0.77
37	-0.496	-0.81	81	0.076	-0.07
38	0.423	0.98	82	0.544	0.92
39	0.536	0.72	83	-0.273	0.20
40	0.384	0.64	84	-1.285	1.02
41	0.951	0.57	85	0.539	0.55

42	-0.368	0.00	86	-1.489	-0.57
43	0.574	0.31	87	0.835	0.49
44	1.025	1.09	88	0.408	0.96

[a] Thrombin/trypsin external test set: compounds 74-88.

Table S5. Experimental differences in binding affinity for the pair thrombin/factor Xa and predicted values from HyPhar H2 model for the set of compounds.

	pK_i			pK_i	
N [a]	Exp	HyPhar H2	N [a]	Exp	HyPhar H2
1	2.968	3.00	45	1.645	2.69
2	3.200	3.56	46	2.337	2.02
3	3.380	3.37	47	0.882	1.19
4	3.821	3.36	48	1.623	1.65
5	4.006	3.77	49	1.496	1.68
6	3.436	3.43	50	1.881	1.73
7	3.000	2.55	51	1.502	1.50
8	3.419	2.88	52	2.221	1.90
9	3.403	2.89	53	1.047	0.95
10	3.363	3.29	54	0.000	-0.01
11	3.607	3.44	55	1.602	1.87
12	3.136	2.88	56	1.727	1.99
13	3.359	3.16	57	0.143	-0.01
14	2.000	2.02	58	0.559	0.84
15	3.436	3.12	59	1.394	1.65
17	3.372	3.59	60	1.652	1.90
18	2.687	3.03	61	1.053	1.26
19	2.711	2.76	62	0.509	0.36
20	1.719	2.21	63	0.659	1.27
21	2.581	2.37	64	1.764	1.60
22	2.781	3.00	65	0.251	0.03
23	3.207	2.78	66	1.886	1.65
24	2.767	2.69	67	0.166	0.23
25	1.426	1.56	68	1.049	1.11
26	3.051	2.77	69	0.931	1.21
27	2.781	2.12	70	0.637	0.45
28	1.503	0.73	71	0.069	0.10
29	2.653	2.70	72	1.163	1.04
30	1.283	1.31	74	2.377	2.16
31	2.683	2.31	75	2.572	3.08
32	2.079	2.00	76	2.865	2.96
33	1.239	1.67	77	2.921	2.91
34	2.677	2.85	78	2.983	2.00
35	2.669	2.62	79	2.818	1.73
36	2.130	2.16	80	2.486	2.48
37	1.586	1.36	81	2.104	2.22
38	1.868	2.45	82	1.422	1.77
39	0.814	0.71	83	1.457	1.84
40	2.165	2.69	84	1.715	3.56
41	0.951	1.01	85	1.840	1.90

42	1.783	1.94	86	1.547	2.12
43	1.609	1.94	87	0.461	-0.08
44	2.724	2.99	88	0.347	1.36

[a] Thrombin/factor Xa external test set: compounds 74-88.

Table S6. Experimental differences in binding affinity for the pair trypsin/factor Xa and predicted values from HyPhar H2 model for the set of compounds.

	pK_i			pK_i	
N [a]	Exp	HyPhar H2	N [a]	Exp	HyPhar H2
1	1.361	1.66	45	1.097	1.51
2	1.629	2.04	46	1.539	1.17
3	1.778	1.79	47	0.183	0.43
4	2.467	2.26	48	1.980	1.96
5	1.994	2.15	49	1.571	1.86
6	2.150	2.22	50	1.874	1.77
7	1.347	1.28	51	1.301	1.42
8	1.824	1.76	52	1.133	1.05
9	3.077	2.47	53	0.807	0.75
10	2.504	2.63	54	0.025	0.26
11	3.585	2.45	55	2.000	1.92
12	1.675	1.92	56	0.66	1.22
13	2.535	2.27	57	0.735	0.94
14	1.493	1.40	58	1.978	2.20
15	2.135	1.90	59	0.858	0.91
17	2.040	1.92	60	0.910	1.16
18	1.840	1.77	61	2.113	2.23
19	1.937	1.95	62	-0.505	-0.62
20	0.626	0.86	63	0.017	0.43
21	1.882	1.59	64	1.352	1.11
22	1.503	2.17	65	0.734	0.48
23	1.376	1.40	66	1.538	1.38
24	2.061	2.01	67	1.342	1.24
25	0.462	0.82	68	0.133	-0.02
26	2.108	1.44	69	0.900	1.18
27	1.421	1.30	70	0.042	-0.20
28	0.398	-0.21	71	0.122	0.04
29	1.079	0.89	72	-0.194	-0.06
30	0.186	0.54	74	1.076	0.69
31	1.160	0.81	75	1.482	2.01
32	1.664	1.68	76	1.557	1.95
33	0.984	0.91	77	1.363	2.73
34	2.677	2.56	78	2.056	1.18
35	1.928	1.95	79	1.384	0.99
36	1.352	1.17	80	1.865	1.62
37	2.082	2.10	81	2.028	1.93
38	1.445	1.44	82	0.878	0.91
39	0.278	0.14	83	1.730	2.06
40	1.781	2.12	84	3.000	2.34
41	0.000	0.20	85	1.301	1.24

42	2.151	2.14	86	3.036	2.51
43	1.035	1.41	87	-0.374	-0.80
44	1.699	1.93	88	-0.061	0.25

[a] Trypsin/factor Xa external test set: compounds 74-88.

Conclusions

A careful monitoring of all the technological and nutritional aspects of foods is necessary to guarantee the nutritional, safety and quality aspects of foods and their impact in global health. Standardized levels of food safety and quality, from the raw material to the final product, have to be carefully assessed through specific operating protocols. In a food-related multi-scale context, computational approaches are valuable in (i) modelling manufacturing processes, (ii) evaluating the risks associated with the food chain (risk assessment), (iii) and analysing the nutritional aspects of food components.

In this context, this doctoral dissertation aims to highlight the potentialities of computational simulations for the analysis of detailed and complex phenomena in the context of food science and technologies. Specifically, two major issues have been considered. First, the direct application of known computational techniques for the analysis of food contaminants. Second, the development of new molecular descriptors that fall within novel computational frameworks, such as HyPhar (Hydrophobic Pharmacophore).

The first work is related to the *in silico* preliminary evaluation of endocrine disrupting effects for some food contaminants (thioxanthone photoinitiators). To this end, structure-based computational protocols permit to analyse the potential toxic events of these chemicals and their *in silico* predicted role as phase I metabolites from a structural point of view, revealing their sub-molecular implications. Five synthetic thioxanthones and 23 (**M1-M23**) phase I metabolites have been analysed. For the first ten (**M1-M10**) an experimental confirmation of their existence was available in literature, whereas the rest of compounds (**M11-M23**) have been hypothesized from *in silico* calculations.

The putative binding modes and the affinity for AR-LBD have been predicted for all these chemicals. The final *in silico* results can be summarized as follows:

- C6-hydroxylation (M12, M16, M17, M20 and M21) seems to greatly enhance AR binding affinity through the formation of H-bonding interactions with Gln711, Met745 and Arg752 within the binding site. With the exception of M12, affinity ratios greater than 2 were obtained for these metabolites. The lower affinity ratio (1.29) observed for M12 could be ascribed to its bulkier C4-propoxy group.
- Sulfoxidation (M13, M14, M15, and M19) is not tolerated due to polar-apolar mismatch within the AR binding site.
- More steric hindrance on C4 (M18, M19, M22 and M23) seems to be poorly tolerated.

All these preliminary data suggest that further *in vitro* analyses should be necessary to correctly estimate the toxicological risk for this class of food contaminants.

Within the well-known context of structure-activity relationships, the second and the third works pursue to develop and validate new molecular descriptors able to (i) approximate the physicochemical molecular features associated to specific biological effects, and (ii) to discriminate their selective behaviour among closely related biomolecular systems. These two works have been motivated by several reasons:

- Overcame the classical limitations of CoMFA models (as incomplete account of entropic contributions, singularities and discontinuities in fields projections and maps representations).
- Introduce new more intuitive and easily interpretable parameters for 3D-QSAR analyses.
- Reach a higher level of consistency in molecular properties description.

