MSc in Bioinformatics
Universitat Autònoma de Barcelona

MASTER THESIS

A Parametrized Approach to LogP-based Hydrophobic Descriptors in Virtual Screening with Pharmscreen®

ELISA GÓMEZ DE LOPE

ACADEMIC TUTOR

Dr. Arnau Cordomí

PROJECT SUPERVISORS

Dr. Enric Herrero

Dra. Tiziana Ginex

UAB
Universitat Autònoma
de Barcelona

Pharmacelera™

JULY 9TH, 2018

A Parametrized Approach to LogP-based hydrophobic descriptors in Virtual Screening with Pharmscreen®

ELISA GÓMEZ DE LOPE

Academic tutor: *Dr. Arnau Cordomí*

Project supervisors: *Dr. Enric Herrero, Dra. Tiziana Ginex*

ABSTRACT

Computational drug design is an evolving interdisciplinary field where new tools and updates of the existing ones are constantly emerging in the aim of improving the effectiveness of predictions to make the early stages of drug discovery more straightforward. Ligand-based approaches take advantage of known biologically-active ligands to design and optimize drug candidates with improved activity. Hydrophobicity/hydrophilicity of molecules is widely considered a critical property regarding ligand-target interactions. In this context, hydrophobicity-based virtual screening is able to scan large databases of compounds and evaluate their similarity to a reference structure. Screened compounds are ranked according to their hydrophobic resemblance to the reference molecule.

In this work, a parametrized approach of atomic MST-model based LogP descriptors for hydrophobicity/hydrophilicity has been developed and tested on a benchmarking database of compounds for virtual screening. Within the framework of Pharmscreen® software, polar and nonpolar $\mathrm{LogP}_{o/w}$-based descriptors have been defined for a classification of atom-types. The suggested pipeline involves identification of atom-types in the query molecules of virtual screening and assignation of hydrophobicity parameters. The effectiveness of this strategy, in terms of (i) execution time for hydrophobicity calculations and (ii) virtual screening performance was also evaluated and compared to Pharmscreen® standard methodology.

The technique described in this work proves substantial reduction of computational cost and execution time in hydrophobicity-based virtual screening while holding performance rates.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

HTS	High-throughput screenings
CADD	Computer-aided drug design
ADME	Absorption, distribution, metabolism and excretion
SAR	Structure-activity relationship
SBDD	Structure based drug design
LBDD	Ligand-based drug design
QSAR	Quantitative structure-activity relationship
VS	Virtual screening
SBVS	Structure-based virtual screening
LBVS	Ligand-based virtual screening
RO5	Rule of five
MEP	Molecular Electrostatic Potential
QM	Quantum mechanics
MST	Miertus-Scrocco-Tomasi
PCM	Polarizable continuum model
MLP	Molecular lipophilicity potential
COSMO-RS	Continuum Solvation Model for Real Solvents
SCRF	Self-Consistent Reaction Field
PAINS	Pan Assay Interference Compounds
DUD	Directory of Useful Decoys
PDB	Protein Data Bank
T	Tanimoto coefficient
ROC	Receiver Operating Characteristic
EF	Enrichment factor
ROCE	ROC Enrichment
SD	Standard deviation
RMSE	Root mean square error
MSE	Mean square error

1. INTRODUCTION

1.1. DRUG DISCOVERY

Pharmaceutical drugs are defined as chemical compounds that are used as medicines (or so do their components) causing therapeutic effects in the purpose of treating, curing or preventing diseases[1,2]. Drug discovery is one of the most far-reaching focus of medicinal chemistry, bioinformatics and pharmaceutical research in the present era. The evolution of computational methods and models together with genomics, biotechnology and biochemistry techniques have utterly changed the concept of drug design. Thanks to the introduction of machine learning algorithms, supercomputers and parallel processing technologies - among other computational advances - in the fields of proteomics and high-throughput screenings (HTS) now we can perform powerful virtual screening over compound libraries in the search for drug candidates. Consequently, the identification of drug targets and the understanding of their biological activity allow the efficient identification and optimization of lead compounds that are subsequently boosted into development candidates.[3]

State-of-the-art: from observations up to computer-aided drug design

The earliest medicinal chemistry started thousands of years ago, when remedies and antidotes were searched in herbs, roots and berries. At that initial stage of the pharmaceutical science, therapeutic agents were used with no knowledge of the underlying biological pathways that were affected by their active pharmaceutical ingredient[1,2]. Drug discovery was rather based on observations of natural phenomena and the consequences of consumption of materials that alleviated distress[4]. Of necessity, the process was empirical. Some therapeutic agents still in use such as opium, licorice, ephedra, marijuana, camellia, alcohol, digitalis, coca, quinine and a host of others predates the rise of modern medicine[1].

The rational design of drugs started to develop around 100-150 years ago[2]. Langley and Ehrlich introduced the concept of *pharmacophore* at the beginning of the 20th century as "*a molecular framework that carries (phoros) the essential features responsible for a drug's (pharmacon) biological activity*", suggesting that not all the molecules would contain the receptors that served as hosts for the drugs[5]. It was also observed that drugs could modify the conduct of enzymes and ligands and indeed have an impact on the chemical reactions in which they are involved[5,6]. In 1894 Emil Fischer[6] first postulated the key-lock theory in which a ligand and a given enzyme were thought to be rigid structures and their action mechanism was explained as an analogy of a key (substrate) and a lock (enzyme). Later on, it was reported that certain properties of enzymes could not be accounted for by the simple relationship proposed by the key-lock hypothesis[6] and the induced-fit theory was postulated. This theory retains the key-lock idea of a fit of the substrate at the active site but states in addition that the binding induces a change in the shape of the enzyme that results in the proper alignment of the catalytic groups on its surface. The "zipper model" is the extreme case of the induced-fit theory. This model takes proteins and compounds as flexible structures that wrap around one another to obtain the proper low-energy conformation of the complex. More recent theories additionally reported that a certain protein can have more than one active site and it can thus bind different molecules[7].

In the first half of the 20th century chemistry and technology experienced colossal advances[5] , and loads of applications were exported to other sciences. This revolution enabled the growth of pharmacology, molecular biology and the clinical sciences that shape drug analysis[5]. During the last decades drug research at laboratories has been increasingly introducing biotechnology techniques and genomics discoveries for the identification of new targets[1,2,8]. In an estimate from 2011, 435 human genome products were identified as therapeutic drug targets of FDA-approved drugs[9]. Additionally, one of the main applications of robotics and detectors in the bio-chemistry field was the development of HTS[10]. HTS allowed drug discovery researchers to test the activity of a potential drug on the target (usually they measure ability to cause a certain effect) and entailed the creation of large chemical libraries that could be screened and eventually tested in vitro.

Besides this breakthrough at *wetlab*, HTS techniques are expensive and arduous to handle due to the big scale of the data. The development of computational algorithms and models applied to life sciences research allowed bioinformatics and cheminformatics to arise as the new way of understanding drug discovery among pharma community: computer-aided drug design (CADD)[1–3,9,11]. In CADD, molecular mechanics and dynamics are used to estimate the existence and strength of molecular interactions between any compound and the target, to predict the conformation of both molecules and to model the binding and subsequent conformational changes the target experiments. Properties of the molecules are estimated by empirical, semi-empirical methods[12], and quantum chemical calculations[13,14], which provide optimized parameters for certain properties of the molecules (hydrophobicity, electrostatic potential, volume, etc) that can cause an impact on the binding affinity of drug candidates[15].

In silico scan of large databases of parametrized compounds using models that weigh their properties in the search for the optimum activity have made the drug discovery process more efficient and low priced at its early steps[16]. In silico screenings also provide information about the action mechanism, allowing optimization of the leads. Nowadays, new computational approaches with increasingly improved performance are constantly arising. The successful application of such techniques demonstrates their key role in the early stages of drug discovery[3].

The complex process of drug design

In the process of drug design, the physicochemical properties of molecules will be deeply tested as they play a crucial role not only in absorption, distribution, metabolism and excretion (ADME) but also in the affinity to a complementary receptor and the following interaction. HTS for ADME properties is nowadays the *status quo* of the drug discovery industry[1,2].

Drug design at its basis aims to predict the ability of a given molecule to bind a given target and the strength of this binding. The discovery and further development of a drug is a complex process that consumes enormous time and cost[3,17–19]. To put it into perspective, it takes an average of 10–15 years with an approximate cost of US$800 million to US$1.8 billion[20–22]. The long, tedious and expensive steps required for a therapeutic agent to reach patients[23] is illustrated in *Figure 1*.

A drug-target is a broad term comprising a wide range of biological entities such as proteins, genes and RNA. Targets should be accessible to the putative drug molecule and bring the pursued biological response upon binding. An ideal “druggable” target would be efficacious, safe and meet clinical and commercial needs[24]. Identified targets must be fully prosecuted. There is a wide range of validation techniques and protocols detailed in literature, including in vitro and in vivo methods[23]. Multi-validation approaches increase the consistency and confidence in the observed results.

Lead identification stage consists on the understanding of the proper mode of interaction target-receptor. While biochemistry is predominantly devoted to the grasp of the mechanism of action of a given chemical modulator, CADD uses the structural knowledge of either the target (structure-based) or known ligands with bioactivity (ligand-based) to determine leads out of libraries of promising candidate drugs by means of virtual screening. These leads ought to be amenable to future modifications and adjustments that will be applied in order to have convenient physicochemical properties that suit the ADME/Tox acronym[25–28].

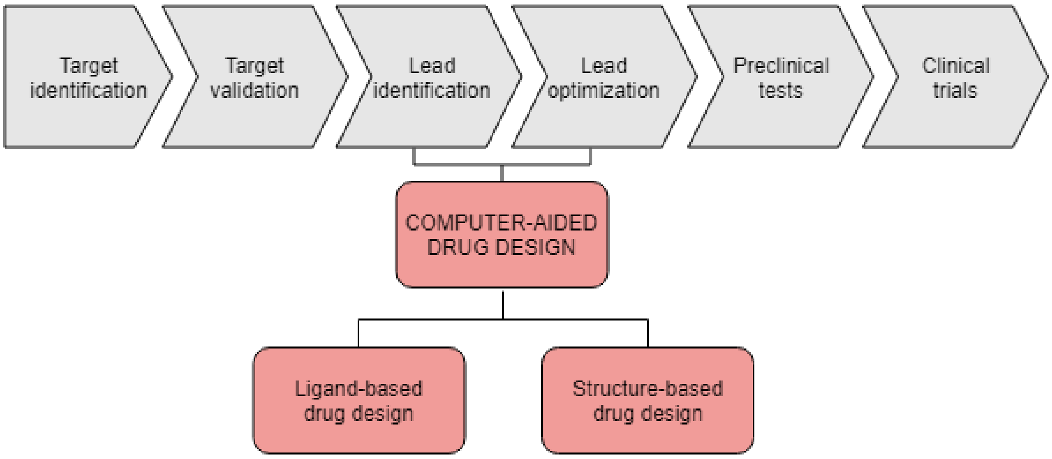


Figure 1. Schematic diagram comprising the main steps in drug design process. The role of computer-aided drug design is evidenced.

Further assays and pre-clinical tests measure structure-activity relationship (SAR), affinity and the pharmacokinetics properties to check leads' efficacy firstly in cell- and tissue-based bioassays and the successful ones in any available *in vivo* models. If there is available structural information of the target, molecular modelling and other structure-based tools, X-ray crystallography and NMR can be applied to enhance the SAR or reveal new binding sites on the targets and other interesting features. It is important to apply activity tests for sites where selectivity might be known or expected to be an issue. Molecules meeting these requirements would be escalated to in vivo assays to test activity, toxicity and ADME properties on other species and on proteins encoded by human DNA[25].

The last phase of the drug development consists on assays under human administration. Even though at this late stage the financial consequences of failure are much higher, only one in 10 candidates is reaching the market. Besides, when a candidate reaches the clinical stage the project has become public knowledge and because termination can influence confidence in the company and shareholder value it turns increasingly difficult to cancel.

Computer-aided drug design

Despite the hardship of drug development, recent advances in computational chemistry and chemoinformatics have already proved to accelerate and overcome limiting factors of drug discovery[29]. Several software tools containing powerful algorithms and models have been developed to accurately model protein structure and function and to identify active sites and interacting ligands, aiming to study protein-ligand interaction[30]. CADD offers an *in-silico* alternative to traditional medicinal chemistry techniques for studying the structure and predicting the biological activity of drug candidates with the appealing advantages of being both cost-effective and relatively rapid.

Structure based drug design (SBDD) and ligand-based drug design (LBDD) have emerged as the two general approaches of computer-aided drug design and are both active areas of research in academic and commercial realms[19,29,30].

SBDD methods are based on the study of the 3D-structure of validated biological targets. There are several databases storing structural information; one of the most used is Protein Data Bank. In case the structure of a certain target is not solved, homology modelling techniques allow to predict the structure of a protein from closely related homologous proteins[31] (MODELLER[32], SWISS-MODEL[33]); other possibilities include folding recognition or *ab initio* protein modelling[20].

The goal is to identify bioactive binding sites that could be used to build putative ligands[19]. Structure-based virtual screening (database scan for ligands) and *de novo* design of ligands (build from scratch) knowing the size and main characteristics of the binding pocket are the main methods to build drug compounds. The ligand (drug candidate) will ideally be an easily synthesized, high-affinity small molecule with desirable pharmacological properties.

LBDD focuses on known ligands' structures for a target so that to establish a SAR. This knowledge is used to design or optimize drug candidates with improved activity[19]. The main advantage of this approach is that not preliminary information about the target structure is required. LBDD is based on the assumption that molecules with similar physico-chemical properties and cause similar biological effects when binding to the target[34]. Hence, the very first step is to compare the known ligand 3D structure with the 3D structure of compounds in available libraries. Current approaches on LBDD are primarily virtual screening (which will be further discussed in next sections of this work) and quantitative structure-activity relationship (QSAR). QSAR aims (although not always possible) to establish a correlation between the molecular structure and a certain bioactivity. 3D-QSAR methods such CoMFA[35] and CoMSiA[36] suggest active compounds by applying a pharmacophore model determined from the relationship between the activities of a set of aligned 3D molecules and properties extracted from their physicochemical descriptors.

1.2. VIRTUAL SCREENING

The pharmacophore concept theorized by Ehrlich, which was already noted in previous sections, represented an important milestone in pharmaceutical science[5]. Among all in silico techniques that have been developed around the idea of the pharmacophore (a molecular framework describing the essential features responsible for a drug's bioactivity), virtual screening (VS) of molecular databases may be considered the most successful and widely-used one.

VS allows to scan large databases of compounds and identify new promising molecules according to their similarity with a reference structure. Its use has been extensive either in SBDD and LBDD[37–39], and the technique lies on the concept of ranking query structures according to their similarity to the reference so that only the top ranked ones will be accounted for further stages of the drug discovery process. In 3D methods the ranking will depend on the molecular alignment, which in turn depends on the type and quality of the molecular description.

Structure-based virtual screening (SBVS) is used to screen large *in silico* compound databases to identify potential ligands to a query target. Well-known docking software examples include DOCK[40], AutoDock[41], AutoDock Vina[42] and SwissDock[43]. They offer different docking and scoring functions whose variances are based on different approximations to describe the chemical environment defined by conformational flexibility and solvation effects.

Compounds used in both structure- and ligand-based VS usually come from publicly accessible in silico databases, like ZINC[44] or PubChem[45]. Also, some pharmaceutical companies have their own libraries of molecules. Once the screening is done, the ranking of the molecules is obtained based on the calculated (or predetermined[46]) free energy of binding $\Delta G_{binding}$.

Ligand-based virtual screening (LBVS) uses one or more molecules that are known to be active for a specific target (references) to extrapolate similar compounds according to their structural and physicochemical properties. Current approaches on LBVS are based on the use of molecular shape or volumes, pharmacophore or molecular fields (interactions)[34].

Shape or volume methods like ROCS[47] from *OpenEye Scientific Software Inc*. do overlap of the structure of the known ligand and query molecules and score the alignment by using a given function that will consider the volume or shape of the reference molecule and the chemical compounds in the library[39]. The pharmacophore method establishes scoring values depending on functional groups like aromatic systems or hydrogen bond acceptors/donors. This method compares patterns of distances between the functional groups in the reference and query molecules, and generates a similarity value[48]. Pharmer[49] from the University of Pittsburgh is a well-known pharmacophore method for VS.

Molecular interaction field methods focus on the characteristics of the molecular fields that exist around the molecules. They were introduced in QSAR modelling, but can also be applied to VS[50]. There is a lot of research being done in regards of molecular fields since methods like CoMFA[35] have revealed better performance than other techniques. Though, a lot of calculations are needed, and it still presents efficiency limitations.

Regarding to interaction field methods, there are three main factors generally considered in *in silico* molecule models: size shape, hydrogen bonding and lipophilicity. They were formalized by Lipinski, who formulated in 1997 his famous Rule of five (RO5) (also known as Lipinski's rule of Five) based on the observation that most medication drugs were relatively small and lipophilic molecules[11,51,52]. Among these three factors, molecular size and shape descriptors are not explicitly used to model ADME(T) profile as they may be implicitly accounted for in the lipophilicity term. Also, hydrogen bonding character may be partially correlated to lipophilicity[49]. Thus, the importance of the role of hydrophobicity/hydrophilicity becomes evident in terms of the description of molecules' behaviours and properties.

1.3. HYDROPHOBICITY IN INTERACTION-FIELD VIRTUAL SCREENING

The relevance of the hydrophobic/hydrophilic properties of molecules is widely known to be crucial for their absorption and pharmacological activity[27,53–56]. Indeed, the solvation and hydrophobic/hydrophilic areas of both ligand and target and has been recognized as one of the major variables that contributes in their interaction[57,58]. Ligand and receptor binding sites are one and the other expected to show complementarity in their distribution of hydrophobic/hydrophilic areas. Even though the use of hydrophobic similarity in LBVS has not been intensely explored, finding compounds with similar hydrophobicity patterns than an active ligand for a given target is an interesting strategy to select promising leads.

Similarity in the chemical space

LBVS aims to disclose compounds from the chemical space being similar to the reference molecule[34,50]. Though, similarity is a subjective and multifaceted concept regardless which objects are considered. Molecular similarity and molecular diversity have various definitions in function of chosen criteria[59,60]. In chemoinformatics, the measure of either molecular similarity and diversity involves the use of three main components: descriptors, coefficients and a weighting scheme.

Descriptors are used to characterize the molecules to be compared. They can be determined from the structure (constitution, configuration and conformation), or the properties (physical, chemical, biological)[61]. Ideally, descriptors used for modelling should represent the chemical reality of a system while being rapidly calculated and easy to interpret by computers and users. They can be obtained from specific definitions, from combinations of other descriptors or be numerical values (usually physicochemical properties). The number and complexity of molecular descriptors has been multiplied in the last years[62]. The basic idea behind the descriptors is that molecules with similar descriptors are supposed to have similar properties.

Similarity coefficients are functions that provide a quantitative measure of the chemical resemblance degree[60]. This is, they transform pairs of compatible molecular representations into numbers. There are global and local similarities; local similarities are described for particular atoms or molecular fragments (atomic charges, bonds polarizabilities, …) whereas global similarities refer to the resemblance between the two whole molecules comparing molecular volume, molecular surface, dipole moment, topological indices, … Similarity coefficients can also be classified into correlation, probabilistic, associative, and distance coefficients. The performance of these similarity coefficients has been widely studied[63,64]. The Tanimoto coefficient[65] is the most utilized one and it is implemented in Pharmacelera's Pharmscreen® [66] software.

The third main component of the similarity computation is the weighting scheme, which is used to assign a certain degree of importance to each of the various components of the representations[60].

When searching similarity between chemical compounds it is important to contemplate the similarity paradox. Similarity searches are usually based on the similarity property principle, which stands that structurally similar molecules are more likely to have resembling properties[67]. However, multiple works have questioned this principle when obtaining contradictory results[61,68]. Molecules can seem structurally similar but when considering the steric, hydrophobic or electrostatic features, the comparisons showed significant differences, particularly with the retained three MEP (Molecular Electrostatic Potential) points for each molecule[60]. Thus, to describe the similarity between molecules it is better to use non-linear variable mapping, where the activity is represented by a non-linear function of structural, topological and molecular descriptors.

Descriptors in field-based VS

Descriptors in interaction field VS are used to represent molecules as they enclose their features and properties[69]. This numerical representation allows molecules to be processed not only in virtual screening, but also in similarity/diversity analysis, library designs and other chemical studies. Descriptors can be classified into categories depending on the criteria. According to the dimensionality of the structure representation, descriptors can be 1D (constitutional descriptors like atom and bond counts, molecular weight,…), 2D (based on molecular topology: topological indices, fragment counts,…) and 3D descriptors (geometrical parameters like molecular surfaces and fields or quantum chemistry parameters)[69,70].

Choosing a particular molecular descriptor for an analysis is a difficult task since literature has shown that different descriptors perform better in certain applications[16,69]. Also, it is important to consider the complexity of the descriptor as the choice is generally limited also by the required computational power. Moreover, some low-complexity descriptors have proved to be sufficient to encode the features of a given compound in certain situations[70]. Consequently, the election is usually based on the experiment itself, the chemical intuition and the criteria of the researcher.

In VS, the descriptors of the query compounds are compared with the descriptors of the reference molecule. An alignment of both descriptors is performed, and a similarity coefficient is obtained. The comparison of the similarity coefficients of all query molecules results in a ranking of the structures in the dataset where the compounds appear sorted according to their similarity to the reference structure[34,50,71].

At the early days of LBDD it was common to apply 2D descriptors such as the number of nitrogen atoms in the molecule or the number of double non-aromatic bonds. These descriptors provided fast screenings but they excessively favoured molecules sharing a common chemotype with the reference molecule, which not always meant higher degree of general similarity[72]. The introduction of 3D descriptors obtained with semiempirical and quantum mechanics methods allowed the VS to sophistically explore the structural diversity of the chemical agents in the library[16,34,50,62,69,70,72]. The idea was that in the proximity of the ligand, when a drug-like molecule approaches its target, the forces are the main agent involved in the ligand-target interaction rather than the topological characteristics.

Field-based approaches compare these reactive properties of the surface of molecules like electrostatic potential or hydrophobic/hydrophilic areas at suitable regions in the surroundings of the molecule. They find correlations between reactive properties of the molecules and generate a model that will predict biological properties of screened compounds. Accurate molecular alignment plays a key role in the scoring of the VS[72], thus the final result will improve when contemplating abstract chemical features represented by molecular fields, since shape-based alignments can fail at superposition if, for example, the dimensions of the aligned compounds show significant differences[39].

Hydrophobicity-based molecular similarity approaches

The hydrophobicity (or lipophilicity) of a drug has been recognized by many authors as one of the main factors influencing the extent of protein binding, metabolism, and absorption[11,51,52,73,74]. Hydrophobicity is therefore an important parameter monitored by medicinal chemists in drug discovery on daily basis as the evidences point out it plays a main role in governing kinetic and dynamic aspects of the binding of ligands to the target receptors[11,13,51,52,73,74]. However, the inclusion of the differential solvation properties of molecules in similarity measurements is scarce. In this work, hydrophobicity descriptors have been used in the VS to describe and compare the distribution of hydrophobic/hydrophilic patterns of compounds.

The hydrophobic/hydrophilic character of molecules can be described by using parameters related to the transfer of the molecule from apolar and polar phases (usually water and octanol)[53]. However, because the receptor sites usually have hydrophobic pockets that will bind the apolar parts of the drug whereas its polar groups interact with hydrophilic sites, general descriptors of the hydrophobic/hydrophilic nature of the whole drug-like compounds are not that useful. Accordingly, not only the global hydrophobic/hydrophilic character of the drug, but rather the distribution pattern of hydrophobicity and hydrophilicity regions is important. Approaches for local 3D representation of the hydrophobicity/hydrophilicity patterns include methods such as qualitative distribution of hydrophobic/hydrophilic regions in the surface (intuitive yet arbitrary)[75], the use of atom-based parameters (atomic charges, interaction energies with probes or the free energy surface density)[57,58], or via lipophilic potentials that combine fragmental hydrophobic contributions with a distance-dependent function[76,77].

The use of empirical methods for the calculation of absolute and relative solvation free energies presents withdrawals as it assumes that the contribution of a given group to solvation is largely independent of its molecular environment. Despite correction terms added *a posteriori* to account for neighbouring influence[78], their ability to properly capture all contributions to solvation is still questioned. Plus, they do not consider the conformation of molecules and have a limited number of defined groups or atom types, so they cannot properly represent all structures. Semi-empirical methods have appeared as an alternative that describes the hydrophobic/hydrophilic patterns by means of a combination of empirical parameters and descriptors of molecular properties defined with quantum mechanics (QM) calculations[53]. Theoretical methods provide another way to obtain absolute and relative free energy of solvation, which is defined as the difference between the reversible works necessary to generate a molecule in the gas phase and in solution. Such methods can be determined at QM levels and using a discrete macroscopic representation of the solvent[13,79,80].

Continuum methods are based on a classical discrete representation of the solute[12], which makes them simple, computationally efficient and therefore convenient for drug-design and molecular modelling packages[81]. Moreover, QM-continuum strategies have reached estimations of free energy of solvation for nonpolar solvents with only a few tenths of kcal/mol error[14].

1.4. QUANTUM MECHANICS MODELS FOR HYDROPHOBICITY DESCRIPTORS

Recently, a new similarity index has been suggested to measure the hydrophobic similarity between molecules[82] based on the partitioning of the free energy of transfer between water and octanol into atomic contributions described by *F.J.. Luque et al. (1999)*[53] within the framework of the Miertus-Scrocco-Tomasi (MST) continuum solvation model[83]. These atomic contributions allow to define global and local measures of hydrophobic similarity based on semiempirical calculations.

The posterior work of *Jordi Muñoz-Muriedas, F. Javier Luque et al (2005)*[84] assessed the suitability of the hydrophobic atomic contributions strategy for two series of compounds (ACAT inhibitors and 5-HT3 receptor agonists). The compounds were aligned to maximize the global hydrophobic similarity using a Monte Carlo - simulated protocol. Inspection of the 3D distribution of hydrophobic/hydrophilic contributions in the aligned molecules concluded that the procedure allows to identify pharmacophoric recognition patterns on regions of very similar hydrophobicity, whereas low similarity regions were associated with structural elements that regulate the differences in activity between molecules[84].

The technique developed in this project benefits from the use of the aforementioned hydrophobicity similarity index proposed by *F.J. Luque, X. Barril & M. Orozco*[53] based on the polarizable continuum model (PCM) developed by Miertus-Scrocco and Tomasi (MST)[85,86]. The method results in 3D pictures of the intrinsic solvation properties of molecules that are treated as a descriptor of the atomic hydrophobicity of molecules screened in VS.

Other approaches to lipophilicity

Piles of computational empirical methods have been developed to estimate LogP_{oct} from contributions determined by molecular fragments or atom types. These empirical approaches to lipophilicity potential include the molecular lipophilicity potential -MLP-[87], or the Hydropathic INTeractions -HINT-[88]. They are based on the concept that the spatial distribution of the empirically determined lipophilicity of molecules - including the differential solvation effects in water and octanol-, provides guidelines about the molecular determinants of ligand binding[13].

The molecular lipophilicity potential (MLP)[87] offers a quantitative 3D description of the lipophilicity potential from all the molecular fragments on the surrounding space of a compound. MLP can also be used to model forces governing interactions between bioactive molecules and receptors.

The HINT method[88] can be considered a direct application of the MLP approach. Formulated by Abraham and Leo[88], it assumed the possibility to scale hydrophobic fragments constants at the atomic level by means of an adaptation of the CLOGP[89] method. An interaction energy term was then defined and applied to score each atom-atom interaction.

QM strategies and semi-empirical methods have been exploited in the purpose of computing lipophilic descriptors. Klamt and collaborators developed the COSMO-RS (Continuum Solvation Model for Real Solvents)[90] model, which accounts for many facets such as electrostatic, H-binding affinity, and hydrophobicity, hence covering enthalpy and entropy of solvation. The method evaluates the total energy and the polarization charge density generated on the surface of the molecule ideally immersed in the solvent, and the charge density is decomposed into fragmental contributions[90].

Hydrophobic contributions under the MST model

Another QM-based approach is Self-Consistent Reaction Field (SCRF) models, which describe the solvent as a continuum polarizable medium (PCM) that reacts with the atom charges projected on the molecular surface[80,91]. The MST atom-based LogP partition method pertains to a revisited version (IEF-PCM) of the polarizable continuum solvation model.

The hydrophobicity of a molecule (M) can be quantified by its partition coefficient - abbreviated *P* - between water and an organic phase (in biology and chemistry the organic solvent is typically octanol)[54,55,92–94]. This equilibrium thermodynamic property measures the ratio of concentrations of the unionized compound between organic and aqueous solvents (Eq. 1)[13,55,95]. *P* can also be regarded as a measure of the difference in solubility of the compound in these two phases and is usually taken as an indicator of intrinsic lipophilicity in the absence of ionization or dissociation of the compound.

$$P = \frac{[M]organic}{[M]water} \; (1)$$

The definition of *P* can be related to the transfer free energy (ΔG_{tr}) of the solute between the immiscible phases (Eq. 2). Therefore, the hydrophobicity can be expressed in terms of the solvation free energy (ΔG_{sol}) of the compound upon transfer from the gas phase to the water and organic phase (*Figure 2*).

$$logP = -\frac{\Delta G_{tr}^{o/w}}{2.303\ RT} = \frac{\Delta G_{tr}^{water} - \Delta G_{tr}^{octanol}}{2.303\ RT} \quad (2)$$

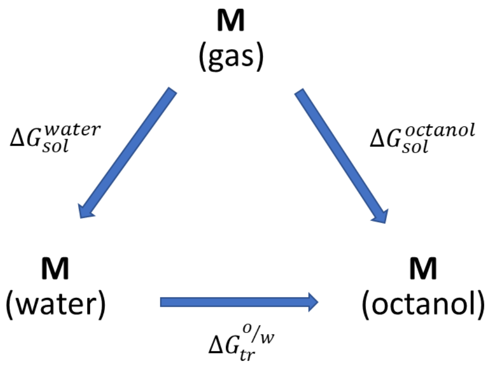


Figure 2. Thermodynamic cycle for the determination of free energy of transfer of a molecule M between two immiscible solvents from the solvation free energies. From Theoretical and quantum chemistry at the dawn of the 21st century, Carbó, R.; Chakraborty, T. (2016).

In the MST model the hydrophobicity of a molecule is partitioned into atomic contributions, which can be further decomposed into three contributions: cavitation, van der Waals, and electrostatic (Eq. 3).

$$\Delta G_{sol} = \Delta G_{ele} + \Delta G_{cav} + \Delta G_{vwa} \; (3)$$

The cavitation term (ΔG_{cav}) represents the work required for creating a cavity shaped to accommodate the solute in the solvent, and the van der Waals term (ΔG_{vW}) accounts for dispersion-repulsion interactions between solute and solvent. Both are non-electrostatic contributions and are computed by means of expressions that depend linearly on the solvent-exposed surface of each atom in the molecule. Hence, they can be directly decomposed into atomic contributions. (Eq. 4 and Eq. 5).

$$\Delta G_{cav} = \sum_{i=1}^{N} \Delta G_{C-P,i} = \sum_{i=1}^{N} \frac{S_i}{S_T} \Delta G_{P,i} \; (4)$$

where $\Delta G_{P,i}$ is the cavitation free energy of atom i determined using Pierotti's formalism[96], whose contribution is weighted by the contribution of the solvent-exposed surface (S_i) of atom i to the total surface (S_T).

$$\Delta G_{vw} = \sum_{i=1}^{N} \Delta G_{vw,i} = \sum_{i=1}^{N} S_i \xi_i \; (5)$$

where ξi denotes the atomic surface tension of atom i, which is determined by fitting the experimental free energy of solvation[83].

The third component is the electrostatic term (ΔG_{ele}), which measures the work needed to build up the solute charge distribution in the solvent. As explained in *J.M. Muriedasa, F. J. Luque et al*

(2005)[84], following the 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 ΔGele into atomic contributions (Eq. 6) following a perturbative description of the solute-solvent electrostatic interaction[97].

$$\Delta G_{ele} = \sum_{i=1}^{N} \Delta G_{ele,i} = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\in i}}^{M} \left\langle \Psi^0 \middle| \frac{1}{2} \frac{q_j^{sol}}{|r_j - r_i|} \middle| \Psi^0 \right\rangle \quad (6)$$

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^0 is the wave function of the solute in the gas phase. $\Delta G_{ele,i}$ represents the electrostatic contribution of atom i, and is calculated from the interaction of the molecular electrostatic potential generated by the whole molecule with the subset of reaction field charges affecting the solvent-exposed surface of atom i.

Consequently, the sum of the three contributions for all atoms in the molecule (Eq. 7) results in the total solvation free energy.

$$\Delta G_{sol} = \sum_{i=1}^{N} \Delta G_{sol,i} = \sum_{i=1}^{N} \left(\Delta G_{ele,i} + \Delta G_{cav,i} + \Delta G_{vW,i} \right) \quad (7)$$

Application of Eq. (7) to the solvation of a given compound in water and octanol leads to octanol/water transfer free energy ($\Delta G_{tr}^{o/w}$). The octanol/water transfer free energy can be expressed as the sum of atomic contributions (Eq. 8).

$$\Delta G_{tr}^{o/w} = \sum_{i=1}^{N} \Delta G_{tr,i}^{o/w} = \sum_{i=1}^{N} \left(\Delta G_{ele,i}^{o/w} + \Delta G_{cav,i}^{o/w} + \Delta G_{vW,i}^{o/w} \right) \quad (8)$$

Finally, the hydrophobicity of a molecule, expressed as logarithm of the octanol/water partition coefficient (*LogP*), can be related to the sum of the atomic contributions (Eq. 9 and 10).

$$logP = \sum_{i=1}^{N} logP_i = \sum_{i=1}^{N} \left(logP_{ele,i} + logP_{cav,i} + logP_{vW,i} \right) \quad (9)$$

$$logP_X = \sum_{i=1}^{N} logP_{X,i} = \sum_{i=1}^{N} - \frac{\Delta G_{X,i}^{o/w}}{2.303\ RT} \quad (X = ele, cav, vW) \quad (10)$$

The strategy of partitioning the *LogP* into atomic contributions allows to quantify the hydrophobic resemblance of two molecules by comparing their hydrophobicity patterns. This is achieved by means of the definition of measures of hydrophobic similarity such as the dot product of the hydrophobicity dipoles[53], or from similarity functions [82], which in turn can be used to derive SARs[84]. Indeed, the MST-based hydrophobic contributions have been used as physicochemical descriptors for 3D-QSAR studies targeting both ligand affinity and receptor selectivity[98]. 3D-QSAR models combining the electrostatic and non-electrostatic components of the octanol/water LogP have showed similar levels accuracy than standard methods CoMFA[35] and CoMSIA[36].

Pharmscreen® is a tool designed by *Pharmacelera S.L.* that performs optimized virtual screening including hydrophobicity descriptors. The methodology for the hydrophobicity-based VS integrates RDKit tool[99] and a locally modified version of MOPAC-6[100] QM package for computing the MST-based atomic contributions to $LogP_{o/w}$[83]. These contributions are utilized to align molecules and to evaluate their similarity with a reference structure. Within this context, this work suggests an alternative pipeline of Pharmscreen® -named "AT"- for performing such hydrophobicity-based virtual screening with parametrized MST-atomic-contributions.

Fragmental and atom-based LogP descriptors.

Previous sections describe how the octanol-water partition coefficient – LogP - is a physical property extensively used to describe a chemical's lipophilic or hydrophobic properties. Accordingly, literature contains many methods for estimating log P[92], most of them classified as fragment constant methods. In such methods a structure is divided into fragments (atom or larger functional groups) and values of each group are summed together - sometimes with structural correction factors - to yield the LogP estimate[101].

The fragment method, originally proposed by Hansch and Leo[56], later refined by others (Klopman[102], Kudo[103],…) and computerized by Chou and Jurs[104] in the CLOGP program, has become a standard calculation of LogP. The atomic contribution method was developed by Broto[105] and improved by Ghose and Crippen[106–108]. This method assigns additive contributions of molecular LogP to the individual atoms in the molecule by classifying atoms into chemically distinct types and fitting the contributions on a dataset of experimentally determined LogP values. The atomic-specific distribution of the LogP values allows to generate a detailed lipophilicity map.

S.A. Wildmann and G.M. Crippen (1999)[93] presented a definition of atomic LogP atom-types and contributions that addressed some negative comments about atomic LogP calculation methods found in literature related to ambiguity in the classification system, unrealistic values of some atom contributions, bias toward underestimation of LogP or large number of atom types. Under the proposed methodology, the partition coefficient (and other descriptors like molar refractivity - MR) of small molecules can be calculated as the sum of the contributions of each of the atoms in the molecules, even though this property is not strictly additive. Intramolecular interactions can be accounted for by classifying atoms into different types based on attached and neighbouring atoms, as described in Eq. 11:

$$P_{calc} = \sum_{i=1}^{N} n_i a_i \quad (11)$$

Where P_{calc} is the property to be calculated (in this case LogP); n_i is the number of atoms of type i present in the molecule, and a_i is the contributions for atoms of type i.

The classification system proposed[93] distinguishes different atom types for atoms with different nearest neighbours (i.e. first carbon of CH_3C is treated different from that of CH_3N). Only atoms with similar chemical nature and the same approximate atomic contribution to log P were grouped together into a single common type. The resulting classification[93] provides 72 atom types, such that each atom present in a given molecule will match one and only one of these atom-types. This atom-type classification has been employed[109,110] in multiple studies and proved to be reliable.

This work suggests "AT", a new pipeline for Pharmscreen® hydrophobicity-based virtual screening in which hydrophobicity/hydrophilicity MST-based atomic values are parametrized according to the mentioned atom-types classification. The proposed approach assigns parametrized atomic hydrophobicity values to specific (previously identified) atom-types in query molecules instead of performing the time- and resources-consuming computation of such parameters for each atom on each molecule.

2. OBJECTIVES

The main goal of the present work lies on the development of “AT”, a new low time-cost but still accurate pipeline to perform effective VS based on the use of parametrized QM-based hydrophobic descriptors. Within the framework of *Pharmacelera S.L*. in-house virtual screening software Pharmscreen®, this project takes advantage of the already demonstrated capability of LogP-based hydrophobic descriptors to model ligand similarity and potential activity against specific targets.