To this end, five molecular sets have been retrieved from the literature and analysed by using the new molecular descriptors in conjunction with the computational protocol named HyPhar. The statistical performances of the models have been compared with those of the reference CoMFA/CoMSIA methodologies. Final comparative results pointed out a satisfactory predictive power, closely comparable to the classical 3D-QSAR methods but with some relevant features:

- The *qualitative* confirmation of suitability for the pharmacophoric elements identified by the new descriptors through a detailed analysis of the protein counterpart, when available.
- The *quantitative* statistical accuracy offered by the classical 3D-QSAR is reached with a lower number of principal components for the cross-validated models and, in case of CoMSIA, with a lower number of property fields, where the latter probably derived from the higher level of accuracy of the new QM-derived hydrophobic descriptors.
- The possibility to analyse and map the sub-molecular features associated to a different behaviour of the same molecule toward similar targets pertaining to the so-called selectivity analysis.

Taking into account all these points, the HyPhar methodology has shown potential to be utilized in a wide domain of applicability and adaptability in food sciences, since the hydrophobic descriptors are more consistent, intuitive and directly linked to experimentally measurable data ($\log P_{o/w}$).

Overall, the results allow us to be confident in the potential impact of *in silico* techniques to disclose the molecular events implicated in nutritional aspects of foods.

References

(1) Ganesh, Vijayalakshmi, and Navam S. Hettiarachchy. (2012) Nutriproteomics: a promising tool to link diet and diseases in nutritional research. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1824, 10, 1107-1117.

(2) Van der Sman, R. G. M. (2012) Soft matter approaches to food structuring. *Adv. Colloid Interface Sci. 176-177*, 18–30.

(3) Limbach, H. J., and Kremer, K. (2006) Multi-scale modelling of polymers: Perspectives for food materials. *Trends Food Sci. Technol. 17*, 215–219.

(4) Alavi, S. H., Rizvi, S. S. H., and Harriott, P. (2003) Process dynamics of starch-based microcellular foams produced by supercritical fluid extrusion. I: model development. *Food Res. Int. 36*, 309–319.

(5) De Gennes, P.-G. (1992) Soft Matter (Nobel Lecture). *Angew. Chemie Int. Ed. English 31*, 842–845.

(6) De Pablo, J. J. (2011) Coarse-grained simulations of macromolecules: from DNA to nanocomposites. *Annu. Rev. Phys. Chem. 62*, 555–74.

(7) Winkler, R. G., Ripoll, M., Mussawisade, K., and Gompper, G. (2005) Simulation of complex fluids by multi-particle-collision dynamics. *Comput. Phys. Commun. 169*, 326–330.

(8) Swift, M. R., Orlandini, E., Osborn, W. R., and Yeomans, J. M. (1996) Lattice Boltzmann simulations of liquid-gas and binary fluid systems. *Phys. Rev. E 54*, 5041–5052.

(9) Rowlinson, J. S. (1979) Translation of J. D. van der Waals' ?The thermodynamik theory of capillarity under the hypothesis of a continuous variation of density? *J. Stat. Phys. 20*, 197–200.

(10) Speck, T., and Seifert, U. (2009) Extended fluctuation-dissipation theorem for soft matter in stationary flow. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. 79*, 040102.

(11) Noguchi, H., and Takasu, M. (2001) Self-assembly of amphiphiles into vesicles: a Brownian dynamics simulation. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. 64*, 041913.

(12) Venturoli, M., and Smit, B. (1999) Simulating the self-assembly of model membranes. *PhysChemComm 2*, 45.

(13) Leermakers, F. A. M., and Lyklema, J. (1992) On the self-consistent field theory of surfactant micelles. *Colloids and Surfaces 67*, 239–255.

(14) Damm, W., Frontera, A., Tirado–Rives, J., and Jorgensen, W. L. (1997) OPLS all-atom force field for carbohydrates. *J. Comput. Chem. 18*, 1955–1970.

(15) Kony, D., Damm, W., Stoll, S., and Van Gunsteren, W. F. (2002) An improved OPLS-AA force field for carbohydrates. *J. Comput. Chem. 23*, 1416–29.

(16) Molinero, V., and Goddard, W. A. (2004) M3B: A Coarse Grain Force Field for Molecular Simulations of Malto-Oligosaccharides and Their Water Mixtures. *J. Phys. Chem. B 108*, 1414–1427.

(17) Brown, D., Marcadon, V., Mélé, P., and Albérola, N. D. (2008) Effect of Filler Particle Size on the Properties of Model Nanocomposites. *Macromolecules 41*, 1499–1511.

(18) Jacobs, M. N. (2004) In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology 205*, 43–53.

(19) Schmidt, R. K., Tasaki, K., and Brady, J. W. (1994) Computer modeling studies of the interaction of water with carbohydrates. *J. Food Eng. 22*, 43–57.

(20) Limbach, H. J., and Kremer, K. (2006) Multi-scale modelling of polymers: Perspectives for food materials. *Trends Food Sci. Technol. 17*, 215–219.

(21) Limbach, H. J., and Ubbink, J. (2008) Structure and dynamics of maltooligomer–water solutions and glasses. *Soft Matter 4*, 1887.

(22) Fransson, S., Peleg, O., Lorén, N., Hermansson, A.-M., and Kröger, M. (2010) Modelling and confocal microscopy of biopolymer mixtures in confined geometries. *Soft Matter 6*, 2713.

(23) Leermakers, F. A. M., Atkinson, P. J., Dickinson, E., and Horne, D. S. (1996) Self-Consistent-Field Modeling of Adsorbed β-Casein: Effects of pH and Ionic Strength on Surface Coverage and Density Profile. *J. Colloid Interface Sci. 178*, 681–693.

(24) Lee, W. B., Mezzenga, R., and Fredrickson, G. H. (2007) Anomalous phase sequences in lyotropic liquid crystals. *Phys. Rev. Lett. 99*, 187801.

(25) Ettelaie, R., Akinshina, A., and Dickinson, E. (2008) Mixed protein–polysaccharide interfacial layers: a self consistent field calculation study. *Faraday Discuss. 139*, 161.

(26) Dickinson, E., Euston, S. R., and Woskett, C. M. (1990) Competitive adsorption of food macromolecules and surfactants at the oil-water interface, in *Surfactants and Macromolecules: Self-Assembly at Interfaces and in Bulk SE - 9* (Lindman, B., Rosenholm, J. B., and Stenius, P., Eds.), pp 65–75. Steinkopff.

(27) Gránásy, L., Börzsönyi, T., and Pusztai, T. (2002) Nucleation and bulk crystallization in binary phase field theory. *Phys. Rev. Lett. 88*, 206105.

(28) Dickinson, E., and Euston, S. R. (1992) Computer simulation model of the adsorption of protein—polysaccharide complexes. *Food Hydrocoll. 6*, 345–357.

(29) Pugnaloni, L. A., Ettelaie, R., and Dickinson, E. (2003) Growth and aggregation of surfactant islands during the displacement of an adsorbed protein monolayer: a Brownian dynamics simulation study. *Colloids Surfaces B Biointerfaces 31*, 149–157.

(30) Bos, M. T. A., and van Opheusden, J. H. J. (1996) Brownian dynamics simulation of gelation and aging in interacting colloidal systems. *Phys. Rev. E 53*, 5044–5050.

(31) Dickinson, E. (2001) Milk protein interfacial layers and the relationship to emulsion stability and rheology. *Colloids Surfaces B Biointerfaces 20*, 197–210.

(32) Van Steijn, V., Kreutzer, M. T., and Kleijn, C. R. (2007) -PIV study of the formation of segmented flow in microfluidic T-junctions. *Chem. Eng. Sci. 62*, 7505–7514.

(33) Kromkamp, J., Van den Ende, D. T. M., Kandhai, D., Van der Sman, R. G. M., and Boom, R. M. (2005) Shear-induced self-diffusion and microstructure in non-Brownian suspensions at non-zero Reynolds numbers. *J. Fluid Mech. 529*, 253–278.

(34) Datta, A. K. (2007) Porous media approaches to studying simultaneous heat and mass transfer in food processes. I: Problem formulations. *J. Food Eng. 80*, 80–95.

(35) Gompper, G., Dhont, J. K. G., and Richter, D. (2008) A unified view of soft matter systems? *Eur. Phys. J. E. Soft Matter 26*, 1–2.

(36) Ginex, T., Seira, C., Estarellas, C., Bidon-Chanal, A. Luque, F. J. (2015) Molecular Dynamics: An Emerging Tool For Exploring the Molecular Events in Food Research, in *From Medicinal Chemistry to Food Science: A Transfer of In Silico Methods Applications* (Cozzini, P., Ed.), p 7x10 – (NBC–C). Nova Publishers.

(37) Ramalakshmi, K., and Raghavan, B. (1999) Caffeine in coffee: its removal. Why and how? *Crit. Rev. Food Sci. Nutr. 39*, 441–56.

(38) Cesaro, A., Russo, E., and Crescenzi, V. (1976) Thermodynamics of caffeine aqueous solutions. *J. Phys. Chem. 80*, 335–339.

(39) Falk, M., Chew, W., Walter, J. A., Kwiatkowski, W., Barclay, K. D., and Klassen, G. A. (1998) Molecular modelling and NMR studies of the caffeine dimer. *Can. J. Chem. 76*, 48–56.

(40) Lilley, T. H., Linsdell, H., and Maestre, A. (1992) Association of caffeine in water and in aqueous solutions of sucrose. *J. Chem. Soc. Faraday Trans. 88*, 2865.

(41) Fritzsche, H., Petri, I., Schütz, H., Weller, K., Sedmera, P., and Lang, H. (1980) On the interaction of caffeine with nucleic acids. III. 1H NMR studies of caffeine--5'-adenosine monophosphate and caffeine-poly(riboadenylate) interactions. *Biophys. Chem. 11*, 109–19.

(42) Davies, D. B., Veselkov, D. A., Djimant, L. N., and Veselkov, A. N. (2001) Hetero-association of caffeine and aromatic drugs and their competitive binding with a DNA oligomer. *Eur. Biophys. J. 30*, 354–66.

(43) Al-Maaieh, A., and Flanagan, D. R. (2002) Salt effects on caffeine solubility, distribution, and self-association. *J. Pharm. Sci. 91*, 1000–8.

(44) Tavagnacco, L., Schnupf, U., Mason, P. E., Saboungi, M.-L., Cesàro, A., and Brady, J. W. (2011) Molecular dynamics simulation studies of caffeine aggregation in aqueous solution. *J. Phys. Chem. B 115*, 10957–66.

(45) Sharma, B., and Paul, S. (2013) Effects of dilute aqueous NaCl solution on caffeine aggregation. *J. Chem. Phys. 139*, 194504.

(46) Tavagnacco, L., Engström, O., Schnupf, U., Saboungi, M.-L., Himmel, M., Widmalm, G., Cesàro, A., and Brady, J. W. (2012) Caffeine and sugars interact in aqueous solutions: a simulation and NMR study. *J. Phys. Chem. B 116*, 11701–11.

(47) Hendon, C. H., Colonna-Dashwood, L., and Colonna-Dashwood, M. (2014) The role of dissolved cations in coffee extraction. *J. Agric. Food Chem. 62*, 4947–50.

(48) Draget, K. I.; Moe, S. T.; Skjåk-Braek, G.; Smidsrød, O. (2006) Alginates, in *Food Polysaccharides and Their Applications* (Stephen, A. M., A., and Phillips, G. O., Eds.). CRC Press.

(49) Brownlee, I. A., Seal, C. J., Wilcox, M., Dettmar, P. W., and Pearson, J. P. (2009) Alginates: Biology and Applications (Rehm, B. H. A., Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg.

(50) Herrero, E. P., Martín Del Valle, E. M., and Peppas, N. A. (2010) Protein Imprinting by Means of Alginate-Based Polymer Microcapsules. *Ind. Eng. Chem. Res. 49*, 9811–9814.

(51) Li, Q., Liu, C.-G., Huang, Z.-H., and Xue, F.-F. (2011) Preparation and characterization of nanoparticles based on hydrophobic alginate derivative as carriers for sustained release of vitamin D3. *J. Agric. Food Chem. 59*, 1962–7.

(52) Song, S., Wang, Z., Qian, Y., Zhang, L., and Luo, E. (2012) The release rate of curcumin from calcium alginate beads regulated by food emulsifiers. *J. Agric. Food Chem. 60*, 4388–95.

(53) Zhang, Y., Jia, X., Wang, L., Liu, J., and Ma, G. (2011) Preparation of Ca-Alginate Microparticles and Its Application for Phenylketonuria Oral Therapy. *Ind. Eng. Chem. Res. 50*, 4106–4112.

(54) Georg Jensen, M., Knudsen, J. C., Viereck, N., Kristensen, M., and Astrup, A. (2012) Functionality of alginate based supplements for application in human appetite regulation. *Food Chem. 132*, 823–829.

(55) Lu, L., Liu, X., Tong, Z., and Gao, Q. (2006) Critical exponents and self-similarity for sol-gel transition in aqueous alginate systems induced by in situ release of calcium cations. *J. Phys. Chem. B 110*, 25013–20.

(56) Fang, Y., Al-Assaf, S., Phillips, G. O., Nishinari, K., Funami, T., Williams, P. A., and Li, L. (2007) Multiple steps and critical behaviors of the binding of calcium to alginate. *J. Phys. Chem. B 111*, 2456–62.

(57) Harper, B. A., Barbut, S., Lim, L.-T., and Marcone, M. F. (2014) Effect of various gelling cations on the physical properties of “wet” alginate films. *J. Food Sci. 79*, E562–7.

(58) Vega, C.; Castells, P. (2012) Spherification. The Kitchen as the Laboratory, in *The Kitchen as the Laboratory* (Vega, C.; Ubbink, J.; van der Linden, E., Ed.), pp 25–32. Columbia University Press.

(59) Grant, G. T., Morris, E. R., Rees, D. A., Smith, P. J. C., and Thom, D. (1973) Biological interactions between polysaccharides and divalent cations: The egg-box model. *FEBS Lett. 32*, 195–198.

(60) Braccini, I., and Pérez, S. (2001) Molecular basis of C(2+)-induced gelation in alginates and pectins: the egg-box model revisited. *Biomacromolecules 2*, 1089–96.

(61) Plazinski, W. (2011) Molecular basis of calcium binding by polyguluronate chains. Revising the egg-box model. *J. Comput. Chem. 32*, 2988–95.

(62) Borgogna, M., Skjåk-Bræk, G., Paoletti, S., and Donati, I. (2013) On the initial binding of alginate by calcium ions. The tilted egg-box hypothesis. *J. Phys. Chem. B 117*, 7277–82.

(63) Plazinski, W., and Drach, M. (2013) Calcium-α-L-guluronate complexes: Ca2+ binding modes from DFT-MD simulations. *J. Phys. Chem. B 117*, 12105–12.

(64) Fu, H., Liu, Y., Adrià, F., Shao, X., Cai, W., and Chipot, C. (2014) From material science to avant-garde cuisine. The art of shaping liquids into spheres. *J. Phys. Chem. B 118*, 11747–56.

(65) Zaveri, N. T. (2006) Green tea and its polyphenolic catechins: medicinal uses in cancer and noncancer applications. *Life Sci. 78*, 2073–80.

(66) Del Rio, D., Rodriguez-Mateos, A., Spencer, J. P. E., Tognolini, M., Borges, G., and Crozier, A. (2013) Dietary (poly)phenolics in human health: structures, bioavailability, and evidence of protective effects against chronic diseases. *Antioxid. Redox Signal. 18*, 1818–92.

(67) Del Rio, D., Stewart, A. J., Mullen, W., Burns, J., Lean, M. E. J., Brighenti, F., and Crozier, A. (2004) HPLC-MSn analysis of phenolic compounds and purine alkaloids in green and black tea. *J. Agric. Food Chem. 52*, 2807–15.

(68) Chung, F.-L., Schwartz, J., Herzog, C. R., and Yang, Y.-M. (2003) Tea and cancer prevention: studies in animals and humans. *J. Nutr. 133*, 3268S–3274S.

(69) Uekusa, Y., Kamihira, M., and Nakayama, T. (2007) Dynamic behavior of tea catechins interacting with lipid membranes as determined by NMR spectroscopy. *J. Agric. Food Chem. 55*, 9986–92.

(70) Tamba, Y., Ohba, S., Kubota, M., Yoshioka, H., Yoshioka, H., and Yamazaki, M. (2007) Single GUV method reveals interaction of tea catechin (-)-epigallocatechin gallate with lipid membranes. *Biophys. J. 92*, 3178–94.

(71) Caturla, N., Vera-Samper, E., Villalaín, J., Mateo, C. R., and Micol, V. (2003) The relationship between the antioxidant and the antibacterial properties of galloylated catechins and the structure of phospholipid model membranes. *Free Radic. Biol. Med. 34*, 648–62.

(72) Sirk, T. W., Brown, E. F., Sum, A. K., and Friedman, M. (2008) Molecular dynamics study on the biophysical interactions of seven green tea catechins with lipid bilayers of cell membranes. *J. Agric. Food Chem. 56*, 7750–8.