An extensive parametrization of the atomic contributions to $\mathrm{LogP_{o/w}}$ aims to be performed in order to extract the electrostatic ($\mathrm{LogP_{ele}}$) and non-electrostatic ($\mathrm{LogP_{cav}}$) contributions to $\mathrm{LogP_{o/w}}$ for each of the atom types included in *S.A. Wildmann and G.M. Crippe*. The final goal is to implement these parametrized descriptors so that $\mathrm{LogP_{o/w}}$ atomic contributions do not have to be iteratively computed for each atom of each of the query compounds in virtual screening but rather assigned with the parametrization. To accomplish this purpose, the following short-term objectives have been addressed:

- ❖ Initial study and comprehension of chemotypes in the atom-type classification system described by S.A. Wildmann and G.M. Crippe and MST-model application for hydrophobicity-based virtual screening.

- ❖ Preparation of training set.

- ❖ Computation of hydrophobic atomic contributions for the training set under Pharmscreen® standard version and identification of atom-types.

- ❖ Parametrization of the hydrophobic contributions for each atom-type.

- ❖ Performance of virtual screening on validation set using standard Pharmscreen® hydrophobic parameters and virtual screening on validation set using “AT” hydrophobic parameters.

- ❖ Analysis of the results: time, performance in VS and accuracy of the parametrization.

3. MATERIALS AND METHODS

“AT” pipeline implements a parametrization of the hydrophobicity contributions that allows to skip the singular calculation of such contributions for each atom of each molecule as the conventional schema of Pharmscreen® does. The suggested strategy performs direct assignation of the parameters to the atom-types instead, saving time and computational cost.

3.1. PREPARATION OF TRAINING SET

The training set for parametrization must be large enough to generate unbiased representations of the atom-type hydrophobicity values. Additionally, training set must be diverse, comprising as many classified atom-types as possible so that most –ideally all- of them are parametrized.

The training set employed in the parametrization is composed of the Specs library[111] and a *Pharmacelera S.L.* internal dataset of compounds. The Specs library included 371,846 molecules and 17,985,122 atoms. This database is publicly available (www.specs.net) and contains new structures exhibiting structural characteristics of biologically active compounds (meeting ADME requirements). Specs compounds are subjected to LC/MS and 1H-NMR analysis and meet strict analytical criteria. The internal library is a set of *duggable* chemical structures containing 209,426,883 atoms belonging to 4,559,318 molecules that were neutralized and clustered to maximize chemical diversity. The library had been previously filtered for Pan Assay Interference Compounds (PAINS) in order to remove structures likely to be false positives in HTS, toxic or cause unwanted side effects. Besides, only one of the possible conformers (the lowest energy structure) was considered for each of the molecules in the set.

An overall amount of 4,931,164 neutralized molecules and 227,412,005 atoms were analysed. The total training set contains 66 out of the 72 atom-types described in *S.A. Wildmann and G.M. Crippen (1999).* The six-remaining atom-types are *C27, N10, O12, Hal, Me1* and *Me2*, which are not parametrized by Pharmscreen® tool either because of their unlikelihood to happen in drug-like molecules.

3.2. COMPUTATION OF HYDROPHOBIC ATOMIC CONTRIBUTIONS FOR TRAINING SET AND IDENTIFICATION OF ATOM-TYPES.

Pharmscreen® can perform the computation of hydrophobic atomic contributions for libraries of molecules. Before operating for the MST solvation calculations, the molecular geometry is optimized for all query compounds by means of the standard geometry optimization procedure implemented in an in-house version of MOPAC-6 methodology. Also, Pharmscreen® is able to use the RDKit toolset to identify the atom-types present in the compounds of the dataset according to the classification of *S.A. Wildmann and G.M. Crippen*.

MST solvation energies in water and in octanol are computed to generate the fractional contributions to the octanol/water transfer free energy. To this purpose, Pharmscreen® employs again the in-house version of MOPAC-6 implementing the semiempirical RAM1[112] method of the MST model[83,85,86]. This methodology was chosen because of its low computational cost compared to *ab initio* methods, which motivates its wide application in the study of large-sized drug-like compounds.

3.3. PARAMETRIZATION OF THE HYDROPHOBICITY PER ATOM-TYPE

The parametrization of LogP atomic contributions is a crucial part of this work. The "AT" implementation of Pharmscreen® integrates the parametrized contributions so that after accessing RDKit tool to identify the atom-types present in the query molecule, the parametrized values are assigned to the atom-types as their hydrophobicity/hydrophilicity atomic contributions instead of being iteratively computed. Therefore, an accurate parametrization plays a critical role in the suitability of the suggested pipeline.

The atomic hydrophobic contributions (electrostatic, cavitational and Van der Waals), and the total atomic hydrophobicity coefficient -abbreviated *$LogP_{ele}$*, *$LogP_{cav}$*, *$LogP_{vwa}$* and *LogP*- of the atom-types identified in training set were statistically analysed. A refined bash pipeline of python scripts with object-oriented techniques (defining classes of objects) was implemented in order to gather data per contribution and atom-type, compute statistical measures of central tendency and diversity, and plot histograms per atom-type of each of the contributions and the total LogP -this is, 66 · 4 = 264 histograms- (see some examples at appendix). Python built-in modules that were addressed include *matplotlib*, *numpy*, *os*, *sys* and *math*. Also, a *mol2_sdf* module programmed *ad hoc* with particular functions and classes was used to ease the handle of mol2-type files. The bioinformatics workflow designed for this parametrization is visually presented in *Figure 23* at section *7. APPENDIX* subsection *F*. Intellectual property free scripts related to parametrization are available at https://github.com/elisagdelope/TFM_scripts-IP-free.git.

The measures of central location that were computed correspond to the mean (average) and the mean of the most populated bin in the plotted histograms (which will be further called "histogram-mode" in this work) considering x axis limited to the range (-10,10) and divided into 100 bins. These statistics were calculated for each contribution of each of the atom-types in the aim of accurately parametrize the values of the LogP contributions. This locally-built statistic "histogram-mode" was created purposely for this parametrization instead of using the classical mode statistic due to the characteristics observed in the distributions subject of study.

Despite the average of data has been widely exploited as an estimation of parameters, in this case its use was refused as it is sensitive to outliers, and it accounts for dispersed values that are not representative, which biased the final parametrization. Indeed, the lines in *Figure 4* show the mean being more utmost than the parametrized value for most atom-types. The reason for deliberately creating and using the "histogram-mode" statistic for the parametrization instead of the classical mode was so as to avoid non-representative modes due to random repetition of certain values. The classical mode takes the most repeated value of a dataset, whatever the distribution of values is. Drawback of using it is it lacks uniqueness, becoming uncertain when several values share the highest frequency. When data is continuous this is particularly problematic, as it is the case, because it is more likely not to have any one value that is more frequent than the other. LogP contributions values have four decimal places (to put it in numbers, there are 2 · 10^5 possible values considering (-10,10) range), and frequencies of atom-types in training set range from magnitude order 10 to 10^8. Indeed, another problem of the mode is that the exact value that is most repeated might not happen to be inside the range where most of the values lie (which was the interest of the parametrization in this work); that is why the mode would not provide a trustworthy measure of central tendency.

For example, *$LogP_{ele}$* of a certain atom-type might show a gaussian distribution of values centred around -0.5. It can randomly happen that no exact value around 0.5 is repeated but the value of

-1.3274 appears twice. If taking the classical definition of mode, the parametrized value of *LogP$_{ele}$* for this atom-type would be -1.3274 instead of being around -0.5. Using the mode to describe the central tendency of *LogP$_{ele}$* would be misleading.

Other measures of diversity that were calculated in order to extract a true overview of every single dataset of contributions were the maximum and minimum values and the standard deviation. This data allowed to define ranges of outlying values. *Table 1* displays the ranges of which outlying values were further discarded. Results of the parametrization can be addressed at section *4.1. PARAMETRIZATION RESULTS.*

Table 1. Ranges of thresholds established in the parametrization of hydrophobic contributions.

Hydrophobic contribution	*Range*
Total LogP	(-10,10)
LogPele	(-8,2)
LogPcav	(-1,1)
LogPvwa	(-1,3)

3.4. VIRTUAL SCREENING BENCHMARKING

Although some literature has been published about the assessment of ligand-based virtual screening[113] techniques, it is not yet a fully regulated field, lacking standard guidelines for dataset preparation, database sharing and quantitative evaluation of methods[114]. Though, benchmarking sets have been created as a solution addressed to this problem. In these datasets of molecules, most of the compounds are known to be inactive (decoys) and a small number of them are known to be bioactive for a certain target (hits)[113]. The active hits and the decoys should have alike physical structure while being chemically different. Indeed, it is essential that molecules in benchmarking libraries are structurally diverse[115].

The virtual screening method subject of evaluation (in this case we have two: VS with standard Pharmscreen® and with "AT" parameters) is supposed to retrieve the hits in the upper positions of the ranking leaving the decoys for the inferior spots[114]. The VS tools are considered better the more hits of the benchmarking set they are able to place at the highest positions in the ranking. Yet, it is also important that VS software suggests lead compounds with different chemical structures that render the same bioactivity on the target.

Regarding the problem of database sharing, datasets for benchmarking tests should be ideally publicly available and used for testing software, so that the results can be compared directly.

Validation set: DUD

The Directory of Useful Decoys (DUD)[115] was employed as the validation set. DUD is a publicly available database (http://dud.docking.org/) for benchmarking virtual screening that was designed to help test docking algorithms. Created by Huang et al.[115], DUD contains benchmarking sets for molecular docking for forty protein targets. Choice of the targets is based on a compromise between the accessibility to their 3D crystal structures information, their known and annotated ligands and other previous docking studies. The DUD database was formerly created for molecular docking software (structure-based virtual screening). However, *Good et al.*[38] adjusted the database to a new optimized version that serves also for benchmarking ligand-based virtual screening. In this work, 13 subsets of the DUD LIB VS 1.0 version described by *Jahn et al.*[34] and *Cheeseright et al.*[37] were downloaded as SDF files (and converted to mol2 for further analysis): *ace, ache, cdk2, cox2, egfr, fxa, hivrt, inha, p38, pde5, pdgfrb, src, vegfr2*. One reference ligand per subset was employed with the exception of cdk2, where two alternative conformations of the same ligand were reportedly found in the crystal structure. Both conformations of this structure were extracted and used as single references for separate virtual screening essays (details of DUD LIB VS 1.0 subsets in *Table 2*).

Proteins in the DUD library are drown from the ZINC database[44] and classified into six categories[115]: nuclear hormone receptors, kinases, serine proteases, folate enzymes, metalloenzymes and other enzymes. The 13 subsets employed here are diverse (six of them are kinases, one is a set of serine proteases, two are metalloenzymes and four subsets belong to the "other enzymes" category) to verify the ability of the tool subject of examination to handle several docking problems and active-site-characteristics. Pharmscreen® treats query structures as rigid molecules so for each dataset a conformational exploration is performed and 100 conformers per molecule are generated by means of RDKit.

3D structures of the proteins employed together with their binding ligands are available in the Protein Data Bank (PDB) except for *pdgfrb* and *vegfr2* (the structure of *pdgfrb* was created with homology modelling and *vegfr2* was crystallised in its apo form). According to *Jahn et al.*[34] and *Cheeseright et al.*[37] works, an overall amount of 1383 active molecules (ranging from 26 to 365 per subset) and 95316 decoys were used. The decoys had to be Lipinski rule-compliant to be included in DUD, and some of them appear in several subsets. Information of targets, active compounds and decoys used in this work can be found in *Table 2*.

Table 2. DUD LIB VS 1.0 subsets. a PDB code of the complexed crystal structures from which the search queries were taken.; b Angiotensine-converting enzyme; c Acetylcholinesterase; d Cyclin-dependent kinase; e Cyclooxygenase-2; f Epidermal growth factor receptor; g Factor Xa; h HI.

Target	*Active compounds*	*Decoys*	*Reference PDB code*[a]
ACE[b]	46	1796	1O86
AChE[c]	99	3859	1EVE
CDK2[d]	47	2070	1CKP
COX-2[e]	212	12606	1CX2
EGFr[f]	365	15560	1M17
FXa[g]	64	2092	1F0R
HIVRT[h]	34	1494	1RT1
InhA[i]	57	2707	1P44
P38[j]	137	6779	1KV2
PDE5[k]	26	1698	1XP0
PDGFrb[l]	124	5603	1T46
SRC[m]	98	5679	2SRC
VEGFr2[n]	74	2647	1FGI

Computation of hydrophobic contributions for validation set under standard Pharmscreen® and under "AT" pipeline

Firstly, hydrophobicity/hydrophilicity contributions were computed for DUD subsets and references (also called queries) with Pharmscreen® tool. MST continuum calculations in water and in octanol were calculated to generate the atomic contributions to the octanol/water transfer free energy for DUD library using Pharmscreen® with the same in-house MOPAC-6 version software and pipeline applied to the training set in the parametrization stage of the project.

The molecular descriptors calculated for each atom of each molecule are the partial charge and the fractional contributions to solvation free energy. Due to the high computational cost of the parameters calculations, for each molecule of the datasets (where the structures are represented by 100 conformations) the parameters were computed only for the lowest energy conformer and set equal for all other conformations. A multi-mol2 format file containing the parameters for each atom of all query molecules is obtained.

Secondly, hydrophobicity/hydrophilicity contributions for DUD subsets and queries were calculated by means of "AT". This pipeline does not involve the time-consuming computation of hydrophobic contributions since it assigns the parametrized values to the atom-types identified in each molecule. The output of "AT" methodology is also a multi-mol2 format file containing the parameters for each atom of all query compounds.

Virtual screenings on validation set

Virtual screening with the same settings was performed twice on the DUD subsets, initially using hydrophobicity computations resulting from standard Pharmscreen®, then with the parameters of "AT". Both experiments were executed in a remote cloud service machine of 16 threads.

Pharmscreen® carries out a molecular alignment and selects the orientation of the molecule that maximizes the hydrophobic overlap with the reference compound. Both reference's and query molecule's hydrophobicity centres are superposed and their hydrophobic dipoles (quantified in matrix-like objects called tensors) are aligned. The virtual screening performed in this work accounted for three terms: electrostatic hydrophobicity, cavitational hydrophobicity, and hydrogen bonds. By means of these three forcefields we consider the energy needed to generate the solute charge distribution of the compounds in the solvent (electrostatic hydrophobicity), the energy related to the steric properties of the molecules (cavitational hydrophobicity) and the potential donor/acceptor of hydrogen bonds character of atoms in the structures subject of comparison. The force fields were weighted 0.15, 0.55 and 0.3 respectively since internal studies developed within *Pharmacelera S.L.* activity framework reported to be the best weighting combination[116]. The work of *T. Ginex et al (2016)*[117] illustrated that the cavitational hydrophobicity component and the Van der Waals one are correlated. Accordingly, introducing Van der Walls hydrophobicity as an extra force field would introduce undesirable noise rather than adding any useful information, hence the Van der Waals component was not accounted for.

Hydrophobic similarity was evaluated for each alignment with the Tanimoto function (Eq. 12) over all points in a 3D grid. The Tanimoto coefficient is equal to the Jaccard coefficient of two sets A and B (also known as intersection over union). For every particular query molecule, a

Tanimoto coefficient (T_k) is obtained from each of the *N* number of forcefields *k* accounted in the virtual screening, together with a final Tanimoto coefficient sum (S) resulting from the addition of the weighted (λ_k) force fields Tanimotos (Eq. 13). Molecules in the libraries are ranked according to their Tanimoto sum *S*. The virtual screening output consists on a ranking of the best conformer of each of the structures screened in the dataset ordered by their Tanimoto sum coefficient.

$$T(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} \quad (12)$$

$$S = \sum_k^N \lambda_k T_k \quad (13)$$

Virtual screening performance metrics

Virtual screening tools perform better the more active compounds of the validation set they rank at the top positions. The evaluation of this performance is not standardized yet, but some metrics have been developed, such as the enrichment factor, the ROC curve and AUC[115].

Enrichment factor (EF). A popular metric that benefits from being both intuitive and straightforward to interpret is the so-called *enrichment factor*[114,115]. The EF was formerly defined as the ratio of the observed active compounds relative to the total number of molecules tested in the top few percent of a virtual screen to that expected by random selection. If *n* is the total number of hits and *N* represents the total number of molecules in the database, then there are n_a actives among the $N_{x\%}$ molecules in the first x% of the database screened and EF is given by Eq. 14:

$$EF_{x\%} = \frac{n_a / N_{x\%}}{n / N} \quad (14)$$

EF is such a universal metric due to its intuitive interpretation related to the purpose of the VS itself, i.e. the ability to select a subset of molecules with promising chances of drug discovery. However, EF value is easily influenced by the number of actives in the dataset. It becomes then a measure of both the method and the experiment, rather than an intrinsic property of the method, which makes it a poor metric[114]. A simple way to overcome the issue is to make the enrichment factor refer only to the fraction of inactive compounds, instead of all the dataset. Therefore, EF can be obtained by dividing the sensitivity of the experiment by the fraction of false positives (decoys) retrieved[34]. This way the new statistic (ROC enrichment factor- ROCE) represents the ability of the test to discriminate two populations, based on a standard statistical measure called Receiver Operating Characteristic (ROC)[118,119].

ROC, ROCE and AUC. In statistics, a receiver operating characteristic curve, i.e. ROC curve, is a plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting true positive rate (hits rate in this work) against false positive rate (decoys rate) at various threshold settings. True-positive rate is also known as sensitivity, recall or probability of detection whereas false-positive rate is also known as the fall-out or probability of false alarm[120] and can be calculated as (1 − specificity).

Hence, the *Y* axis of the ROC curve displays the sensitivity of the test (discovered hits fraction) while the *X* axis shows 1 - specificity (discovered decoys fraction)[34]. X% is the false positive rate or decoys rate found in a chosen range of the experiment. *Figure 3* presents examples of random enrichment, which generates a graph resembling a linear function f(x) = x, together with better and worse outcomes of tests. ROCE curves resulting from Standard and AT virtual screening on validation set can be found at *Figure 21* and *Figure 22* in section *7. APPENDIX*, subsections *D* and *E*.

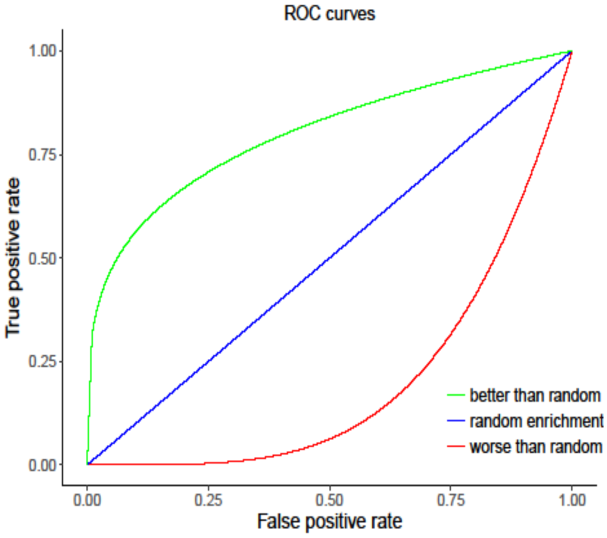


Figure 3. ROC curves illustrating an approximation of random enrichment, better than random and worse than random enrichment.

ROC Enrichment (ROCE) is then obtained by dividing the *y* value by the *x* value of a ROC curve at certain decoys fractions (Eq. 15). ROCE decoys fractions (placed on the *X* axis) at which values are suggested to be reported are 0.5%, 1.0%, 2.0%, 5.0% [114] in order to render the early enrichment of the test, and so has been reported from the virtual screenings performed in this work.

$$ROCeX\% = \frac{\frac{hits}{hits+decoys}}{\frac{decoys}{hits+decoys}} = \frac{sensitivity}{1-specificity} = \frac{y\ value\ ROC\ point}{x\ value\ ROC\ point} \quad (15)$$

Nicholls[114] strongly advocates the use of ROC[118] together with the area under the ROC (AUC)[118], which are commonly applied in other fields including statistical analysis, data mining, or machine learning techniques[120]. Yet, because of the universality of EF, in this work both metrics have been considered.