(73) Nagle, D. G., Ferreira, D., and Zhou, Y.-D. (2006) Epigallocatechin-3-gallate (EGCG): chemical and biomedical perspectives. *Phytochemistry 67*, 1849–55.

(74) Bennett, W. F. D., and Tieleman, D. P. (2014) The importance of membrane defects-lessons from simulations. *Acc. Chem. Res. 47*, 2244–51.

(75) Martinez-Seara, H., Róg, T., Karttunen, M., Vattulainen, I., and Reigada, R. (2010) Cholesterol induces specific spatial and orientational order in cholesterol/phospholipid membranes. *PLoS One 5*, e11162.

(76) Bennett, W. F. D., MacCallum, J. L., and Tieleman, D. P. (2009) Thermodynamic analysis of the effect of cholesterol on dipalmitoylphosphatidylcholine lipid membranes. *J. Am. Chem. Soc. 131*, 1972–8.

(77) Zhu, Q., Cheng, K. H., and Vaughn, M. W. (2007) Molecular dynamics studies of the molecular structure and interactions of cholesterol superlattices and random domains in an unsaturated phosphatidylcholine bilayer membrane. *J. Phys. Chem. B 111*, 11021–31.

(78) Mondal, S., and Mukhopadhyay, C. (2007) Molecular insight of specific cholesterol interactions: A molecular dynamics simulation study. *Chem. Phys. Lett. 439*, 166–170.

(79) Alwarawrah, M., Dai, J., and Huang, J. (2010) A molecular view of the cholesterol condensing effect in DOPC lipid bilayers. *J. Phys. Chem. B 114*, 7516–23.

(80) Hakobyan, D., and Heuer, A. (2014) Key molecular requirements for raft formation in lipid/cholesterol membranes. *PLoS One 9*, e87369.

(81) Khelashvili, G., Johner, N., Zhao, G., Harries, D., and Scott, H. L. (2014) Molecular origins of bending rigidity in lipids with isolated and conjugated double bonds: the effect of cholesterol. *Chem. Phys. Lipids 178*, 18–26.

(82) Pan, J., Cheng, X., Heberle, F. A., Mostofian, B., Kučerka, N., Drazba, P., and Katsaras, J. (2012) Interactions between ether phospholipids and cholesterol as determined by scattering and molecular dynamics simulations. *J. Phys. Chem. B 116*, 14829–38.

(83) Hsieh, C.J., Chen, Y.W., and Hwang, D. W. (2013) Effects of cholesterol on membrane molecular dynamics studied by fast field cycling NMR relaxometry. *Phys. Chem. Chem. Phys. 15*, 16634–40.

(84) Rosetti, C., and Pastorino, C. (2012) Comparison of ternary bilayer mixtures with asymmetric or symmetric unsaturated phosphatidylcholine lipids by coarse grained molecular dynamics simulations. *J. Phys. Chem. B 116*, 3525–37.

(85) Plesnar, E., Subczynski, W. K., and Pasenkiewicz-Gierula, M. (2012) Saturation with cholesterol increases vertical order and smoothes the surface of the phosphatidylcholine bilayer: a molecular simulation study. *Biochim. Biophys. Acta 1818*, 520–9.

(86) Sugár, I. P., and Chong, P. L.-G. (2012) A statistical mechanical model of cholesterol/phospholipid mixtures: linking condensed complexes, superlattices, and the phase diagram. *J. Am. Chem. Soc. 134*, 1164–71.

(87) Choubey, A., Nomura, K., Kalia, R., Nakano, A., and Vashishta, P. (2012) Cholesterol Flip-Flop Dynamics in a Phospholipid Bilayer: All Atom Molecular Dynamics Simulations. *Biophys. J. 102*, 241a.

(88) Jo, S., Rui, H., Lim, J. B., Klauda, J. B., and Im, W. (2010) Cholesterol flip-flop: insights from free energy simulation studies. *J. Phys. Chem. B 114*, 13342–8.

(89) Plesnar, E., Subczynski, W. K., and Pasenkiewicz-Gierula, M. (2013) Comparative computer simulation study of cholesterol in hydrated unary and binary lipid bilayers and in an anhydrous crystal. *J. Phys. Chem. B 117*, 8758–69.

(90) Saito, H., and Shinoda, W. (2011) Cholesterol effect on water permeability through DPPC and PSM lipid bilayers: a molecular dynamics study. *J. Phys. Chem. B 115*, 15241–50.

(91) Shigematsu, T., Koshiyama, K., and Wada, S. (2014) Molecular dynamics simulations of pore formation in stretched phospholipid/cholesterol bilayers. *Chem. Phys. Lipids 183*, 43–9.

(92) Xu, W., Wei, G., Su, H., Nordenskiöld, L., and Mu, Y. (2011) Effects of cholesterol on pore formation in lipid bilayers induced by human islet amyloid polypeptide fragments: a coarse-grained molecular dynamics study. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. 84*, 051922.

(93) Fernández, M. L., Marshall, G., Sagués, F., and Reigada, R. (2010) Structural and kinetic molecular dynamics study of electroporation in cholesterol-containing bilayers. *J. Phys. Chem. B 114*, 6855–65.

(94) Pourmousa, M., Róg, T., Mikkeli, R., Vattulainen, L., Solanko, L. M., Wüstner, D., List, N. H., Kongsted, J., and Karttunen, M. (2014) Dehydroergosterol as an analogue for cholesterol: why it mimics cholesterol so well-or does it? *J. Phys. Chem. B 118*, 7345–57.

(95) Robalo, J. R., do Canto, A. M. T. M., Carvalho, A. J. P., Ramalho, J. P. P., and Loura, L. M. S. (2013) Behavior of fluorescent cholesterol analogues dehydroergosterol and cholestatrienol in lipid bilayers: a molecular dynamics study. *J. Phys. Chem. B 117*, 5806–19.

(96) Cournia, Z., Ullmann, G. M., and Smith, J. C. (2007) Differential effects of cholesterol, ergosterol and lanosterol on a dipalmitoyl phosphatidylcholine membrane: a molecular dynamics simulation study. *J. Phys. Chem. B 111*, 1786–801.

(97) Czub, J., and Baginski, M. (2006) Modulation of amphotericin B membrane interaction by cholesterol and ergosterol--a molecular dynamics study. *J. Phys. Chem. B 110*, 16743–53.

(98) Czub, J., and Baginski, M. (2006) Comparative molecular dynamics study of lipid membranes containing cholesterol and ergosterol. *Biophys. J. 90*, 2368–82.

(99) Robalo, J. R., Ramalho, J. P. P., and Loura, L. M. S. (2013) NBD-labeled cholesterol analogues in phospholipid bilayers: insights from molecular dynamics. *J. Phys. Chem. B 117*, 13731–42.

(100) Róg, T., Pasenkiewicz-Gierula, M., Vattulainen, I., and Karttunen, M. (2009) Ordering effects of cholesterol and its analogues. *Biochim. Biophys. Acta 1788*, 97–121.

(101) Róg, T., Vattulainen, I., Jansen, M., Ikonen, E., and Karttunen, M. (2008) Comparison of cholesterol and its direct precursors along the biosynthetic pathway: effects of cholesterol, desmosterol and 7-dehydrocholesterol on saturated and unsaturated lipid bilayers. *J. Chem. Phys. 129*, 154508.

(102) Oates, J., and Watts, A. (2011) Uncovering the intimate relationship between lipids, cholesterol and GPCR activation. *Curr. Opin. Struct. Biol. 21*, 802–7.

(103) Hanson, M. A., Cherezov, V., Griffith, M. T., Roth, C. B., Jaakola, V.-P., Chien, E. Y. T., Velasquez, J., Kuhn, P., and Stevens, R. C. (2008) A specific cholesterol binding site is established by the 2.8 A structure of the human beta2-adrenergic receptor. *Structure 16*, 897–905.

(104) Khelashvili, G., Grossfield, A., Feller, S. E., Pitman, M. C., and Weinstein, H. (2009) Structural and dynamic effects of cholesterol at preferred sites of interaction with rhodopsin identified from microsecond length molecular dynamics simulations. *Proteins 76*, 403–17.

(105) Hibbs, R. E., and Gouaux, E. (2011) Principles of activation and permeation in an anion-selective Cys-loop receptor. *Nature 474*, 54–60.

(106) Cheng, M. H., Xu, Y., and Tang, P. (2009) Anionic lipid and cholesterol interactions with alpha4beta2 nAChR: insights from MD simulations. *J. Phys. Chem. B 113*, 6964–70.

(107) Hénin, J., Salari, R., Murlidaran, S., and Brannigan, G. (2014) A predicted binding site for cholesterol on the GABAA receptor. *Biophys. J. 106*, 1938–49.

(108) Buck, L., and Axel, R. (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell 65*, 175–87.

(109) Kaupp, U. B. (2010) Olfactory signalling in vertebrates and insects: differences and commonalities. *Nat. Rev. Neurosci. 11*, 188–200.

(110) Crasto, C. J. (2009) Computational Biology of Olfactory Receptors. *Curr. Bioinform. 4*, 8–15.

(111) Lai, P. C., and Crasto, C. J. (2012) Beyond modeling: all-atom olfactory receptor model simulations. *Front. Genet. 3*, 61.

(112) Don, C. G., and Riniker, S. (2014) Scents and sense: in silico perspectives on olfactory receptors. *J. Comput. Chem. 35*, 2279–87.

(113) Singer, M. S. (2000) Analysis of the Molecular Basis for Octanal Interactions in the Expressed Rat I7 Olfactory Receptor. *Chem. Senses 25*, 155–165.

(114) Lai, P. C., Singer, M. S., and Crasto, C. J. (2005) Structural activation pathways from dynamic olfactory receptor-odorant interactions. *Chem. Senses 30*, 781–92.

(115) Araneda, R. C., Kini, A. D., and Firestein, S. (2000) The molecular receptive range of an odorant receptor. *Nat. Neurosci. 3*, 1248–55.

(116) Anselmi, C., Buonocore, A., Centini, M., Facino, R. M., and Hatt, H. (2011) The human olfactory receptor 17-40: requisites for fitting into the binding pocket. *Comput. Biol. Chem. 35*, 159–68.

(118) Charlier, L., Topin, J., Ronin, C., Kim, S.-K., Goddard, W. A., Efremov, R., and Golebiowski, J. (2012) How broadly tuned olfactory receptors equally recognize their agonists. Human OR1G1 as a test case. *Cell. Mol. Life Sci. 69*, 4205–13.

(119) Lai, P. C., Guida, B., Shi, J., and Crasto, C. J. (2014) Preferential binding of an odor within olfactory receptors: a precursor to receptor activation. *Chem. Senses 39*, 107–23.

(120) Becker, O. M., MacKerell Jr, A. D., Roux, B., & Watanabe, M. (2001) Computational biochemistry and biophysics (Becker, O. M., MacKerell Jr, A. D., Roux, B., & Watanabe, M., Ed.). CRC Press.

(121) Leach, A. R. (2001) Molecular modelling: principles and applications (Leach, A. R., Ed.). Pearson education.

(122) Schlick, T. (2002) Molecular Modeling and Simulation. An interdisciplinary Guide (Schlick, T., Ed.). Springer-Verlag New York, Inc. New York.

(123) Frenkel, D. and Smit, B. (2002) Understanding Molecular Simulation (Frenkel, D. and Smit, B., Ed.). Academic Press.

(124) Allinger, N.L., Yuh, Y. H., and Lii, J.-H. (1989) Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1. J. Am. Chem. Soc 111, 8551-8565.

(125) Hwang, M. J., Stockfisch, T. P., and Hagler, A. T. (1994) Derivation of Class II Force Fields. 2. Derivation and Characterization of a Class II Force Field, CFF93, for the Alkyl Functional Group and Alkane Molecules. *J. Am. Chem. Soc. 116*, 2515–2525.

(126) Cramer, C. J. (2013) Essentials of computational chemistry: theories and models (Cramer, C. J., Ed.). John Wiley & Sons.

(127) MacKerell Jr, A. D., Wiorkiewicz-Kuczera, J., and Karplus, M. (1995) An all-atom empirical energy function for the simulation of nucleic acids. *J. Am. Chem. Soc. 117*, 11946–11975.

(128) Schlenkrich, M., Brickmann, J., MacKerell Jr, A. D., and Karplus, M. (1996) An empirical potential energy function for phospholipids: criteria for parameter optimization and applications. *Biol. Membr.* 31–81.

(129) MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B 102*, 3586–616.

(130) Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., ... & Kollman, P. A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc. 117*, 5179–5197.

(131) Halgren, T. A. (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem. 17*, 490–519.

(132) Allinger, N. L. (1977) Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc. 99*, 8127–8134.

(133) Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem. 26*, 1668–88.

(134) Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins 65*, 712–25.

(135) Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins 78*, 1950–8.

(136) Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T. E., Laughton, C. A., and Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J. 92*, 3817–29.

(137) Zgarbová, M., Otyepka, M., Sponer, J., Mládek, A., Banáš, P., Cheatham, T. E., and Jurečka, P. (2011) Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput. 7*, 2886–2902.

(138) Yildirim, I., Stern, H. A., Kennedy, S. D., Tubbs, J. D., and Turner, D. H. (2010) Reparameterization of RNA chi Torsion Parameters for the AMBER Force Field and Comparison to NMR Spectra for Cytidine and Uridine. *J. Chem. Theory Comput. 6*, 1520–1531.

(139) Zgarbová, M., Luque, F. J., Sponer, J., Cheatham, T. E., Otyepka, M., and Jurečka, P. (2013) Toward Improved Description of DNA Backbone: Revisiting Epsilon and Zeta Torsion Force Field Parameters. *J. Chem. Theory Comput. 9*, 2339–2354.

(140) Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., and Woods, R. J. (2008) GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem. 29*, 622–55.

(141) Tessier, M. B., Demarco, M. L., Yongye, A. B., and Woods, R. J. (2008) Extension of the GLYCAM06 Biomolecular Force Field to Lipids, Lipid Bilayers and Glycolipids. *Mol. Simul. 34*, 349–363.

(142) Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004) Development and testing of a general amber force field. *J. Comput. Chem. 25*, 1157–74.

(143) Zhu, X., Lopes, P. E. M., and Mackerell, A. D. (2012) Recent Developments and Applications of the CHARMM force fields. *Wiley Interdiscip. Rev. Comput. Mol. Sci. 2*, 167–185.

(144) Poger, D., Van Gunsteren, W. F., and Mark, A. E. (2010) A new force field for simulating phosphatidylcholine bilayers. *J. Comput. Chem. 31*, 1117–25.

(145) Schmid, N., Eichenberger, A. P., Choutko, A., Riniker, S., Winger, M., Mark, A. E., and van Gunsteren, W. F. (2011) Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J. 40*, 843–56.

(146) Luque, F. J., Dehez, F., Chipot, C., and Orozco, M. (2011) Polarization effects in molecular interactions. *Wiley Interdiscip. Rev. Comput. Mol. Sci. 1*, 844–854.

(147) Anisimov, V. M., Lamoureux, G., Vorobyov, I. V., Huang, N., Roux, B., and MacKerell, A. D. (2005) Determination of Electrostatic Parameters for a Polarizable Force Field Based on the Classical Drude Oscillator. *J. Chem. Theory Comput. 1*, 153–168.

(148) Baker, C. M., Anisimov, V. M., and MacKerell, A. D. (2011) Development of CHARMM polarizable force field for nucleic acid bases based on the classical Drude oscillator model. *J. Phys. Chem. B 115*, 580–96.

(149) Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Schnieders, M. J., Haque, I., Mobley, D. L., Lambrecht, D. S., DiStasio, R. A., Head-Gordon, M., Clark, G. N. I., Johnson, M. E., and Head-Gordon, T. (2010) Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B 114*, 2549–64.

(150) Wang, J., Cieplak, P., Li, J., Hou, T., Luo, R., and Duan, Y. (2011) Development of polarizable models for molecular mechanical calculations I: parameterization of atomic polarizability. *J. Phys. Chem. B 115*, 3091–9.

(151) Wang, J., Cieplak, P., Li, J., Cai, Q., Hsieh, M.-J., Luo, R., and Duan, Y. (2012) Development of polarizable models for molecular mechanical calculations. 4. van der Waals parametrization. *J. Phys. Chem. B 116*, 7088–101.

(152) Harder, E., Mackerell, A. D., and Roux, B. (2009) Many-body polarization effects and the membrane dipole potential. *J. Am. Chem. Soc. 131*, 2760–1.

(153) Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. . (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys. 23*, 327–341.

(154) Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997) LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem. 18*, 1463–1472.

(155) Sagui, C., and Darden, T. A. (1999) Molecular dynamics simulations of biomolecules: long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct. 28*, 155–79.