AUC stands for the Area Under the ROC Curve. AUC is used to evaluate the performance of the VS software across the entire ranking of hits[37]. A test producing random enrichment would get an AUC ~ 0.5; Better-than-random enrichment will get AUC > 0.5 as there is a higher probability of assigning high scores to true hits; worse-than-random experiments would get AUC < 0.5. An ideal distribution of active structures and decoys would get a value of 1.0. Though, AUC should not be used as an independent metric for comparing VS experiments since it does not reveal any information about the ROC curves. From two experiments showing the same AUC value, one may perform better among the early top ranked spots and the other in lower ranges of decoys fraction, but this information could not be told from AUC. However, in this work it was considered an appropriate metric when accompanied by ROC curves and enrichment factors, providing information about the general performance of the virtual screening.

The two virtual screenings performed generated rankings of compounds for each DUD subsets were processed with a python script belonging to *Pharmacelera S.L.* intellectual property that cannot be publicly available. This script computed the VS performance metrics detailed above. ROCE and EF at 0.5%, 1%, 2% and 5% were calculated and a ROCE curve was created for each subset. This script addresses python modules such as Matplotlib, OpenPyXL, os, sys, getpass, subprocess, argparse, numpy and datetime module. The metrics are printed in a XLSX format file with tables and automatically-created bar charts to ease an overview and a first comparison between VS performance from standard PharmScreen® and the new pipeline.

3.5. PROCESSING AND ANALYSIS OF RESULTS.

Several aspects of the resulting data were processed and analysed for a correct understanding of the outcomes.

Regarding the VS performance, the differential results on ROCE and EF on Standard Pharmscreen® and "AT" experiments were analysed. The differences between the number of hits, and which hits were positioned where in the rankings, together with the differences within the different percentages of ROCE and EF that were reported (0.5%, 1%, 2%, 5%) was examined using Pymol and Python parsing scripts in the search of chemical groups, parametrization errors or molecular features that could explain these differences (see *Figure 10* and *Table 3* in section *4.3. VIRTUAL SCREENING PERFORMANCE*). The force fields projections of the molecules subject of inspection were computed and scrutinized as well. R and Rstudio were also utilized in the search of correlation, significance of differences, statistics computations and plotting. R and Rstudio were used also for computing speedup and execution time plots.

Additionally, Python scripts were programmed in order to analyse in which extent "AT" is assuming a bias between the conventionally computed hydrophobicity values by Pharmscreen® and the atom-type-parametrized ones implemented in "AT" pipeline. To study the error rate that was assumed in the parametrization for each atom-type, deviation rates were calculated for the DUD experiment hydrophobic values under Standard Pharmscreen® computations. There are several statistics to measure the diversity and error or deviation rate of samples in respect to the parameter assumed to be an acceptable prediction of the real values. Standard deviation (SD) is widely used in diversity studies to measure the spread of data around the mean (Eq. 16). However, in this study it was not the mean the statistic taken as the estimation of the hydrophobic contributions, but the "histogram-plot". Root-mean-square- error (RMSE) of an estimator $\hat{\theta}$, also called root-mean-square-deviation (RMSD) is defined as the root of the mean square error (MSE) of an estimator $\hat{\theta}$ (Eq .17). Its formula is similar to the SD (Eq. 18) as both are square roots of squared differences between some values. Nonetheless, RMSE is used to measure distance between some values and estimation for those values. This statistic is generally applied in the purpose of analysing the error of a certain estimation. In this project, it was used to measure how much the parametrized values differ from the computed ones. Since RMSE is defined as the root of the MSE (Eq. 18), this statistic was also computed and analysed.

$$SD = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} \quad (16)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\theta_i - \hat{\theta})^{\,2} \quad (17)$$

$$RMSD(\hat{\theta}) = \sqrt{MSE(\hat{\theta})} \quad (18)$$

Indeed, RMSE has been extensively applied in bioinformatics to measure the average distance between the atoms of superimposed proteins, or to measure the difference between a crystal conformation of a protein and its docking prediction.

The script for computing MSE and RMSE and some other scripts involved in the analysis of molecules with high values of RMSE are intellectual property free and thus publicly available at the following github repository: https://github.com/elisagdelope/TFM_scripts-IP-free.git

4. RESULTS AND DISCUSSION

4.1. PARAMETRIZATION RESULTS

Hydrophobicity/hydrophilicity atomic contributions of the molecules in the training set were computed with the locally-modified version of MOPAC6 as explained in section *3.2. COMPUTATION OF HYDROPHOBIC ATOMIC CONTRIBUTIONS FOR TRAINING SET AND IDENTIFICATION OF ATOM-TYPES*.. The data obtained for each atom-type and each of the electrostatic, cavitational and van der Waals contributions was studied in the aim of parametrization. The highlights of the statistical analysis of $LogP_{ele}$ and $LogP_{cav}$ (the contributions that were further used in the VS) are displayed in *Figure 4* and *Figure 5*. Additionally, in order to compute the parametrized value (‘*histogram-mode’*) – defined in *3.3. PARAMETRIZATION OF THE HYDROPHOBICITY PER ATOM-TYPE* -, a histogram per atom-type and per LogP contribution was displayed. A couple of examples can be found in *7. APPENDIX*, subsection *A*.

The lines in *Figure 4* and *Figure 5* graphs represent the mean and the parametrized value, which corresponds to a statistic defined on purpose as the average of the values belonging to the most populated range of a histogram x-limited (-10,10) divided into 100 bins. The standard deviation and the percentage of outliers are displayed as shadows illustrating the diversity among data of each atom-type.

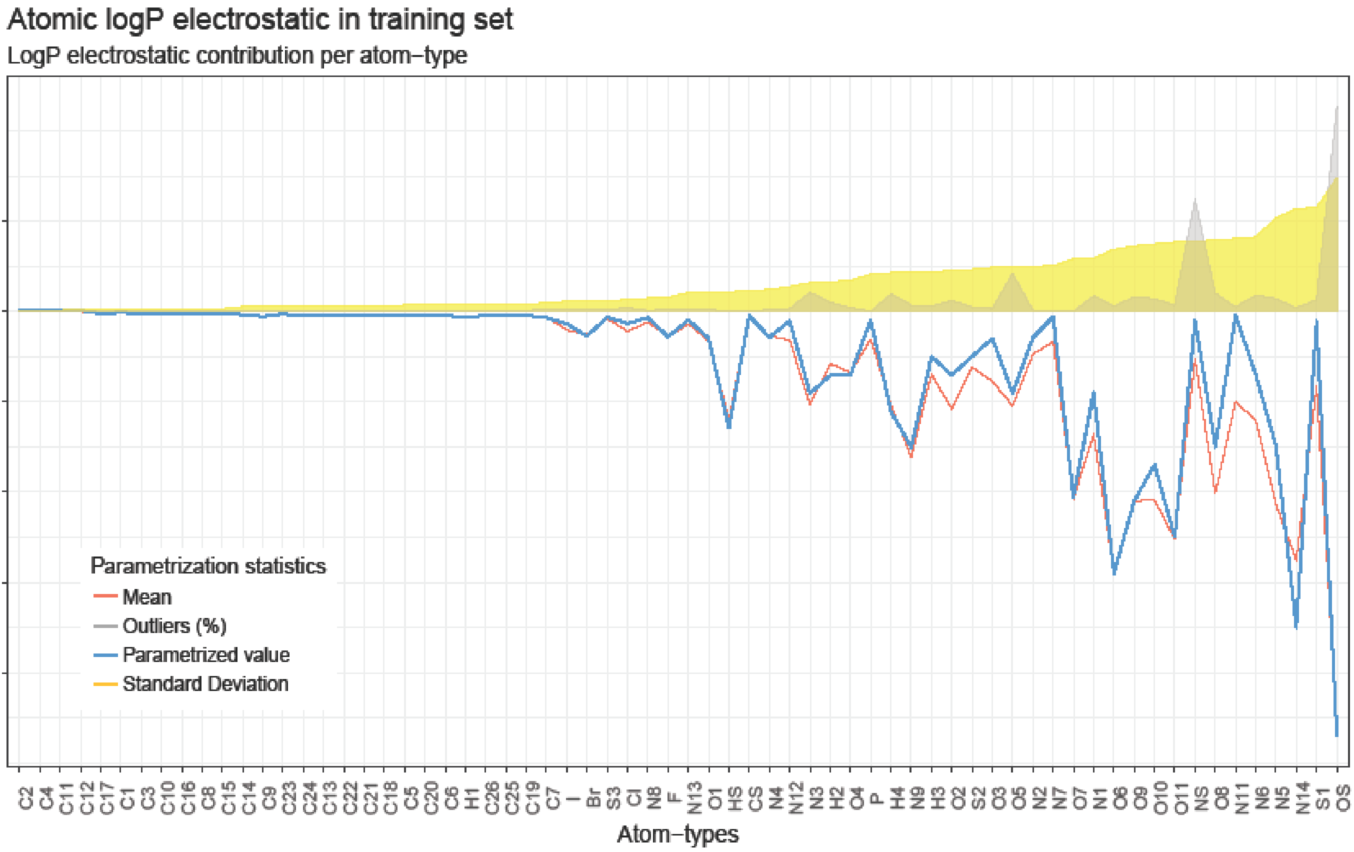


Figure 4. Parametrization statistics along atom-types LogP electrostatic data. The mean and the parametrization value are displayed in lines; outliers ratio and standard deviation show as shadows.

Figure 4 of LogP electrostatic parametrization statistics reveals how well characterized are the Carbon atom-types, since they have little standard deviation among their electrostatic hydrophobicity/hydrophilicity. However, heteroatoms show a notable standard deviation and outliers ratio (considering the outlying threshold settled (-8,2) was still generous for LogP electrostatic standards). This fact supports the reason why the total mean was not taken for the parametrization, as it would have accounted for those very dispersed values that are not representative and could have biased the final parametrization. Indeed, the lines in the plot show the mean being more utmost than the parametrized value for most of the atom-types.

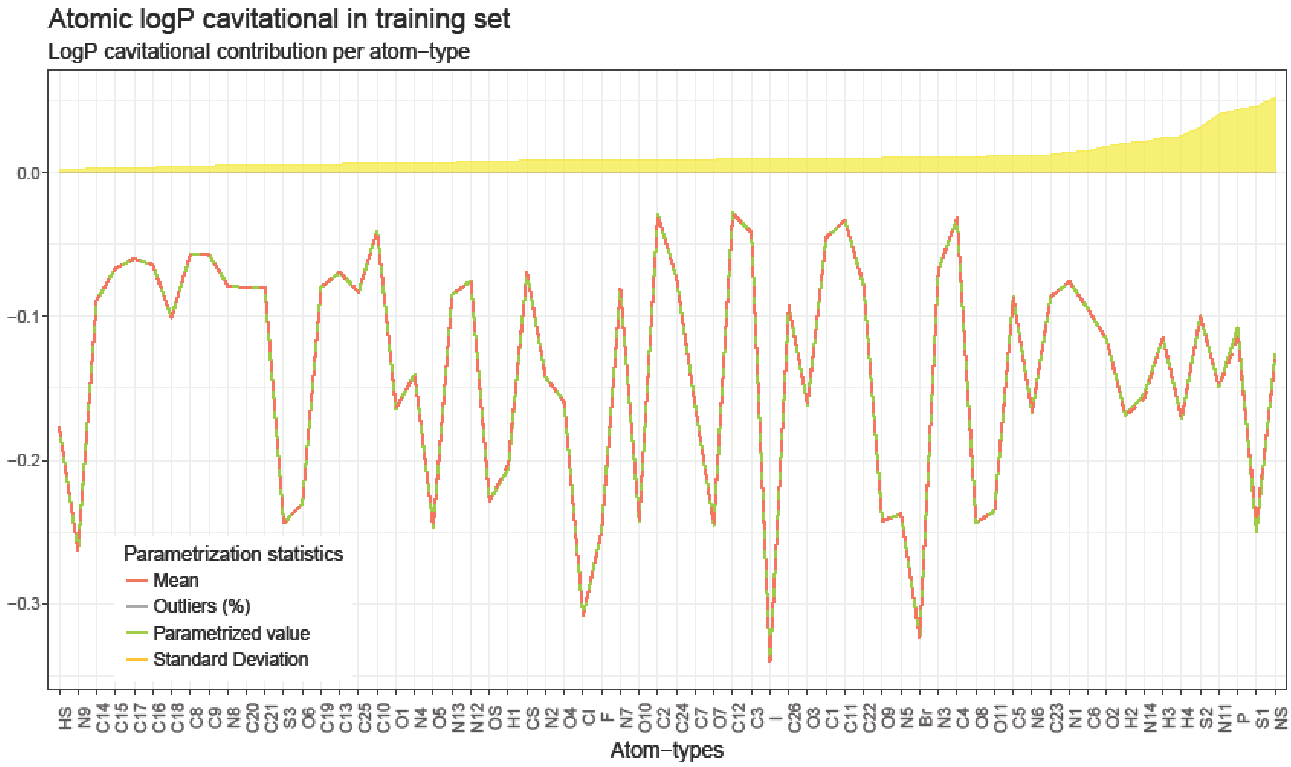


Figure 5. Parametrization statistics along atom-types LogP cavitational data. The mean and the parametrization value are displayed in lines; outliers ratio and standard deviation show as shadows.

Figure 5 exhibits LogP cavitational values are extremely homogenous. Outliers shadow does not appear in the picture since there are no outliers, hence the line is covered by y = 0 axis. Standard deviation is also very low (maximum value of SD = 0.05), and the mean and the parametrized value are so close the lines overlap for most of the atom-types data. Let's note that the cavitational component of the partitioned hydrophobicity/hydrophilicity estimations accounts for the work required for creating a cavity shaped to accommodate the solute in the solvent. This definition explains how LogP$_{cav}$ is affected by the close environment of each atom, which is defined by the atom-type. Therefore, the variance within LogP$_{cav}$ values of each atom-type was expected to be low as it resulted.

4.2. COMPUTATION OF HYDROPHOBICITY PARAMETERS SPEEDUP

The new "AT" pipeline was expected to shorten the otherwise long time required for the procurement of hydrophobicity/hydrophilicity values. Execution time for the computation of these values was reported and plotted in *Figure 20*, which can be found at section *7. APPENDIX*, subsection C. The bar chart below (*Figure 6*) displays speedup of hydrophobicity/hydrophilicity values calculation of "AT" experiment over the Standard hydrophobicity calculation. Speedup was computed according to the formula described in Eq. (19):

$$S = \frac{Execution\ time_{Standard}}{Execution\ time_{AT}} \quad (19)$$

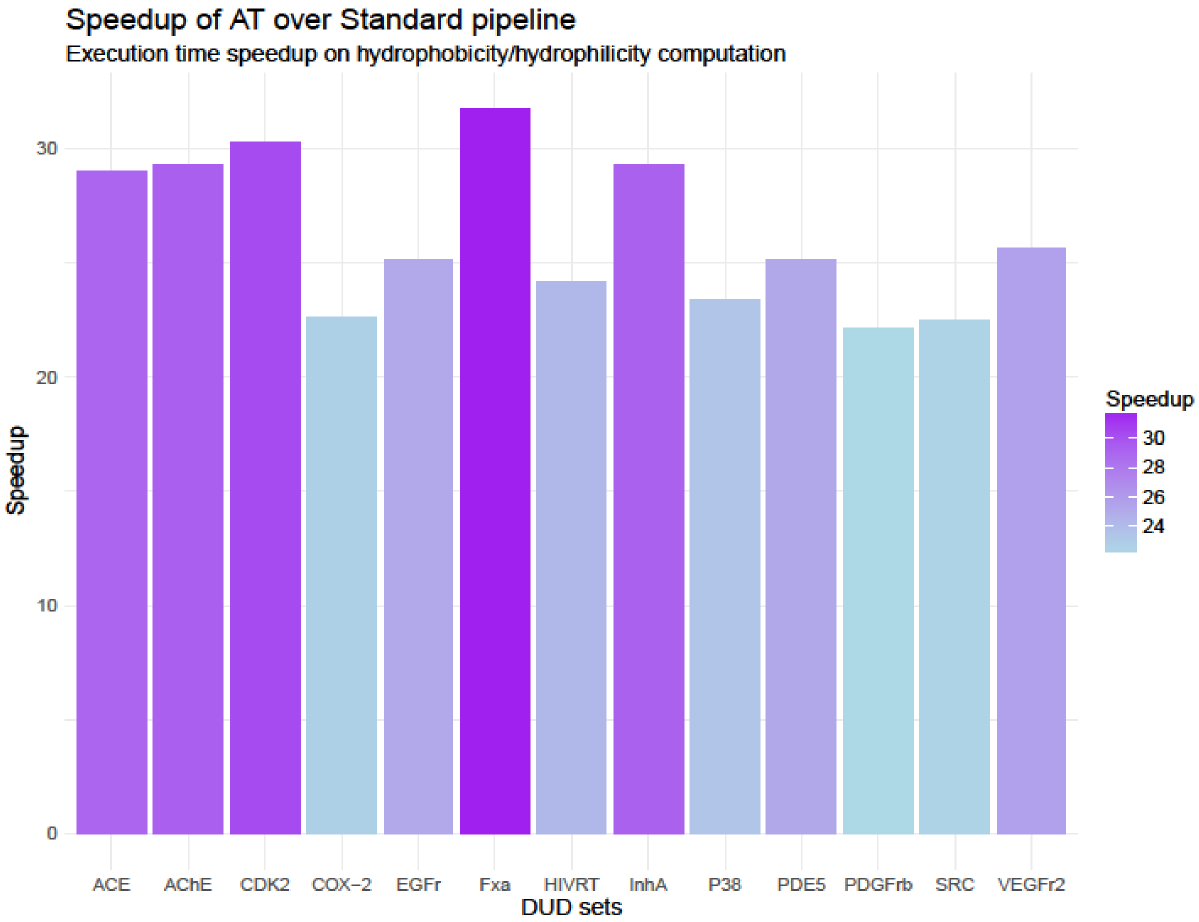


Figure 6. Speedup of AT parameters assignation in respect to standard Pharmscreen® hydrophobicity computation time.

The calculation of parameters in AT experiment proves to be outrageously faster than Standard Pharmscreen® parameters computation, which fits with the expectation, since in the AT experiment parameters are actually assigned and not computed. In *Fxa* set the speedup achieves ratios of 30x, and all of them overcome the 20x speedup. The interpretation of these results is that assigning parametrized values can turn the procurement of the hydrophobicity/hydrophilicity values up to 30 times faster. Differences within the speedup of the DUD sets might be due to the number of molecules and atoms per set and the conformation of molecules (the lower energy the structure has the less time MOPAC needs to compute hydrophocity/hydrophilicity). Indeed, *COX-2* and *PDGFrb* sets include compounds with a substantial number of aromatic rings, which are considered rigid and stable structures. Hence, the computation time using the AT pipeline does not improve as much as in other sets.

Also, the fact that MOPAC sometimes fails to compute the parameters of certain conformers may impact the speedup. In such failure cases, the algorithm does not converge but still takes up to 2 hours of computation per failed molecule. If MOPAC happens to have a different failure ratio among the subsets, let's put the example it fails to compute hydrophobicity parameters for several molecules that randomly belong to a certain set, the execution time for computing parameters of this given set soars and so would do the Speedup. When benchmarking AT (or any other technique for computing hydrophobic parameters) in other validation sets, MOPAC failure ratio is an important detail to note when assessing Speedup.