(156) Rocchia, W.; Masetti, M.; Cavalli, A. (2012) Enhanced Sampling Methods in Drug Design., in *Physico-Chemical and Computational Approaches to Drug Discovery.* (Luque, F. J.; Barril, X., Ed.), pp 273–301. Royal Society of Chemistry, Drug Discovery Series, No. 23, Cambridge.

(158) Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem. 26*, 1781–802.

(159) Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B., and Lindahl, E. (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics 29*, 845–54.

(160) Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., and Shaw, D. E. (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys. 41*, 429–52.

(161) Götz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S., and Walker, R. C. (2012) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput. 8*, 1542–1555.

(162) Le Grand, S., Götz, A. W., and Walker, R. C. (2013) SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun. 184*, 374–380.

(163) Harvey, M. J., Giupponi, G., and Fabritiis, G. De. (2009) ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput. 5*, 1632–1639.

(164) Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., and Wriggers, W. (2010) Atomic-level characterization of the structural dynamics of proteins. *Science 330*, 341–6.

(165) Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011) How Fast-Folding Proteins Fold. *Science (80-.). 334*, 517–520.

(166) Dror, R. O., Pan, A. C., Arlow, D. H., Borhani, D. W., Maragakis, P., Shan, Y., Xu, H., and Shaw, D. E. (2011) Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci. U. S. A. 108*, 13118–23.

(167) Shan, Y., Kim, E. T., Eastwood, M. P., Dror, R. O., Seeliger, M. A., and Shaw, D. E. (2011) How does a drug molecule find its target binding site? *J. Am. Chem. Soc. 133*, 9181–3.

(168) Buch, I., Giorgino, T., and De Fabritiis, G. (2011) Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A. 108*, 10184–9.

(169) Nieve, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. (2004) Coarse Grain Models and the Computer Simulation of Soft Materials. *J. Phys. Condens. Matter 16*, R481–R512.

(170) Baaden, M., and Marrink, S. J. (2013) Coarse-grain modelling of protein-protein interactions. *Curr. Opin. Struct. Biol. 23*, 878–86.

(171) Darré, L., Machado, M. R., and Pantano, S. (2012) Coarse-grained models of water. *Wiley Interdiscip. Rev. Comput. Mol. Sci. 2*, 921–930.

(172) Dans, P.; Zeida, A.; Machado, M. R.; Pantano, S. (2010) A Coarse Grained Model for Atomic-detailed DNA Simulations with Explicit Electrostatics. *J. Chem. Theory Comput.* 6, 1711-1715.

(5173) Tepper, H. L., and Voth, G. A. (2005) A coarse-grained model for double-helix molecules in solution: spontaneous helix formation and equilibrium properties. *J. Chem. Phys. 122*, 124906.

(174) Tozzini, V. (2005) Coarse-grained models for proteins. *Curr. Opin. Struct. Biol. 15*, 144–50.

(175) Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and de Vries, A. H. (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B 111*, 7812–24.

(176) Shih, A. Y., Freddolino, P. L., Arkhipov, A., and Schulten, K. (2007) Assembly of lipoprotein particles revealed by coarse-grained molecular dynamics simulations. *J. Struct. Biol. 157*, 579–92.

(177) Arkhipov, A., Freddolino, P. L., and Schulten, K. (2006) Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure 14*, 1767–77.

(178) Arkhipov, A., Yin, Y., and Schulten, K. (2008) Four-scale description of membrane sculpting by BAR domains. *Biophys. J. 95*, 2806–21.

(179) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* 2007, 111, 7812-7824.

(180) Marrink, S. J., and Tieleman, D. P. (2013) Perspective on the Martini model. *Chem. Soc. Rev. 42*, 6801–22.

(181) Limbach, H. J., Arnold, A., Mann, B. A., and Holm, C. (2006) ESPResSo—an extensible simulation package for research on soft matter systems. *Comput. Phys. Commun. 174*, 704–727.

(182) Reynwar, B. J., Illya, G., Harmandaris, V. A., Müller, M. M., Kremer, K., and Deserno, M. (2007) Aggregation and vesiculation of membrane proteins by curvature-mediated interactions. *Nature 447*, 461–4.

(183) Selassie, C., & Verma, R. P. (2003). History of quantitative structure–activity relationships. *Burger's Medicinal Chemistry and Drug Discovery* (Abraham, D. J., Ed.) sixth edit. John Wiley & Sons, Inc., Hoboken, NJ, USA.

(184) Ramsden, C.A., E. (1990) Quantitative Drug Design (Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds) (Hansch. C., Sammers, P.G., and Taylor. J. B., E., Ed.). Pergamon Press.

(185) Kubinyi, H. (1993) QSAR: Hansch Analysis and Related Approaches. Wiley-VCH Verlag GmbH, Weinheim, Germany.

(186) Hansch, C. & Leo, A. (1995) Exploring QSAR. Fundamentals and Applications in Chemistry and Biology, ACS.

(187) Hansch, C., Maloney, P. P., Fujita, T., & Muir R. M. (1962) Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature 194*, 178–180.

(188) Hammett, L. P. (1970) Physical Organic Chemistry. McGraw-Hill, New York.

(189) Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc. 110*, 5959–67.

(190) Salum, L. B., and Andricopulo, A. D. (2009) Fragment-based QSAR: perspectives in drug design. *Mol. Divers. 13*, 277–85.

(191) Polanski, J., Bak, A., Gieleciak, R., and Magdziarz, T. (2006) Modeling robust QSAR. *J. Chem. Inf. Model. 46*, 2310–8.

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

(193) Sutherland, J. J., O'Brien, L. A., and Weaver, D. F. (2004) A comparison of methods for modeling quantitative structure-activity relationships. *J. Med. Chem. 47*, 5541–54.

(194) Dudek, A., Arodz, T., and Galvez, J. (2006) Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. *Comb. Chem. High Throughput Screen. 9*, 213–228.

(195) Lemmen, C., and Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput. Aided. Mol. Des. 14*, 215–232.

(196) Polanski, J., and Bak, A. (2003) Modeling steric and electronic effects in 3D- and 4D-QSAR schemes: predicting benzoic pK(a) values and steroid CBG binding affinities. *J. Chem. Inf. Comput. Sci. 43*, 2081–92.

(197) Vedani, A., Zbinden, P., Snyder, J. P., & Greenidge, P. A. (1995). Pseudoreceptor modeling: The construction of three-dimensional receptor surrogates. *Journal of the American Chemical Society*, *117*(17), 4987-4994.

(198) Vedani, A., Descloux, A.-V., Spreafico, M., and Ernst, B. (2007) Predicting the toxic potential of drugs and chemicals in silico: a model for the peroxisome proliferator-activated receptor gamma (PPAR gamma). *Toxicol. Lett. 173*, 17–23.

(199) Mannhold, Raimund, Hugo Kubinyi, and G. F. (2006) Pharmacophores and Pharmacophore Searches (Langer, T., and Hoffmann, R. D., Eds.). Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG.

(200) Alexander, D. L. J., Tropsha, A., and Winkler, D. A. (2015) Beware of R(2): Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model. 55*, 1316–22.

(201) Berman, Helen M., et al. (2000) The protein data bank. *Nucleic acids research* 28, 1, 235-242.

(202) Huang, N., Shoichet, B. K., and Irwin, J. J. (2006) Benchmarking sets for molecular docking. *J. Med. Chem. 49*, 6789–801.

(203) Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res. 34*, D668–72.

(204) Irwin, J. J., and Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model. 45*, 177–82.

(205) Goodford, P. J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem. 28*, 849–857.

(206) DesJarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D., and Venkataraghavan, R. (1988) Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem. 31*, 722–729.

(207) Taylor, R. D., Jewsbury, P. J., and Essex, J. W. A review of protein-small molecule docking methods. *J. Comput. Aided. Mol. Des. 16*, 151–166.

(208) Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem. 47*, 1739–49.

(209) Kramer, B., Metz, G., Rarey, M., and Lengauer, T. (1999) Ligand docking and screening with FLEXX *9*, 463–478.

(210) Oshiro, C. M., Kuntz, I. D., and Dixon, J. S. (1995) Flexible ligand docking using a genetic algorithm. *J. Comput. Aided. Mol. Des. 9*, 113–130.

(211) Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997) Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol. 267*, 727–48.

(212) Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem. 19*, 1639–1662.

(213) Charifson, P. S., Corkery, J. J., Murcko, M. A., and Walters, W. P. (1999) Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem. 42*, 5100–9.

(214) Brylinski, M. (2013) Nonlinear scoring functions for similarity-based ligand docking and binding affinity prediction. *J. Chem. Inf. Model. 53*, 3097–112.

(215) Gao, C., Thorsteinson, N., Watson, I., Wang, J., and Vieth, M. (2015) Knowledge-Based Strategy to Improve Ligand Pose Prediction Accuracy for Lead Optimization. *J. Chem. Inf. Model. 55*, 1460–8.

(216) Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A., and Vieth, M. (2004) Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem. 47*, 45–55.

(217) Slynko, I., Scharfe, M., Rumpf, T., Eib, J., Metzger, E., Schüle, R., Jung, M., and Sippl, W. (2014) Virtual screening of PRK1 inhibitors: ensemble docking, rescoring using binding free energy calculation and QSAR model development. *J. Chem. Inf. Model. 54*, 138–50.

(218) Wang, J., Morin, P., Wang, W., and Kollman, P. A. (2001) Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc. 123*, 5221–30.

(219) Gilson, M. K., and Zhou, H.-X. (2007) Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct. 36*, 21–42.

(220) Mobley, D. L., and Dill, K. A. (2009) Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure 17*, 489–98.

(221) Michel, J., and Essex, J. W. (2010) Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aided. Mol. Des. 24*, 639–58.

(222) Hettiarachchy, N. S., & Ziegler, G. R. (1994) Protein functionality in food systems. CRC Press.

(223) Nicholls, A., Sharp, K. A., and Honig, B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins 11*, 281–96.

(224) Eugene Kellogg, G., and Abraham, D. J. Hydrophobicity: is LogP(o/w) more than the sum of its parts? *Eur. J. Med. Chem. 35*, 651–61.

(225) Hansch C., L. A. J. (1979) Substituent Constants for Correlation Analysis in Chemistry and Biology. John Wiley and Sons Inc., NY.

(226) Israelachvili, J., and Pashley, R. (1982) The hydrophobic interaction is long range, decaying exponentially with distance. *Nature 300*, 341–342.

(227) Tirado-Rives, J., and Jorgensen, W. L. (2006) Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J. Med. Chem. 49*, 5880–4.

(228) Monticelli, L.; Salonen, E. Biomolecular Simulations. Methods and Protocols. Volume 924, 2013. Springer Protocols. Humana Press.

(229) Hohenberg, P. (1964) Inhomogeneous Electron Gas. *Phys. Rev. 136*, B864–B871.

(230) Thomas, L. H. (2008) The calculation of atomic fields. *Math. Proc. Cambridge Philos. Soc. 23*, 542.

(231) Fermi, E. (1927) Un metodo statistico per la determinazione di alcune proprietà dell'atomo. *Rend Accad Naz Lincei 6*, 602–607.

(232) Kohn, W., and Sham, L. J. (1965) Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev. 140*, A1133–A1138.

(233) Becke, A. D. (1993) A new mixing of Hartree–Fock and local density-functional theories. *J. Chem. Phys. 98*, 1372.

(234) Pople, J. A., Santry, D. P., and Segal, G. A. (1965) Approximate Self-Consistent Molecular Orbital Theory. I. Invariant Procedures. *J. Chem. Phys. 43*, S129.

(235) Dewar, M. J. S., and Thiel, W. (1977) Ground states of molecules. 38. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc. 99*, 4899–4907.

(236) Dewar MJS, Zoebisch EG, Healy EF, S. J. (1985) Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J Am Chem Soc 107*, 3902–3909.

(237) Stewart, J. J. P. (1989) Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem. 10*, 209–220.

(238) Stewart, J. J. P. (2007) Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J. Mol. Model. 13*, 1173–213.

(239) Ridley, J., and Zerner, M. (1973) An intermediate neglect of differential overlap technique for spectroscopy: Pyrrole and the azines. *Theor. Chim. Acta 32*, 111–134.

(240) Zheng, X., and Stuchebrukhov, A. A. (2003) Electron Tunneling in Proteins: Implementation of ZINDO Model for Tunneling Currents Calculations. *J. Phys. Chem. B 107*, 6621–6628.

(241) Hayashi, T., and Stuchebrukhov, A. A. (2010) Electron tunneling in respiratory complex I. *Proc. Natl. Acad. Sci. U. S. A. 107*, 19157–62.

(242) Ball, D. M., Buda, C., Gillespie, A. M., White, D. P., and Cundari, T. R. (2002) Can semiempirical quantum mechanics be used to predict the spin state of transition metal complexes? An application of de novo prediction. *Inorg. Chem. 41*, 152–6.

(243) Gorelsky, S. I., and Lever, A. B. P. (2001) Electronic structure and spectra of ruthenium diimine complexes by density functional theory and INDO/S. Comparison of the two methods. *J. Organomet. Chem. 635*, 187–196.

(244) Tomasi, J., Mennucci, B., and Cammi, R. (2005) Quantum mechanical continuum solvation models. *Chem. Rev. 105*, 2999–3093.

(245) Miertuš, S., Scrocco, E., and Tomasi, J. (1981) Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chem. Phys. 55*, 117–129.

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

(247) Colominas, C., Luque, F. J., and Orozco, M. (1999) Dimerization of Formamide in Gas Phase and Solution. An Ab Initio MC−MST Study. *J. Phys. Chem. A 103*, 6200–6208.

(248) Luque, J. F., Curutchet, C., Munoz-Muriedas, J., Bidon-Chanal, A., Soteras, I., Morreale, A., Gelpì, J. L., and Orozco, M. (2003) Continuum solvation models: Dissecting the free energy of solvation. *Phys. Chem. Chem. Phys. 5*, 3827.

(249) Morreale, A., Gelpí, J. L., Luque, F. J., and Orozco, M. (2003) Continuum and discrete calculation of fractional contributions to solvation free energy. *J. Comput. Chem. 24*, 1610–23.

(250) Klamt, A., and Schuurmann, G. (1993) COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin Trans. 2* 799.

(251) Klamt, A. (1995) Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem. 99*, 2224–2235.

(252) Chipman, D. M. (2000) Reaction field treatment of charge penetration. *J. Chem. Phys. 112*, 5558–5565.

(253) Luque, F. J., Barril, X., and Orozco, M. Fractional description of free energies of solvation. *J. Comput. Aided. Mol. Des. 13*, 139–152.

(254) Pierotti, R. A. (1976) A scaled particle theory of aqueous and nonaqueous solutions. *Chem. Rev. 76*, 717–726.

(255) Helferich, William, and C. K. W. (2000) Food Toxicology. CRC Press.

(256) Carvalho, Fernando P. (2006) Agriculture, pesticides, food security and food safety. *Environmental science & policy* 9, 685-692.

(257) van Schothorst, M., and Jongeneel, S. (1994) Line monitoring, HACCP and food safety. *Food Control 5*, 107–110.

(258) Tarcsay, Á., and Keseru, G. M. (2011) In silico site of metabolism prediction of cytochrome P450-mediated biotransformations. *Expert Opin. Drug Metab. Toxicol. 7*, 299–312.

(259) Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T., and Vianello, R. (2005) MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem. 48*, 6970–9.

(260) Klopman, G., Dimayuga, M., and Talafous, J. (1994) META. 1. A Program for the Evaluation of Metabolic Transformation of Chemicals. *J. Chem. Inf. Model. 34*, 1320–1325.

(261) ECB REACH Guidelines. EC.
http://ec.europa.eu/environment/chemicals/reach/reviews_en.htm

(262) Schaafsma, G., Kroese, E. D., Tielemans, E. L. J. P., Van de Sandt, J. J. M., and Van Leeuwen, C. J. (2009) REACH, non-testing approaches and the urgent need for a change in mind set. *Regul. Toxicol. Pharmacol. 53*, 70–80.

(263) Gleeson, M. P., Modi, S., Bender, A., Robinson, R. L. M., Kirchmair, J., Promkatkaew, M., Hannongbua, S., and Glen, R. C. (2012) The challenges involved in modeling toxicity data in silico: a review. *Curr. Pharm. Des. 18*, 1266–91.

(264) EFSA (European Food Safety Authority). (2005) Opinion of the scientific panel of food additives, flavourings, processing aids and materials in contact with food on a request from the commission related to 2-isopropylthioxanthone (ITX) and 2-ethylhexyl-4-dimethylaminobenzoate (EHDAB) in food contact m. *EFSA J 293*, 1–15.

(265) Aprile, S., Del Grosso, E., and Grosa, G. (2011) In vitro metabolism study of 2-isopropyl-9H-thioxanthen-9-one (2-ITX) in rat and human: evidence for the formation of an epoxide metabolite. *Xenobiotica. 41*, 212–25.