4.3. VIRTUAL SCREENING PERFORMANCE

The virtual screening results from the Standard and the AT experiments were analysed so that to compare performance metrics. The EF and ROCe were reported at 0.5%, 1%, 2% and 5%. Additionally, AUC (ROCe 100%) was computed and ROC curves were plotted (see *Figure 21* and *Figure 22* at section *7. APPENDIX*, subsections *D* and *E*). The following bar charts compare EF at 0.5%, ROCe at 0.5% and AUC in the Standard and the AT experiment in each of the DUD sets. Additional EF and ROCe plots for 1%, 2% and 5% of virtual screening can be found in *Figures Figure* 14 *- Figure* 19 at section *7. APPENDIX*, subsection *B* for more detail. Additionally, Wilcoxon Signed-Rank Test was performed in R and Rstudio for EF 0.5%, ROCe 0.5% and AUC values to statistically support the significance of the differential or similar appearance between Standard and AT performance. Choice of the Wilcoxon signed-rank test was because it is a paired difference, non-parametric statistical hypothesis test. It was used as an alternative to the paired Student's t-test because the population (performance metrics values in the different DUD sets) could not be assumed to be normally distributed.

The Wilcoxon signed-rank test can determine whether two dependent samples were selected from populations having the same distribution, meaning if H0 is not rejected, the distribution of the tested metric value is the same for Standard and AT experiment (this is, no significant differences in their performance).

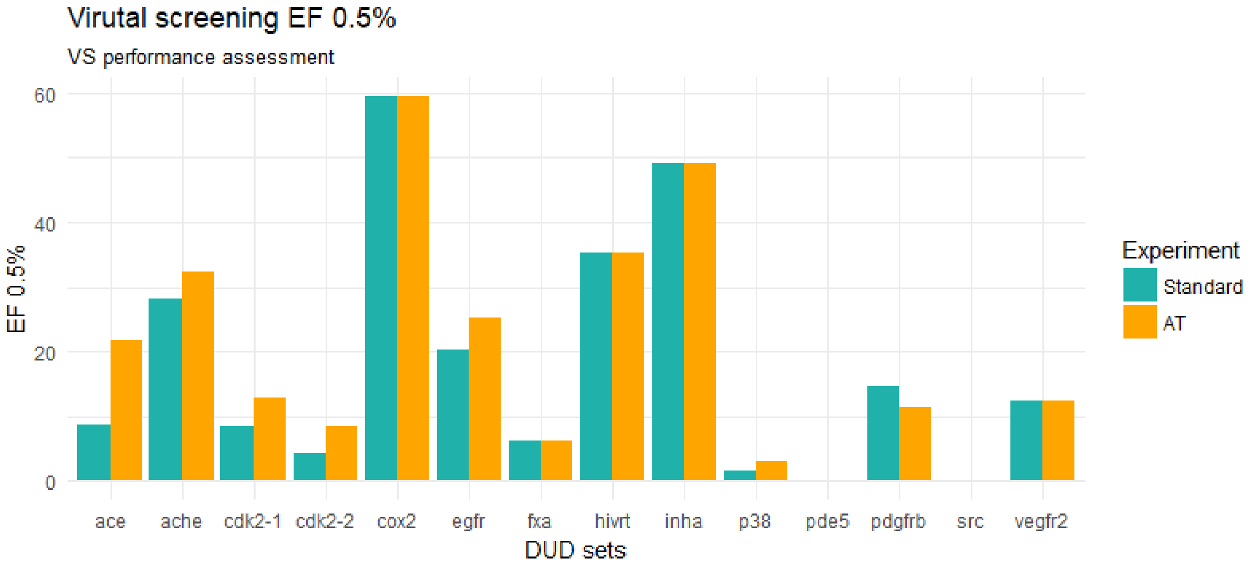


Figure 7. Comparison of EF 0.5% on Standard and AT virtual screening experiments on validation set.

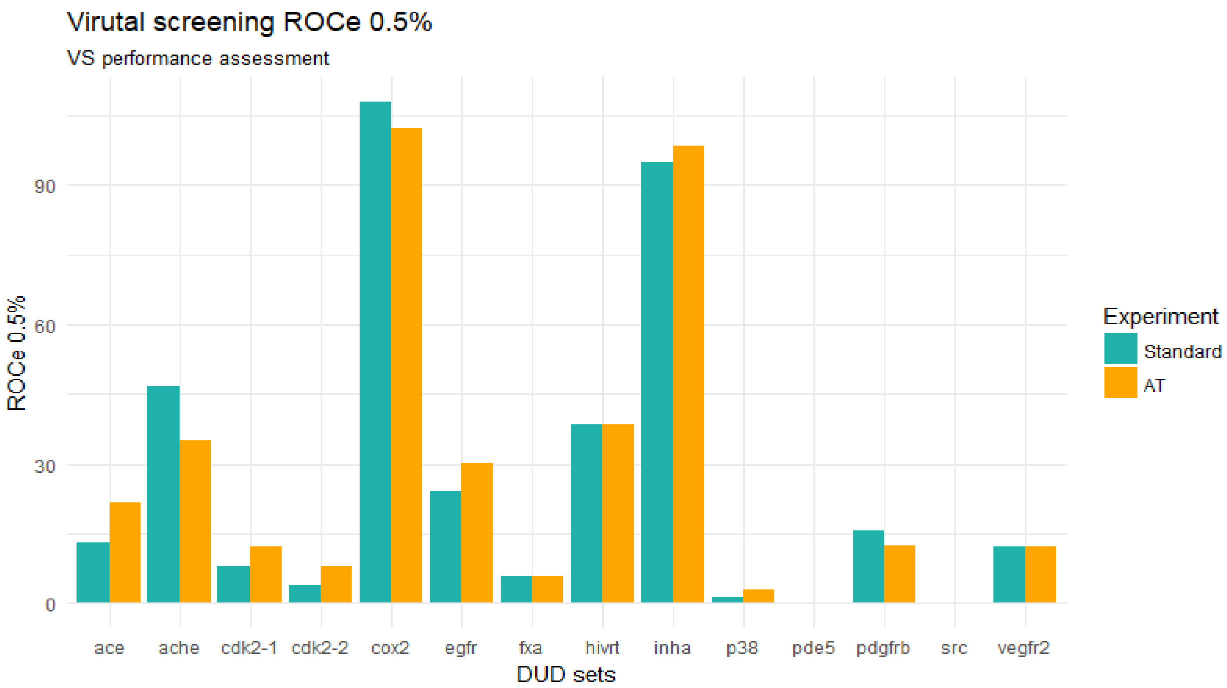


Figure 8. Comparison of ROCe 0.5% on Standard and AT virtual screening experiments on validation set.

Both EF 0.5% and ROCe 0.5% charts (*Figure 7* and *Figure 8*) display the results at the early positions of the VS ranking. This is, the ability of the VS to place the active molecules of each set at the top 0.5% spots of the ranking. In these graphs one methodology and the other show to perform very much alike. Wilcoxon signed-rank test for EF 0.5% data along the sets resulted in a p-value of 0.052, whereas p-value from Wilcoxon signed-rank test for ROCe 0.5% data was 0.55. Both cases showed p-value > 0.05, hence the null hypothesis cannot be rejected, and it is assumed that the distribution of EF 0.5% values on Standard and AT virtual screenings is not different, and same applies to ROCe 0.5% distribution of values on Standard and AT experiments.

Yet, the difference of one order of magnitude between both p-values, while bar charts looking similar, drew our attention. This difference is due to the procedure of the Wilcoxon signed-rank test. Sample size is already small in both EF 0.5% and ROC 0.5% (only 14 pairs of data). But, since the test excludes pairs whose difference is zero, this is, the ones for which the metric has the same value, when testing for EF 0.5% data, the test is only considering difference between 7 sets (cox2, fxa, hivrt, inha, pde5, src and vegfr2 are excluded for sharing the same value in both experiments). In the case of ROCe 0.5%, five sets (fxa, hivrt, pde5, src and vegfr2) are excluded for sharing the same value, so the test is considering differences between 9 sets. These two sets are thought to be causing the differential p-value between both tests. Though, the important fact is that in both metrics, the performance of AT in respect to the Standard pipeline is not significantly different.

AUC in *Figure 9* reveals that for most of the DUD subsets the overall performance of the hydrophobicity/hydrophilicity-based VS is again similar for both Standard and AT implementation. Wilcoxon signed-rank test was applied and resulted on a p-value = 1, meaning the null hypothesis cannot be rejected again and so the distribution of AUC values along the different DUD sets on the Standard Pharmscreen® is not significantly different from the distribution of AUC values along the DUD sets on the AT virtual screening.

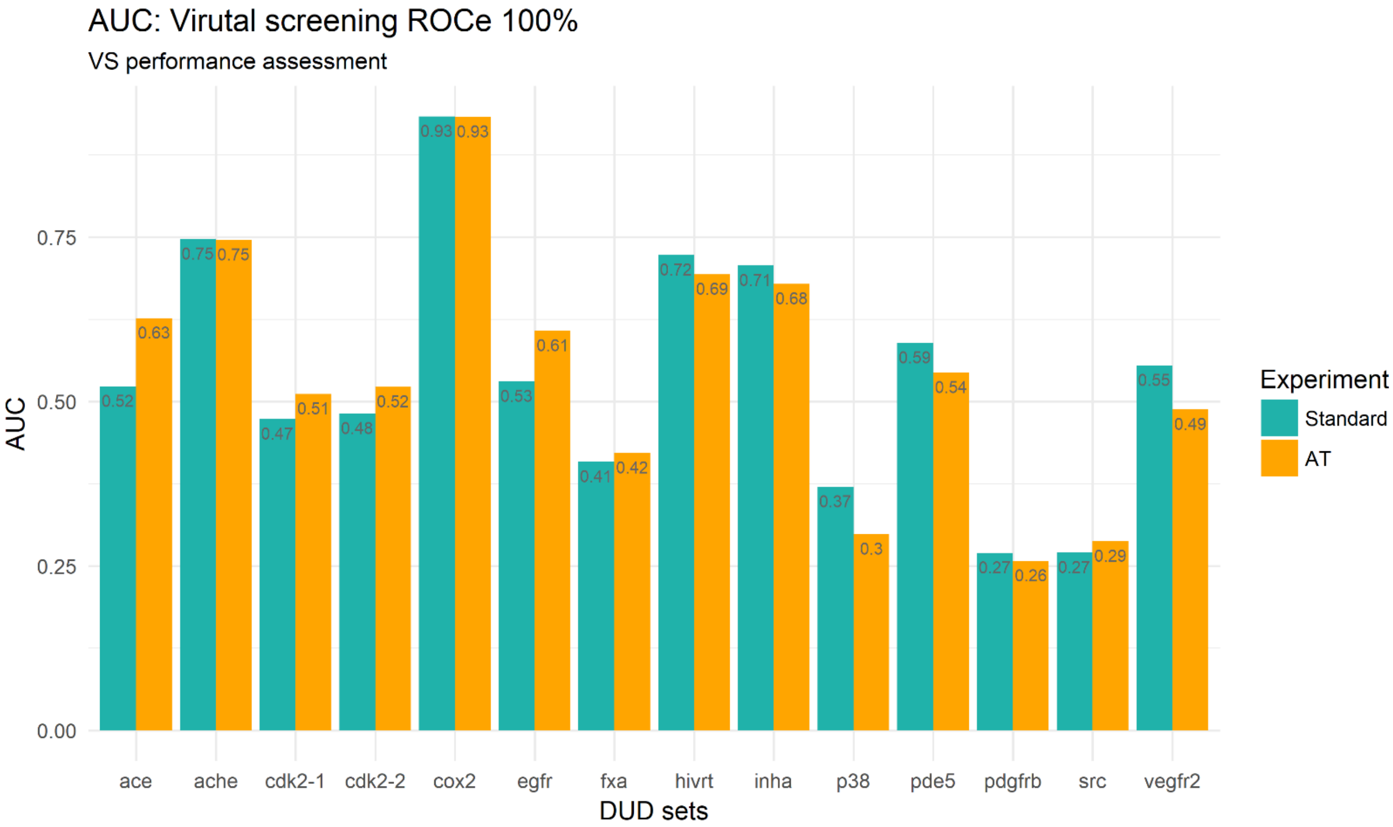


Figure 9. Comparison of AUC on Standard and AT virtual screening experiments on validation set.

Such results of no significant difference among the performance metrics are considered to be very positive as the vast savings in computation time do not correspond to loss in accuracy.

Surprisingly, the analysis of the performance exposes that in some sets the AT experiment happens to achieve even more hits at the first positions. Such is the case of ACE test, which was treated and deeply examined to track the positions of those hits that were found in the first 0.5% spots at the AT ranking but in lower positions at the Standard ranking in the search of specific chemical features that could be responsible of this differential sorting. These compounds (see case of ZINC03809801_1 in Figure 10 and Table 3) were further studied in Pymol to review the alignment and distribution of the force fields used in the virtual screening.

This check-up exposed that active molecules in ACE set include, besides a carbon-based scaffold, a good number of heteroatoms in their structure while the reference structure (PDB code: 1O86) does not. Therefore, the heteroatoms (usually having lower utmost $LogP_{ele}$ values -considering electrostatic hydrophilicity is negative-) from the active molecules will typically align with Carbon atoms (whose $LogP_{ele}$ values are generally close to zero) of the reference ligand. Besides, the parametrization of $LogP_{ele}$ values exposed some atom-types (mainly heteroatoms) were not optimally characterized and showed diversity within their distribution of hydrophilicity values (together with a good number of outlying values), see case of N11 in *Figure 13*. The statistic taken for parametrization (histogram-mode) was generally a value closer to zero than the mean, which was sensitive to noisy lower values (check *Figure 4*).

Accordingly, the AT prediction of hydrophobicity/hydrophilicity values (parametrized values) of heteroatoms is close to the estimated ones for the hydrophobic carbon chain, whereas MST-model quantum-mechanic computations from MOPAC can give more utmost $LogP_{ele}$ values to heteroatoms that will significantly differ from the ones given to a hydrophobic carbon chain. As a result, the force field of electrostatic hydrophilicity differs more in the Standard implementation, where hydrophobicity/hydrophilicity is computed from scratch.

Figure 10 illustrates the case of the active compound ZINC03809801_1 and is useful to show the underlying explanation for finding higher number of active molecules in the top positions of the ranking, this is, presenting higher similarity. The figure evidences resemblance in the alignment of the structures of the ACE reference and an active compound (named ZINC03809801_1) that appears in the 8th position (thus belonging to the first 0.5%) in the AT ranking but falls to the 71th position in the Standard ranking. However, the Sulfur in ZINC03809801_1 has an estimated value of $LogP_{ele}$ = -0.1079 and aligns to a Carbon with $LogP_{ele}$ = -0.0122 in the AT experiment, whereas in the Standard VS the Sulfur has $LogP_{ele}$ = -0.5423 and the Carbon it aligns to has a value of $LogP_{ele}$ = -0.0073. The similarity regarding electrostatic hydrophilicity is therefore higher between the active compound and the reference in the AT experiment than in the Standard. The corresponding force field has been projected over the structures in *Figure 10* and it is straightforward to realise how the electrostatic hydrophilicity is more similar in the AT experiment than in the Standard.

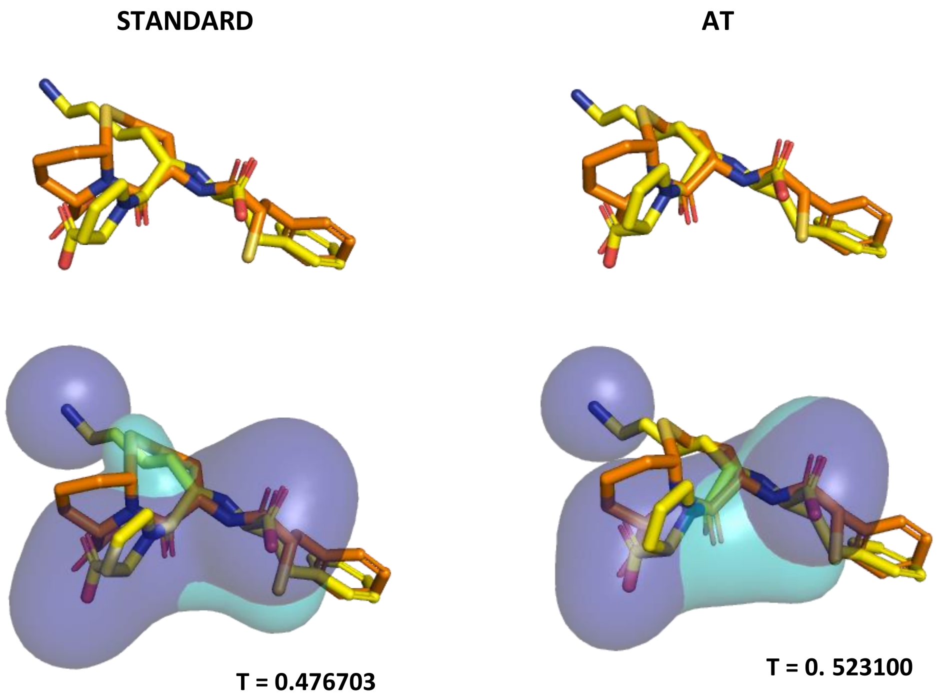


Figure 10. Alignment of the reference (yellow) and ZINC03809801_1 (orange) on Standard and AT experiments. Electrostatic force field projections of the reference (dark blue) and ZINC03809801_1 (teal blue) in Standard and AT together with their weighted tanimoto (T) are show as well.

This fact is thus reflected in the Tanimoto coefficient, which improves significantly in its electrostatic hydrophobicity component (check *Table 3*) with the AT estimation and that is the reason why more active compounds are found at the first spots in the AT ranking than in the Standard ranking.

Table 3. Tanimoto coefficients and ranking positions of molecule ZINC03809801_1 in Standard and AT virtual screening experiments on validation set.

Experiment	*Ranking spot*	*Hele Tanimoto*	*Hcav Tanimoto*	*HBonds Tanimoto*	*Weighted Tanimoto*
AT	8	0.529121	0.591120	0.395388	0.523100
Standard	71	0.387513	0.540115	0.405043	0.476703

4.4. PARAMETRIZATION ASSESSMENT IN VALIDATION SET

The MSE and RMSE between the parametrization of hydrophobicity/hydrophilicity contributions in AT and the Standard experimental values per atom-type were computed in the purpose of analysing the error assumed in the AT approach for the validation set. $LogP_{cav}$ and $LogP_{vwa}$ showed very little errors (mean of $RMSE_{cav}$ per atom-type = 0.01; mean of $RMSE_{ele}$ per atom-type = 0.03), hence the parametrization for these two is assumed accurate (as small as the SD present in training set, see *Figure 5*). Though, $LogP_{ele}$ exhibits significant diversity for certain atom-types (mean of $RMSE_{ele}$ per atom-type = 0.49). The RMSE of LogP electrostatic was plotted in the *Figure 11*.