(266) Reitsma, M., Bovee, T. F. H., Peijnenburg, A. A. C. M., Hendriksen, P. J. M., Hoogenboom, R. L. A. P., and Rijk, J. C. W. (2013) Endocrine-disrupting effects of thioxanthone photoinitiators. *Toxicol. Sci. 132*, 64–74.

(267) Cozzini, P., Fornabaio, M., Marabotti, A., Abraham, D. J., Kellogg, G. E., and Mozzarelli, A. (2002) Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water. *J. Med. Chem. 45*, 2469–83.

(268) Pripp, A. H., Isaksson, T., Stepaniak, L., Sørhaug, T., and Ardö, Y. (2005) Quantitative structure activity relationship modelling of peptides and proteins as a tool in food science. *Trends Food Sci. Technol.* 16, 484-494.

(269) Khan, T. H. (2012) Novel Tyrosinase Inhibitors From Natural Resources - Their Computational Studies. *Curr. Med. Chem. 19*.

(270) Huang, X., Liu, T., Gu, J., Luo, X., Ji, R., Cao, Y., Xue, H., Wong, J. T.-F., Wong, B. L., Pei, G., Jiang, H., and Chen, K. (2001) 3D-QSAR Model of Flavonoids Binding at Benzodiazepine Site in GABA A Receptors. *J. Med. Chem. 44*, 1883–1891.

(271) Matoba, T., and Hata, T. (1972) Relationship between Bitterness of Peptides and their Chemical Structures. *Agric. Biol. Chem. 36*, 1423–1431.

(272) Collantes, E. R., and Dunn, W. J. (1995) Amino Acid Side Chain Descriptors for Quantitative Structure-Activity Relationship Studies of Peptide Analogs. *J. Med. Chem. 38*, 2705–2713.

(273) Lejon, T., Strøm, M. B., and Svendsen, J. S. (2001) Antibiotic activity of pentadecapeptides modelled from amino acid descriptors. *J. Pept. Sci. 7*, 74–81.

(274) Nakai, S., and Li-Chan, E. (2009) Recent advances in structure and function of food proteins: QSAR approach. *Crit. Rev. Food Sci. Nutr.*

(275) Si W, Gong J, Tsao R, Zhou T, Yu H, Poppe C, Johnson R, Du Z. (2006) Antimicrobial activity of essential oils and structurally related synthetic food additives towards selected pathogenic and beneficial gut bacteria. *J Appl Microbiol*. 100, 296-305.

(276) J. C. Fishbein and J. Heilman. (2014) Advances in Molecular Toxicology. Elsevier.

(277) Böhm, M., Strzebecher, J., and Klebe, G. (1999) Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and

comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa *J. Med. Chem.* 42, 458-477.

Tiziana Ginex

Date of birth: 13/03/1987

Nationality: Italian

Foreign experiences: Visiting PhD student at the University of Barcelona (ES), Campus de l'Alimentació de Torribera, Santa Coloma de Gramenet under the supervision of Prof. F. Javier Luque (05/2014-07/2014, 04/2015-12/2015).

Education

PhD Student in Food Sciences and Technology **January 2013 – December 2015**

Department of Food Science, University of Parma, Italy

Bachelor's Degree in Medicinal Chemistry (110/110) **October 2006 – July 2012**

Faculty of Pharmacy, University of Parma, Italy

Supervised by Prof. A. Mozzarelli and Prof. P. Cozzini

Thesis Title: "ANALISI COMPARATIVA "IN SILICO" DEI RECETTORI NUCLEARI ERa, ERb, AR E PR PER VIRTUAL SCREENING DI ADDITIVI ALIMENTARI "

Publications

Submitted **January 2016**

Ginex Tiziana, Jordi Muñoz-Muriedas, Enric Herrero, Enric Gibert, Pietro Cozzini, and F. Javier Luque. *"Application of the Quantum Mechanical IEF/PCM-MST Hydrophobic Descriptors to Selectivity in Ligand Binding"* **Journal of Molecular Modeling**.

In press **December 2015**

Ginex Tiziana, Jordi Muñoz-Muriedas, Enric Herrero, Enric Gibert, Pietro Cozzini, and F. Javier Luque. *"Development and Validation of Hydrophobic Molecular Fields Derived from the Quantum Mechanical IEF/PCM-MST Solvation Model in 3D-QSAR"* **Journal of Computational Chemistry**.

September 2015

Belma Z. Kurt, Isil Gazioglu, Livia Basile, Fatih Sonmez, Tiziana Ginex, Mustafa Kucukislamoglu, and Salvatore Guccione. *"Potential of aryl–urea–benzofuranylthiazoles hybrids as multitasking agents in Alzheimer's disease."* **European Journal of Medicinal Chemistry** 102: 80-92.

August 2015

Stefano Lorenzetti, Emilio Benfenati, Pietro Cozzini, Luca Dellafiora, Ferdinando Fiorino, Tiziana Ginex, Monica Giulivo, Pierpaolo La Fauci, Livia Basile, Daniele Marcoccia, Elisa Perissutti, Alessandra Roncaglioni and Alberto Mantovani. *"A LIFE-EDESIA project update: An animal-free approach to search for SVHC alternatives."* **Reproductive Toxicology** 56: 26.

April 2014

Ginex Tiziana, Francesca Spyrakis, and Pietro Cozzini. *"FADB: a food additive molecular database for in silico screening in food toxicology."* **Food Additives & Contaminants: Part A** 31(5): 792-798.

February 2014

Ginex Tiziana, Chiara Dall'Asta, and Pietro Cozzini. *"Preliminary hazard evaluation of androgen receptor-mediated endocrine-disrupting effects of thioxanthone metabolites through structure-based molecular docking."* **Chemical Research in Toxicology** 27(2): 279-289.

Oral Presentations, Posters and Books

Book chapter **14-16 October 2015**

"Molecular Dynamics: an emerging tool for exploring the molecular events in Food Research." *Ginex, T., Seira, C., Estarellas, C., Bidon-Chanal, A., Luque, FJ.*

In book: From Medicinal Chemistry to Food Science: A Transfer of In Silico Methods Applications.

Publisher: NOVA Publisher

Editors: Pietro Cozzini.

Poster presentation **14-16 October 2015**

EFSA's 2nd Scientific Conference. Shaping the Future of Food Safety, Together.

Location: Milan, Italy

Poster Title: "*In silico preliminary hazard evaluation of androgen receptor-mediated endocrine-disrupting effects of Thioxanthone photoinitiators and their metabolites*." Ginex T., Dall'Asta C., and Cozzini P.

Poster presentation **14-16 October 2015**

EFSA's 2nd Scientific Conference. Shaping the Future of Food Safety, Together.

Location: Milan, Italy

Poster Title: "*The substitution principle in the LIFE-EDESIA animal free approach: an in silico project update.*" Lorenzetti S, Giulivo M, Benfenati E, Roncaglioni A, La Fauci P, Cozzini P, Dellafiora L, Ginex T, Basile L, Fiorino F, Perissutti E, Marcoccia D, Mantovani A.

Istituto Superiore di Sanità (ISS), Italy.

Oral Presentation **23-25 September 2015**

XX Workshop on the Developments in the Italian PhD Research on Food Science, Technology and Biotechnology.

Location: Perugia, Italy.

Presentation Title: *"QM-based analysis of host-guest interactions in food sciences: Application of 3D analysis of molecular hydrophobicity."*

Poster Presentation **6-9 September 2015**

NMMC, Salerno 2015. XXIII National Meeting on Medicinal Chemistry, 9th Young Medicinal Chemists Symposium.

Location: Campus di Fisciano, University of Salerno, Italy.

Poster Title: *"Novel coumarin derivatives as selective acetylcholinesterase inhibitors."* Basile L., Kurt B.Z., Sonmez F., Gazioglu I., Kucukislamoglu M., Ginex T., Cappello V., and Guccione S.

Book chapter **August 2015**

"How Protein Flexibility Can Influence Docking/Scoring Simulations." Cozzini P., Dellafiora L., Ginex T., and Spyrakis F.

In book: In Silico Drug Discovery and Design Theory, Methods, Challenges, and Applications.

Publisher: CRC Press, pp558

Editors: Claudio N. Cavasotto.

Poster Presentation **26-28 February 2014**

Annual Meeting of Bioinformatics Italian Society (BITS2014)

Location: Faculty of Physics, Università “La Sapienza” di Roma, Italy.

Poster Title: *“A food additive 3D repository for in silico screening in food chemistry: the case of androgen receptor.”* Ginex T., and Cozzini P.