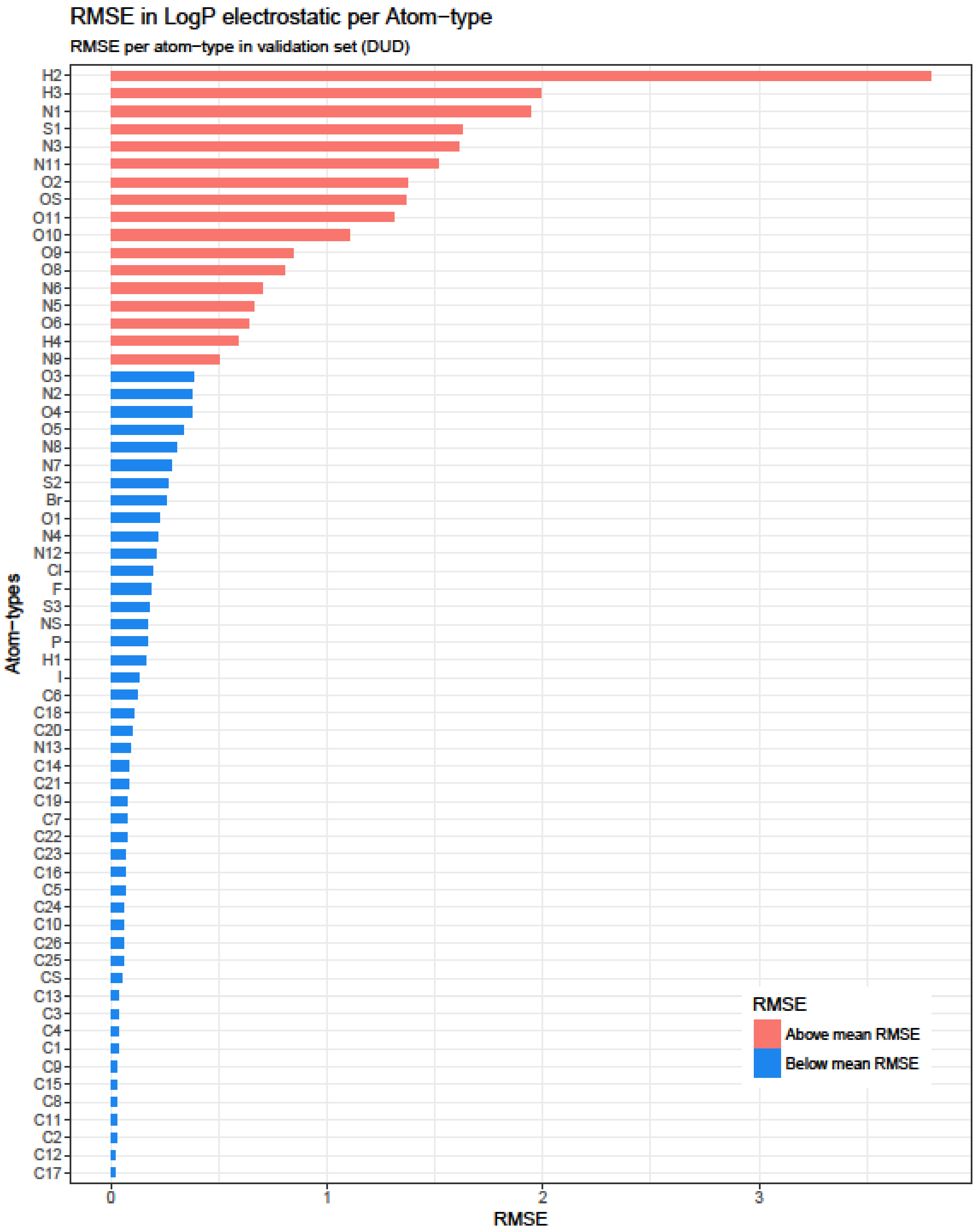
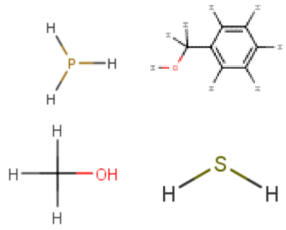
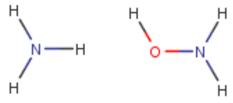
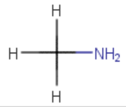
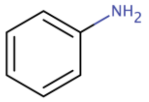
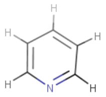
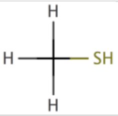


Figure 11. RMSE in LogP electrostatic contribution per atom-type on validation set.

Some of these atom-types were in fact expected to have a considerable MSE and RMSE since they already showed notable SD and/or outliers ratio in the training set (being the case of N1, S1, N11, OS, N3 or O11, see these atom-types statistics in training set in *Figure 4*). Yet H2 and H3 appear to be the atom-types with the highest RMSE (and indeed highest MSE) while having little SD and outliers ratio in the training set (*Figure 4*). Both H2 and H3 atom-types were tracked within the DUD compounds in order to find whether this surprisingly high RMSE was due to a biased parametrization value or to an error in the Standard hydrophobicity/hydrophilicity computation developed by the locally-modified version of MOPAC6. A description of the atom-types showing higher RMSE can be found in *Table 4*.

The study of molecules having extreme outlying values for LogP electrostatic detected that MOPAC was computing incorrect values of electrostatic LogP for certain atom-types in some molecules (extremely biased values such as 275.7216 or 280.0602 for H2 $LogP_{ele}$ and -224.8309 for H3 $LogP_{ele}$). This is the reason why these atom-types show such a large RMSE in the validation set whereas they showed small SD in the training set. As a matter of facts, MOPAC hydrophobicity calculations depend on the elected conformation (typically the one with the lowest energy), and these parameters are assigned to its conformers. Thanks to this inspection, this work detected that in some cases the lowest-energy-conformation can converge to LogP values that do not correspond to a successful MOPAC computation.

Table 4. Description of atom-types with higher RMSE in validation set.

Atom-type	*Description*	*SMARTS code*	*Structure(s) including the atom-type*
H2	Alcohol	‘[#1]O[CX4]’,‘[#1]Oc’,‘[#1]O[!(C,N,O,S)]’,[#1][!C,N,O)]’	
H3	Amine	‘[#1][#7]’, ‘[#1]O[#7]’	
N1	1º amine	‘[NH2+0]A’	
N3	1º aromatic amine	‘[NH2+0]a’	
N11	Unprotonated aromatic	‘[n+0]’	
OS	Oxygen supplemental	‘[#8]’ not matching any basic O type	-
S1	Aliphatic	‘[S-0]’	

5. CONCLUSIONS

The conducted work illustrates a new version of Pharmscreen® hydrophobicity-based virtual screening pipeline called "AT" in which the hydrophobicity/hydrophilicity contributions are assigned from parametrized values. The proposed parametrization was accomplished with data from a training dataset of compounds, then benchmarked with a validation set and compared to the standard Pharmscreen® virtual screening. Accordingly, the present project concluded the following statements:

- The parametrization defined works generally properly, yet the characterization of certain heteroatoms could be improved.

- LogP electrostatic component of the MST-model hydrophobicity/hydrophilicity contributions plays a key role in hydrophobicity-based virtual screening. LogP_{ele} presents a wider range of values since it is influenced by the whole molecule while LogP_{cav} and LogP_{vwa} are not.

- The general performance of AT pipeline and the standard Pharmscreen® are similar in the validation set.

- AT pipeline allows to save large amounts of time in the hydrophobicity parameters computation stage of hydrophobicity-based virtual screening achieving speedup levels up to 30x.

The main objective of this project was accomplished, as the designed new low time-consuming pipeline for hydrophobicity-based virtual screening has demonstrated to be certainly fast and effective without significant loss of accuracy.

6. FURTHER PERSPECTIVES

Although the goals of the project were achieved with success, further analysis and research on the characterization of the atom-types and its impact on the performance of virtual screening would improve the methodology developed in this work. Several perspectives are proposed for future research directions.

This work studied the RMSD per atom-types between the parametrization and validation set. An interesting approach would be to compute the relative frequency of atom-types per set and check if the atom-types having both higher RMSD in validation and higher SD in training set are more frequent in those sets for which poor performance was reported in virtual screening. From the results obtained in this work (see *Figure 4*, *Figure 11*), such is the case of N1, S1, N11, OS, N3 or O11. If parametrization errors are correlated to performance, the mentioned atom-types are suggested as candidates for a better characterization. Other candidates to improve characterization would be the ones with low presence in training set (the parametrization risks to be biased due to small sample size).

These atom-types might present a significant diversity among their electrostatic hydrophobicity/hydrophilicity values (see histogram of N11 in *Figure 13*). The atom-types whose distribution of LogP electrostatic is unimodal, i.e. distributed around one modal value (case A in *Figure 12*) are considered well characterized with a single hydrophobic partitioned parameter (this is the case of most of carbon atom-types, check histogram of C1 in *Figure 13*). Though, for some atom-types, the distribution of LogP electrostatic happens to present bimodality (case B in *Figure 12* and N11 in *Figure 13*). For each of those cases, this work rises the suggestion of creating two atom-type-subcategories that should be characterized depending on the chemical context of the sub-atom-types and their corresponding adjusted hydrophobic value. Accordingly, in case of a given atom-type showing bimodality, two parameters would be extra added in order to generate the decisive library of atom types with their partial contributions to $LogP_{o/w}$.

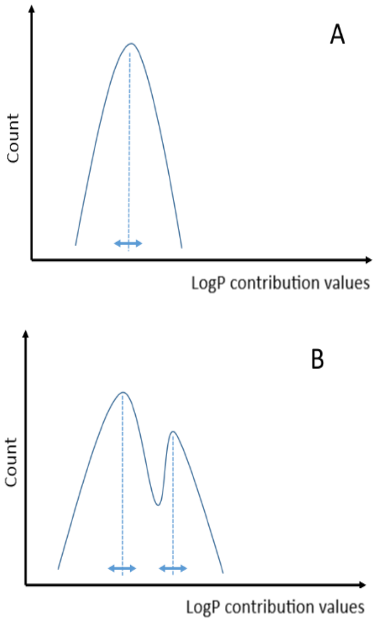


Figure 12. Unimodal and bimodal approximations for the distribution of hydrophobic contributions within atom-types.

Additionally, as a suggestion regarding the case of H2 and H3 biased contributions computed by MOPAC (see RMSE in *Figure 11)*, this work raises for MOPAC the possibility of checking for this kind of errors and in case the formerly elected conformer meets this event, choose an alternative conformer for further computation of hydrophobicity.

Finally, it would be interesting to compute the RMSD of the alignments of molecules using the parametrized descriptors for hydrophobicity defined in this work, together with statistics per atom-type. This study would allow to check if the alignment itself (not only in terms of similarity for virtual screening pipeline) improves or worsens generally and specifically per atom-type to detect more cases of improvable parametrization.

7. APPENDIX

A. Parametrization histograms

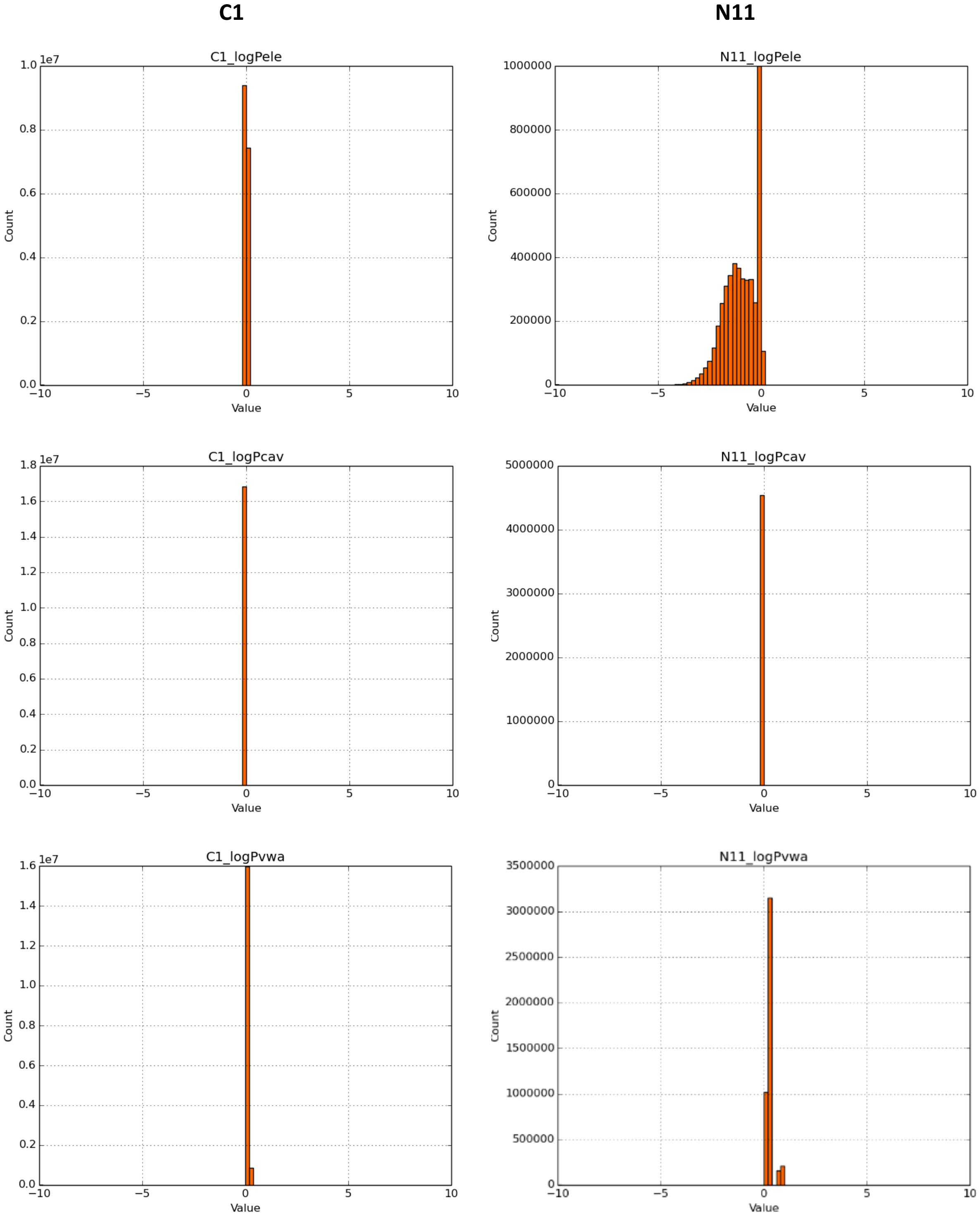


Figure 13. C1 and N11 examples of the histograms computed to study the distribution of LogP contributions for each atom-type.

B. Additional performance metrics

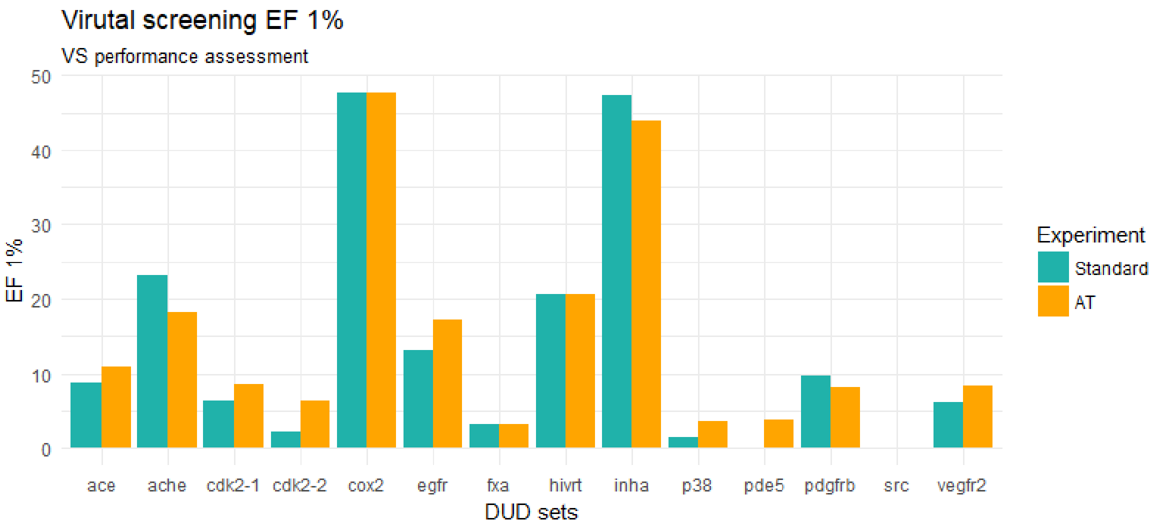


Figure 14. Comparison of EF 1% on Standard and AT virtual screening experiments on validation set.

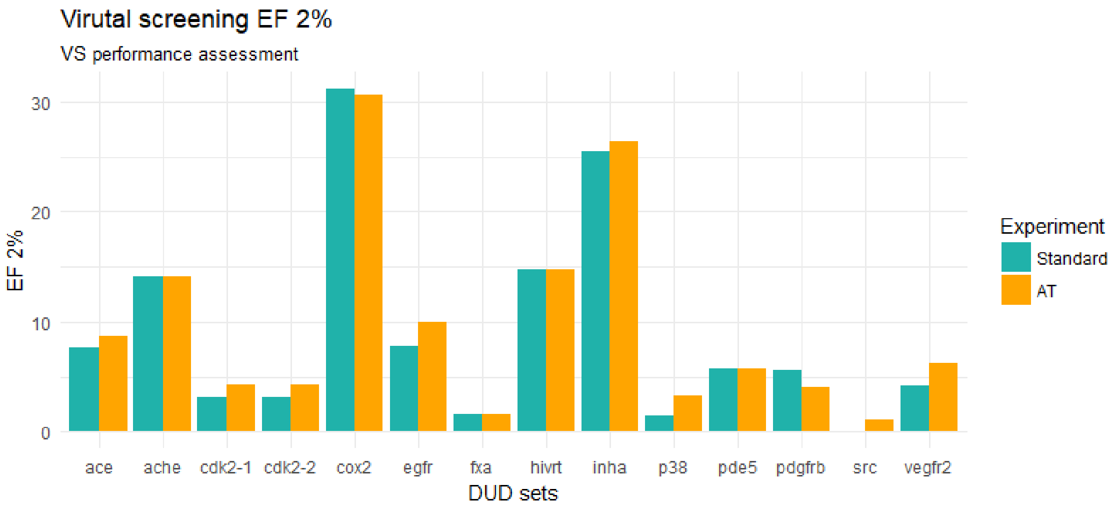


Figure 15. Comparison of EF 2% on Standard and AT virtual screening experiments on validation set.

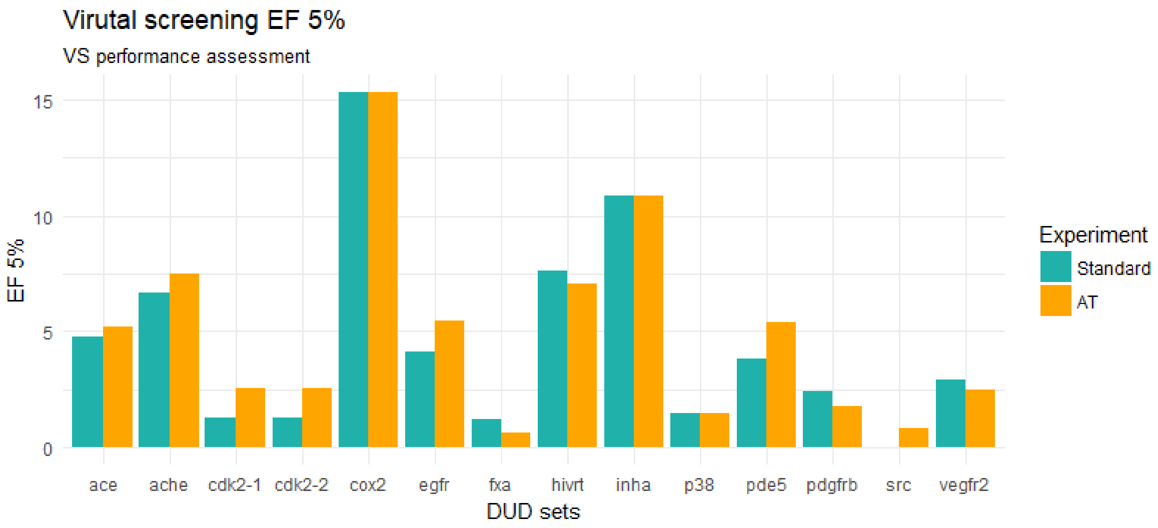


Figure 16. Comparison of EF 5% on Standard and AT virtual screening experiments on validation set.

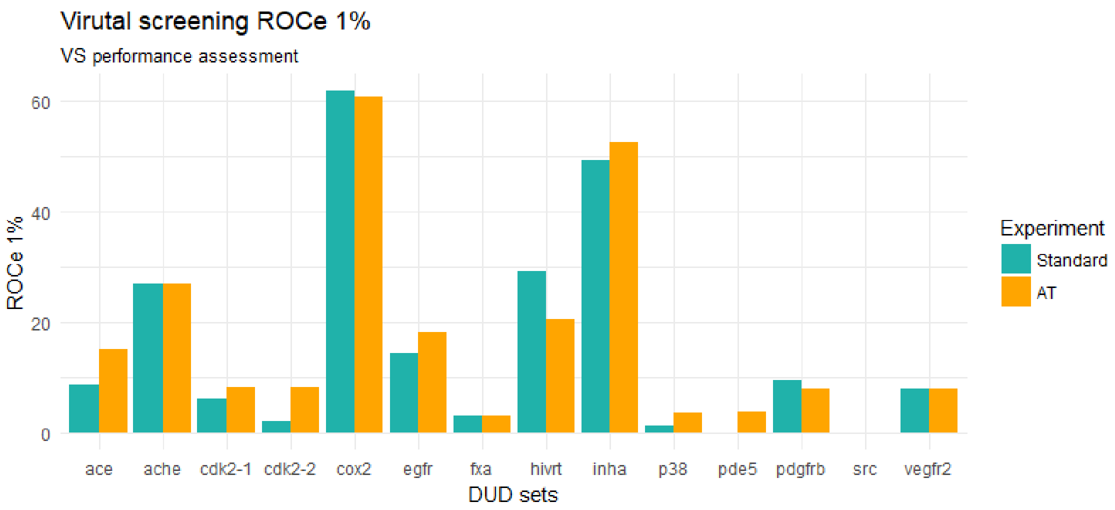


Figure 17. Comparison of ROCe 1% on Standard and AT virtual screening experiments on validation set.

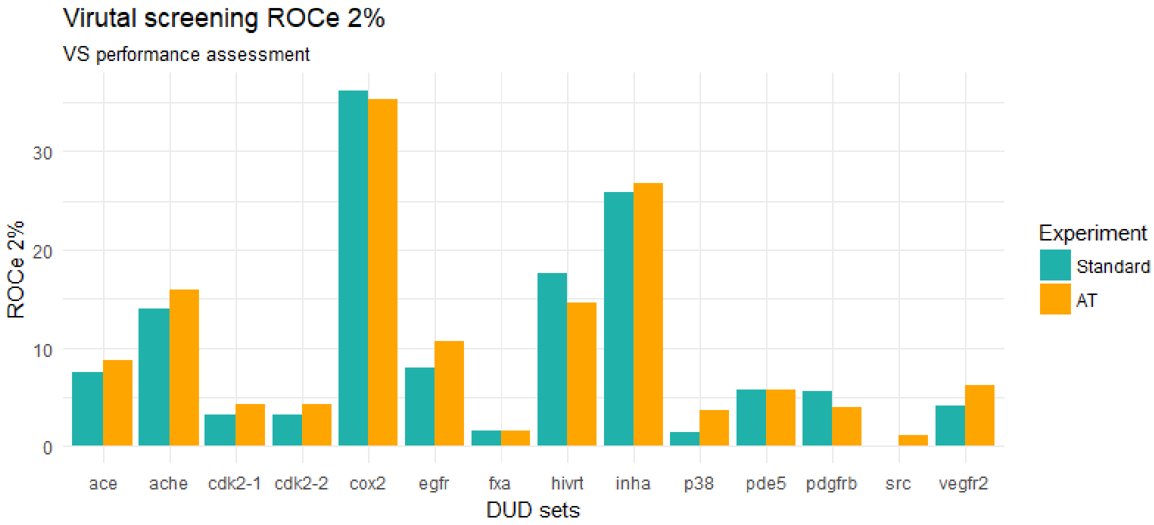


Figure 18. Comparison of ROCe 2% on Standard and AT virtual screening experiments on validation set.

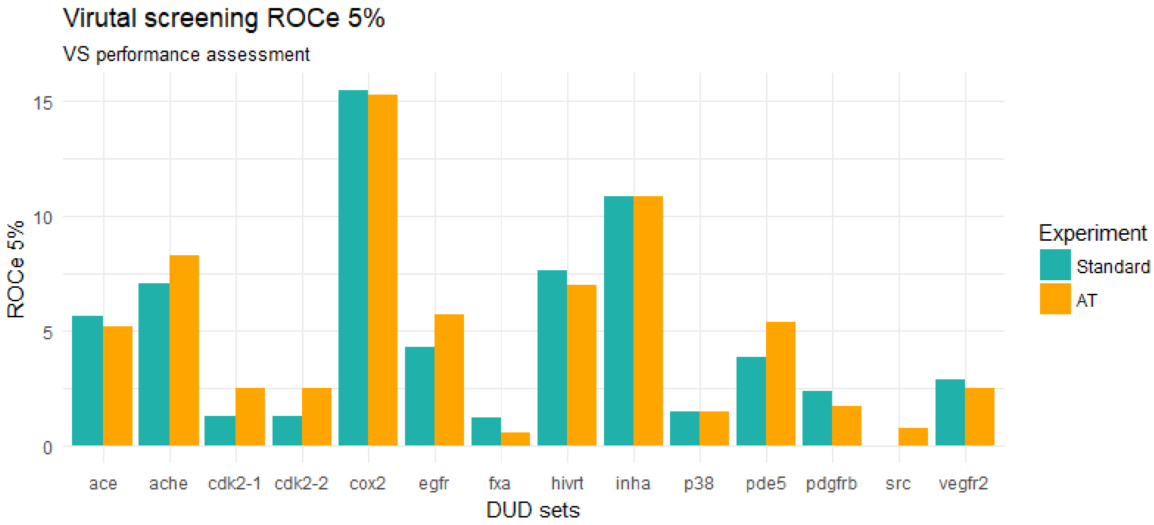


Figure 19. Comparison of ROCe 5% on Standard and AT virtual screening experiments on validation set.

C. Execution time for hydrophobicity parameters

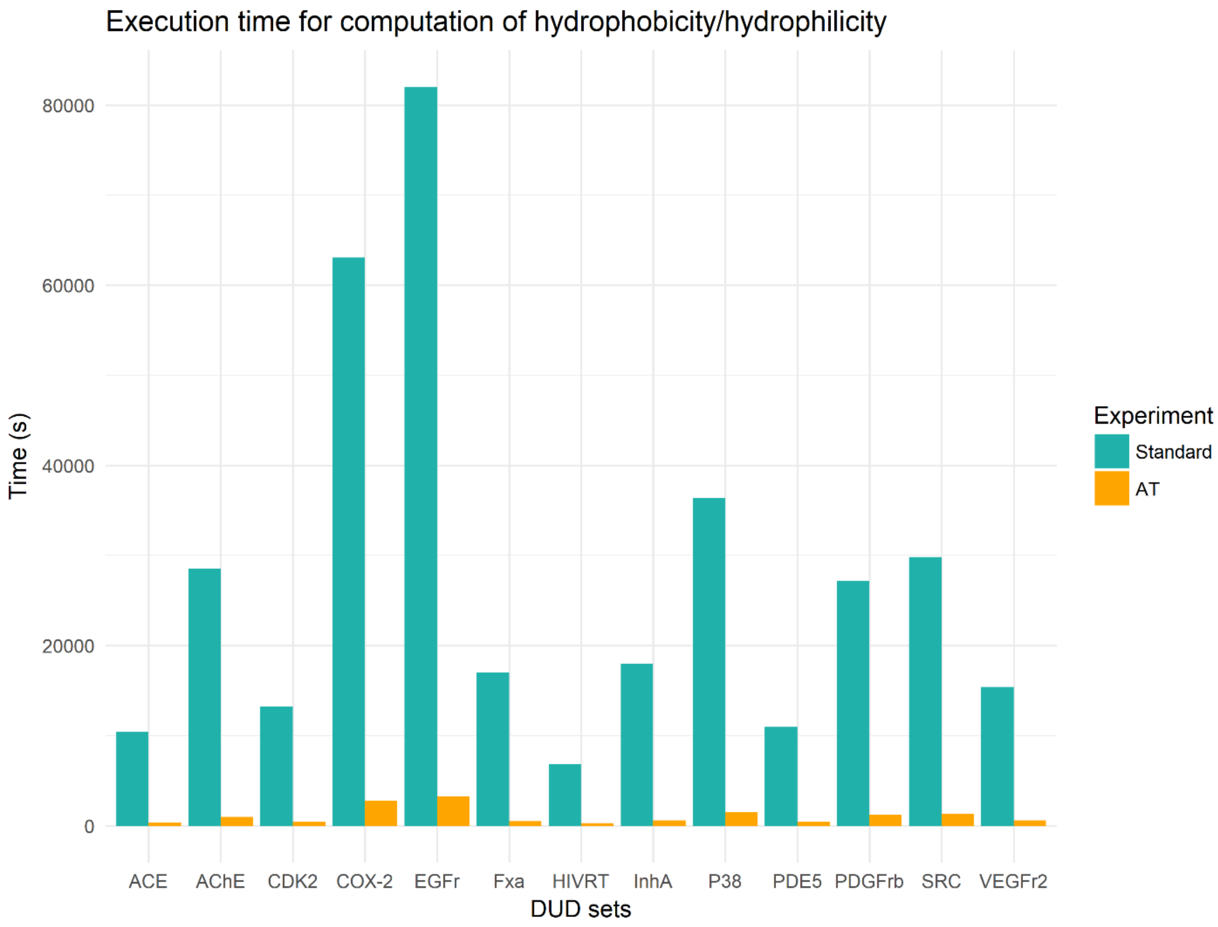
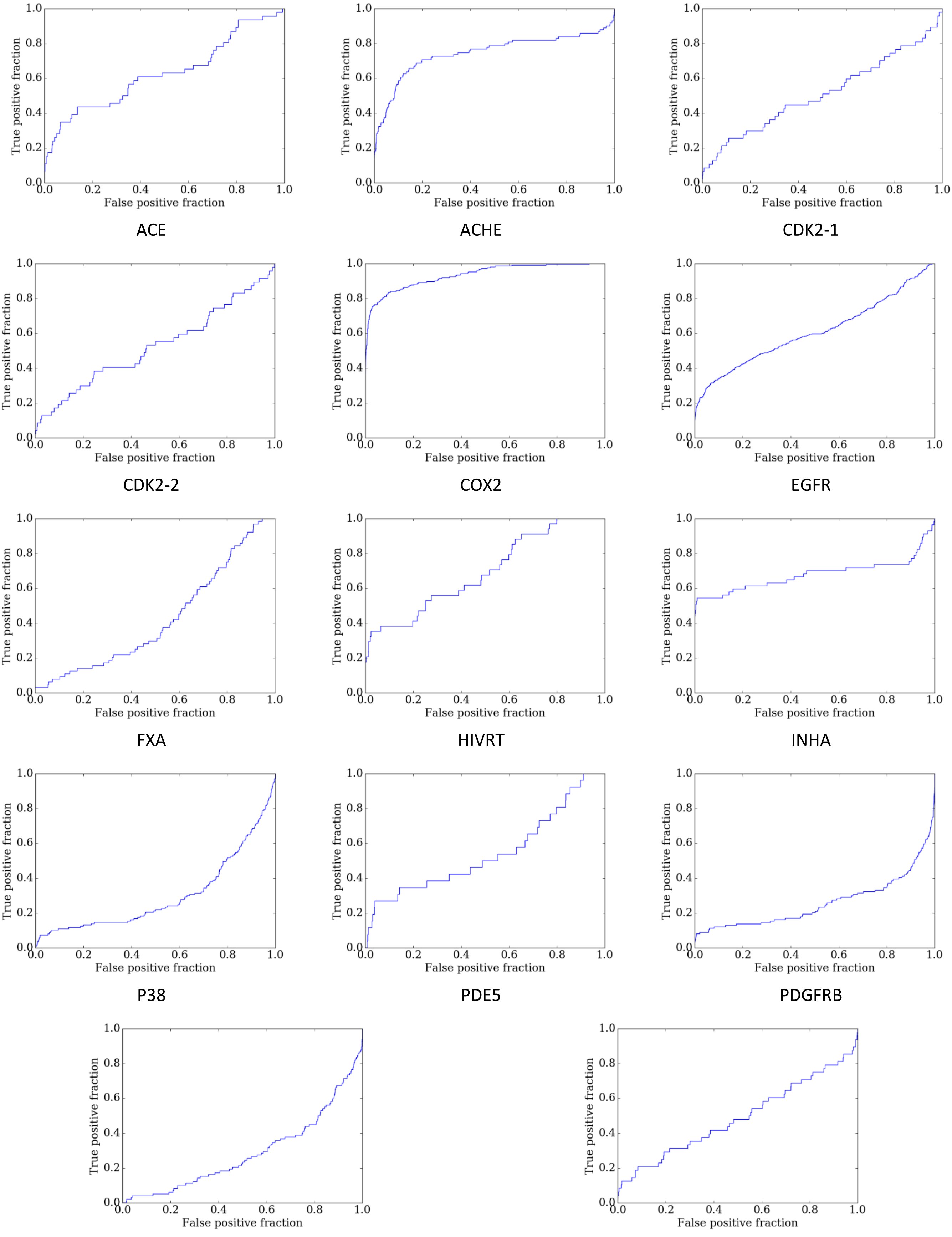


Figure 20. Differential execution time in Standard Pharmscreen® computation of hydrophobicity parameters and AT assignation of parameters.

D. ROC curves for AT virtual screening



ACE

ACHE

CDK2-1

CDK2-2

COX2

EGFR

FXA

HIVRT

INHA

P38

PDE5

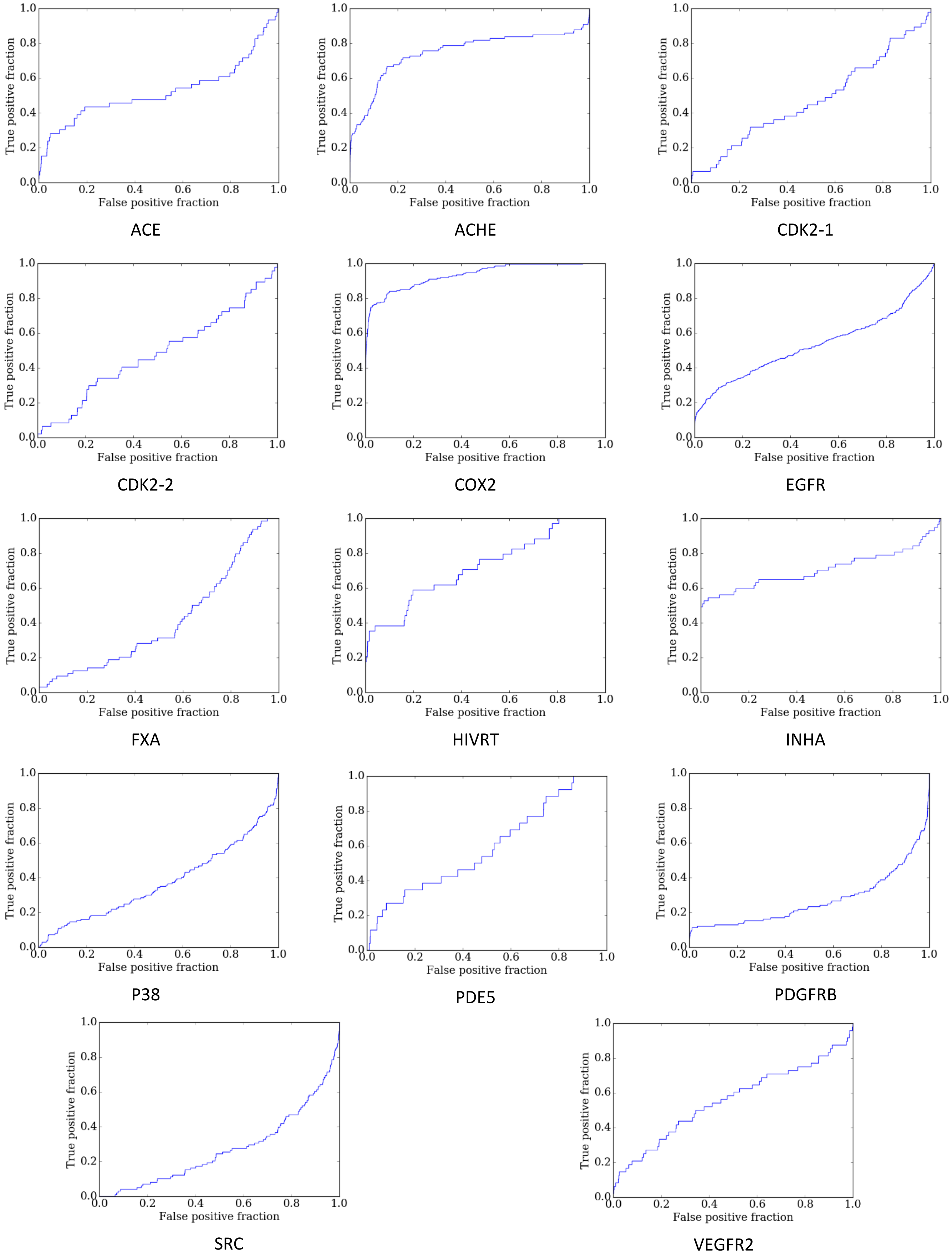
PDGFRB

SRC

VEGFR2

Figure 21. Multi-plot figure of ROC curves from AT virtual screening of validation subsets (DUD sets).

E. ROC curves for Standard Pharmscreen® virtual screening



ACE

ACHE

CDK2-1

CDK2-2

COX2

EGFR

FXA

HIVRT

INHA

P38

PDE5

PDGFRB

SRC

VEGFR2

Figure 22.Multi-plot figure of ROC curves from Standard Pharmscreen® virtual screening of validation subsets (DUD sets).

F. Pipeline of scripts and files processing in parametrization workflow

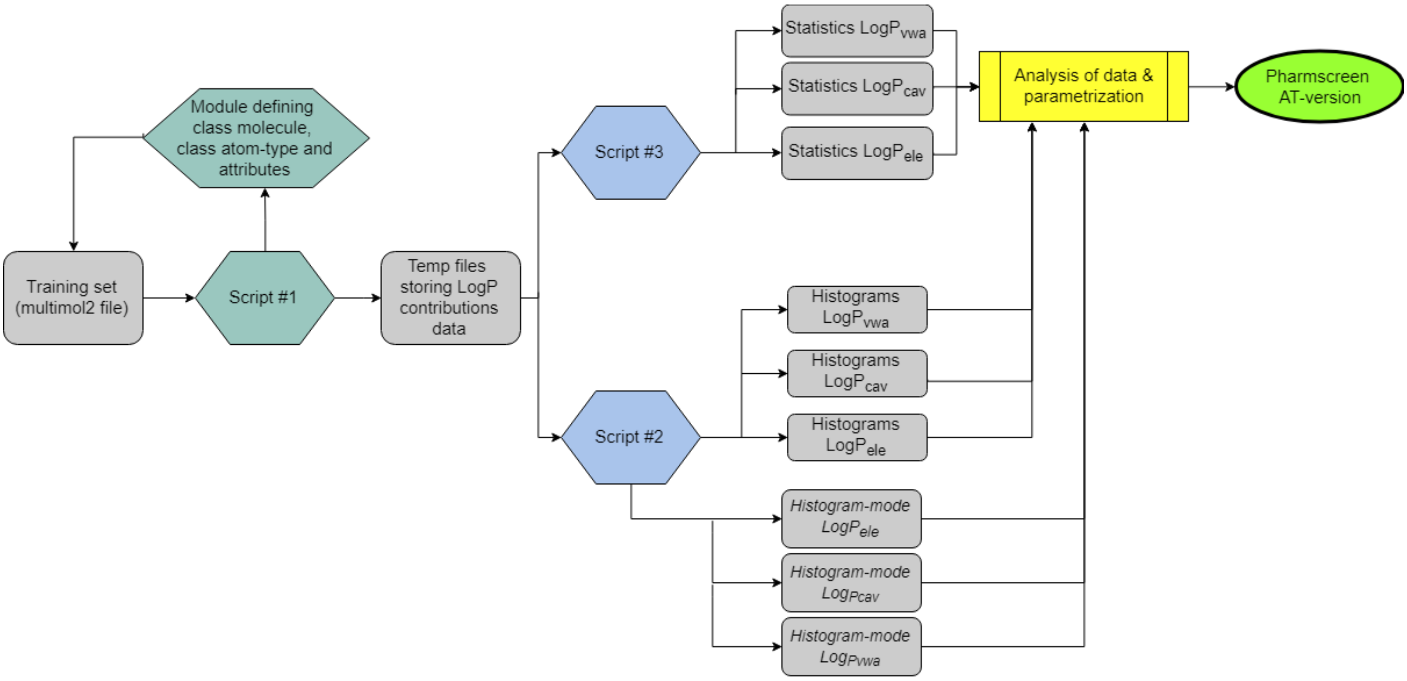


Figure 23. Diagram of the pipeline of scripts and files processing designed for parametrization of LogP contributions per atom-type.

The workflow of parametrization included a combination of several python scripts calls that processed temporary files storing hydrophobicity/hydrophilicity data of the atom-types identified in training set. The outcome of this routine included histograms plots, *histogram-mode* parameter and other statistics per atom-type for LogPele, LogPcav and LogPvwa. This information was analysed, and the parametrized values were implemented in "AT" version of Pharmscreen® software. Scripts #1, #2 and #3 are intellectual property-free and can be found at github repository: https://github.com/elisagdelope/TFM_scripts-IP-free.git.

8. BIBLIOGRAPHY

1. Krogsgaard-Larsen, P., Strømgaard, K. & Madsen, U. *Textbook of drug design and discovery*. (2010).

2. Silverman, R. B. *Chapter 7 - Drug Metabolism BT - The Organic Chemistry of Drug Design and Drug Action (Second Edition)*. (2004). doi:http://dx.doi.org/10.1016/B978-0-08-051337-9.50012-2

3. Macalino, S. J. Y., Gosu, V., Hong, S. & Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch. Pharm. Res.* **38,** 1686–1701 (2015).

4. Takenaka, T. Classical vs reverse pharmacology in drug discovery. *BJU Int.* **88,** 7–10 (2008).

5. Maehle, A.-H., Prüll, C.-R. & Halliwell, R. F. The emergence of the drug receptor theory. *Nat. Rev. Drug Discov.* **1,** 637–641 (2002).

6. Koshland, D. E. The Key–Lock Theory and the Induced Fit Theory. *Angew. Chemie Int. Ed. English* **33,** 2375–2378 (1995).

7. Kovermann, M., Grundström, C., Sauer-Eriksson, A. E., Sauer, U. H. & Wolf-Watz, M. Structural basis for ligand binding to an enzyme by a conformational selection pathway. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 6298–6303 (2017).

8. Drews, J. Drug discovery: a historical perspective. *Science* **287,** 1960–4 (2000).

9. Rask-Andersen, M., Almén, M. S. & Schiöth, H. B. Trends in the exploitation of novel drug targets. *Nat. Rev. Drug Discov.* **10,** 579–590 (2011).

10. Inglese, J. & Auld, D. S. High Throughput Screening (HTS) Techniques: Applications in Chemical Biology. in *Wiley Encyclopedia of Chemical Biology* 1–15 (John Wiley & Sons, Inc., 2008). doi:10.1002/9780470048672.wecb223

11. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46,** 3–26 (2001).

12. Orozco, M. *et al.* Theoretical representation of solvent effects in the study of biochemical systems. *J. Mol. Struct. THEOCHEM* **371,** 269–278 (1996).

13. Carbó, R. & Chakraborty, T. *Theoretical and quantum chemistry at the dawn of the 21st century*. (Apple Academic Press, Inc, 2016).

14. Giesen, D. J., Hawkins, G. D., Liotard, D. A., Cramer, C. J. & Truhlar, D. G. A universal model for the quantum mechanical calculation of free energies of solvation in non-aqueous solvents. *Theor. Chem. Accounts Theory, Comput. Model. (Theoretica Chim. Acta)* **98,** 85–109 (1997).

15. Lewis, R. A. Chapter 4. The Development of Molecular Modelling Programs: The Use and Limitations of Physical Models. in 88–107 (2011). doi:10.1039/9781849733410-00088

16. Bajorath†, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. (2001). doi:10.1021/CI0001482

17. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3,** 711–716 (2004).

18. Adams, C. P. & Brantner, V. V. Estimating The Cost Of New Drug Development: Is It Really $802 Million? *Health Aff.* **25,** 420–428 (2006).

19. Krusemark, C. J. *Drug Design: Structure- and Ligand-Based Approaches*. *The Quarterly Review of Biology* **87,** (2012).

20. Song, C. M., Lim, S. J. & Tong, J. C. Recent advances in computer-aided drug design. *Brief. Bioinform.* **10,** 579–591 (2009).

21. DiMasi, J. A., Hansen, R. W. & Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **22,** 151–185 (2003).

22. Paul, S. M. *et al.* How to improve R&amp;D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9,** 203–214 (2010).

23. Hughes, J., Rees, S., Kalindjian, S. & Philpott, K. Principles of early drug discovery. *Br. J. Pharmacol.* **162,** 1239–1249 (2011).

24. Yang, Y., Adelstein, S. J. & Kassis, A. I. Target discovery from data mining approaches. *Drug Discov. Today* **14,** 147–154 (2009).

25. Li, A. P. Screening for human ADME/Tox drug properties in drug discovery. *Drug Discov. Today* **6,** 357–366 (2001).

26. Lombardo, F., Gifford, E. & Shalaeva, M. Y. In silico ADME prediction: data, models, facts and myths. *Mini Rev. Med. Chem.* **3,** 861–75 (2003).

27. Pliska, V., Testa, B. & Van De Waterbeemd, H. Lipophilicity in Drug Action and Toxicology. (1996).

28. Kubinyi, H. Lipophilicity and drug activity. in *Progress in Drug Research / Fortschritte der Arzneimittelforschung / Progrès des recherches pharmaceutiques* 97–198 (Birkhäuser Basel, 1979). doi:10.1007/978-3-0348-7105-1_5

29. Gore, M. & Desai, N. S. Computer-Aided Drug Designing. in *Methods in molecular biology (Clifton, N.J.)* **1168,** 313–321 (2014).

30. Huang, H.-J. *et al.* Current developments of computer-aided drug design. *J. Taiwan Inst. Chem. Eng.* **41,** 623–635 (2010).

31. Martí-Renom, M. A. *et al.* Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).

32. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **234,** 779–815 (1993).

33. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42,** W252–W258 (2014).

34. Jahn, A., Hinselmann, G., Fechner, N. & Zell, A. Optimal assignment methods for ligand-based virtual screening. *J. Cheminform.* **1,** 14 (2009).

35. Cramer, R. D., Patterson, D. E. & Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110,** 5959–5967 (1988).

36. Klebe, G., Abraham, U. & Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **37,** 4130–46 (1994).

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

38. Good, A. C. & Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput. Aided. Mol. Des.* **22,** 169–178 (2008).

39. Kirchmair, J. *et al.* How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **49,** 678–692 (2009).

40. Ewing, T. J., Makino, S., Skillman, A. G. & Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided. Mol. Des.* **15,** 411–28 (2001).

41. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30,** 2785–2791 (2009).

42. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31,** 455–61 (2010).

43. Grosdidier, A., Zoete, V. & Michielin, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* **39,** W270–W277 (2011).

44. Irwin, J. J. & Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45,** 177–82

45. Wang, Y. *et al.* PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **37,** W623-33 (2009).

46. Kollman, P. A. *et al.* Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **33,** 889–97 (2000).

47. Paul C. D. Hawkins, *, A. Geoffrey Skillman, and & Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. (2006). doi:10.1021/JM0603365

48. von Korff, M., Freyss, J. & Sander, T. Flexophore, a New Versatile 3D Pharmacophore Descriptor That Considers Molecular Flexibility. *J. Chem. Inf. Model.* **48,** 797–810 (2008).

49. Koes, D. R. & Camacho, C. J. Pharmer: Efficient and Exact Pharmacophore Search. *J. Chem. Inf. Model.* **51,** 1307–1314 (2011).

50. Ahlström, M. M., Ridderström, M., Luthman, K. & Zamora, I. Virtual Screening and Scaffold Hopping Based on GRID Molecular Interaction Fields. *J. Chem. Inf. Model.* **45,** 1313–1323 (2005).

51. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23,** 3–25 (1997).

52. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* **1,** 337–341 (2004).

53. Luque, F. J., Barril, X. & Orozco, M. Fractional description of free energies of solvation. *J. Comput. Aided. Mol. Des.* **13,** 139–52 (1999).

54. Mannhold, R., Poda, G. I., Ostermann, C. & Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of LogP Methods on more than 96,000 Compounds. *J. Pharm. Sci.* **98,** 861–893 (2009).

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

56. Leo, A., Jow, P. Y. C., Silipo, C. & Hansch, C. Calculation of hydrophobic constant (log P) from .pi. and f constants. *J. Med. Chem.* **18,** 865–868 (1975).

57. Jäger, R., Kast, S. M. & Brickmann, J. Parametrization Strategy for the MolFESD Concept: Quantitative Surface Representation of Local Hydrophobicity. *J. Chem. Inf. Comput. Sci.* **43,** 237–247 (2003).

58. Kantola, A., Villar, H. O. & Loew, G. H. Atom based parametrization for a conformationally dependent hydrophobic index. *J. Comput. Chem.* **12,** 681–689 (1991).

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

60. Maldonado, A. G., Doucet, J. P., Petitjean, M. & Fan, B.-T. Molecular similarity and diversity in chemoinformatics: From theory to applications. *Mol. Divers.* **10,** 39–79 (2006).

61. Nikolova, N. & Jaworska, J. Approaches to Measure Chemical Similarity– a Review. *QSAR Comb. Sci.* **22,** 1006–1026 (2003).

62. Todeschini, R., Consonni, V. & Wiley InterScience (Online service). *Handbook of molecular descriptors*. (Wiley-VCH, 2000).

63. Willett, P. & Winterman, V. A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity Measures of Inter-Molecular Structural Similarity. *Quant. Struct. Relationships* **5,** 18–25 (1986).

64. Holliday, J. D., Hu, C.-Y. & Willett, P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screen.* **5,** 155–66 (2002).

65. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7,** 20 (2015).

66. S.L, P. Phamscreen. (2018).

67. Yvonne C. Martin, *,†, James L. Kofron, ‡ and & Traphagen‡, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? (2002). doi:10.1021/JM020155C

68. Doucet, J. P. & Panaye, A. 3D Structural Information: From Property Prediction to Substructure Recognition with Neural Networks. *SAR QSAR Environ. Res.* **8,** 249–272 (1998).

69. Xue, L. & Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screen.* **3,** 363–72 (2000).

70. Oprea, T. I. & Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* **8,** 349–358 (2004).

71. Willett*, P., and, J. M. B. & Downs, G. M. Chemical Similarity Searching. (1998). doi:10.1021/CI9800211

72. Giganti, D. *et al.* Comparative Evaluation of 3D Virtual Ligand Screening Methods: Impact of the Molecular Alignment on Enrichment. *J. Chem. Inf. Model.* **50,** 992–1004 (2010).

73. Kwon, Y. *Handbook of essential pharmacokinetics, pharmacodynamics and drug metabolism for industrial scientists*. (Kluwer Academic Publishers, 2001).

74. Testa, B., Crivori, P., Reist, M. & Carrupt, P.-A. The influence of lipophilicity on the pharmacokinetic behavior of drugs: Concepts and examples. *Perspect. Drug Discov. Des.* **19,** 179–211 (2000).

75. Katrin Palm *et al.* Evaluation of Dynamic Polar Molecular Surface Area as Predictor of Drug Absorption: Comparison with Other Computational and Experimental Predictors. (1998). doi:10.1021/JM980313T

76. Carrupt, P.-A., Testa, B. & Gaillard, P. Computational Approaches to Lipophilicity: Methods and Applications. in 241–315 (2007). doi:10.1002/9780470125885.ch5

77. Heiden, W., Moeckel, G. & Brickmann, J. A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces. *J. Comput. Aided. Mol. Des.* **7,** 503–514 (1993).

78. Meylan, W. M. & Howard, P. H. Atom/fragment contribution method for estimating octanol-water partition coefficients. *J. Pharm. Sci.* **84,** 83–92 (1995).

79. Qishi Du, *,†,‡, Peng-Jun Liu, ‡ and & Mezey§, P. G. Theoretical Derivation of Heuristic Molecular Lipophilicity Potential: A Quantum Chemical Description for Molecular Solvation. (2005). doi:10.1021/CI049707L

80. Modesto Orozco*, † and & F. Javier Luque*, ‡. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. (2000). doi:10.1021/CR990052A

81. Biosym Technologies. DELPHI computer program. (1992).

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

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

84. Muñoz-Muriedas, J. *et al.* Hydrophobic molecular similarity from MST fractional contributions to the octanol/water partition coefficient. *J. Comput. Aided. Mol. Des.* **19,** 401–419 (2005).

85. Miertuš, S. & Tomasi, J. Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes. *Chem. Phys.* **65,** 239–245 (1982).

86. Miertuš, S., Scrocco, E. & Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **55,** 117–129 (1981).

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

88. Kellogg, G. E., Semus, S. F. & Abraham, D. J. HINT: A new method of empirical

hydrophobic field calculation for CoMFA. *J. Comput. Aided. Mol. Des.* **5,** 545–552 (1991).

89. Leo, A. J. ClogP. *Daylight Chemical Information Systems: Irvine, CA* (1991). Available at: http://www.daylight.com/dayhtml/doc/clogp/. (Accessed: 2nd July 2018)

90. Andreas Klamt, *, Volker Jonas, †, Thorsten Bürger, and & Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. (1998). doi:10.1021/JP980017S

91. Tomasi, J., Mennucci, B. & Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **105,** 2999–3094 (2005).

92. Carrupt, P.-A.; Testa, B.; Gaillard, P. Computational Approaches to Lipophilicity: Methods and Applications. in *Reviews in Computational Chemistry, Vol. 11* (ed. Lipkowitz, K. B., Boyd, D. B.) 576 (Wiley: New York, 1997).

93. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **39,** 868–873 (1999).

94. Vrakas, D., Tsantili-Kakoulidou, A. & Hadjipavlou-Litina, D. Exploring the consistency of logP estimation for substituted coumarins. *QSAR Comb. Sci.* **22,** 622–629 (2003).

95. Sangster, J. *Octanol-water partition coefficients : fundamentals and physical chemistry*. (Wiley, 1997).

96. Pierotti, R. A. A scaled particle theory of aqueous and nonaqueous solutions. *Chem. Rev.* **76,** 717–726 (1976).

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

98. Ginex, T. *et al.* Application of the quantum mechanical IEF/PCM-MST hydrophobic descriptors to selectivity in ligand binding. *J. Mol. Model.* **22,** 136 (2016).

99. RDKit: Open-source cheminformatics software. (2006).

100. MOPAC6.0. Version locally modified by Luque, F.J. and Orozco, M. (Universidad de Barcelona).

101. Meylan, W. M. & Howard~, P. H. AtomlFragment Contribution Method for Estimating Octanol-Water Partition Coefficients. (1993).

102. Klopman, G., Li, J.-Y., Wang, S. & Dimayuga, M. Computer Automated log P Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Model.* **34,** 752–781 (1994).

103. Suzuki, T. & Kudo, Y. Automatic log P estimation based on combined additive modeling methods. *J. Comput. Aided. Mol. Des.* **4,** 155–198 (1990).

104. Chou, J. T. & Jurs, P. C. Computer-Assisted Computation of Partition Coefficients from Molecular Structures Using Fragment Constants. *J. Chem. Inf. Model.* **19,** 172–178 (1979).

105. P, B., G, M. & C, V. *European journal of medicinal chemistry. Eur J Med Chem* **19,** (Elsevier Masson SAS, 1984).

106. Ghose, A. K. & Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition

Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **7,** 565–577 (1986).

107. Ghose, A. K. & Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **27,** 21–35 (1987).

108. Ghose, A. K., Pritchett, A. & Crippen, G. M. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. *J. Comput. Chem.* **9,** 80–90 (1988).

109. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **4,** 268–276 (2018).

110. Walters, W. P. & Murcko, M. A. Prediction of 'drug-likeness'. *Adv. Drug Deliv. Rev.* **54,** 255–71 (2002).

111. SPECS. specs.net. (2017). Available at: www.specs.net.

112. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. & Stewart, J. J. P. 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 (1985).

113. Geppert, H., Vogt, M. & Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **50,** 205–216 (2010).

114. Jain, A. N. & Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput. Aided. Mol. Des.* **22,** 133–139 (2008).

115. Niu Huang, Brian K. Shoichet, * and & Irwin*, J. J. Benchmarking Sets for Molecular Docking. (2006). doi:10.1021/JM0608356

116. Vázquez, J., Deplano, A., Herrero, E. & Luque, F. J. Development and Validation of Molecular Overlays Derived From 3D Hydrophobic Similarity with PharmScreen. in *255th ACS National Meeting & Exposition, American Chemistry Society* (2018).

117. Ginex, T. *et al.* Development and validation of hydrophobic molecular fields derived from the quantum mechanical IEF/PCM-MST solvation models in 3D-QSAR. *J. Comput. Chem.* **37,** 1147–1162 (2016).

118. Witten, I. H. (Ian H. ., Frank, E. & Hall, M. A. (Mark A. *Data mining : practical machine learning tools and techniques*. (Morgan Kaufmann, 2011).

119. Nicolas Triballeau, *,†,‡, Francine Acher, †, Isabelle Brabet, §, Jean-Philippe Pin, § and & Bertrand‡, H.-O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. (2005). doi:10.1021/JM049092J

120. *Detector Performance Analysis Using ROC Curves - MATLAB &amp; Simulink*.