3D-Pharma, A Ligand-based Virtual Screening tool using 3D Pharmacophore Fingerprints

Bernardo F. Domingues, Andrelly Martins-José, and Júlio C. D. Lopes

Chemoinformatics Group NEQUIM, Departamento de Química, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

E-mail: jlopes.ufmg@gmail.com

Abstract

In this work, we introduced 3D-Pharma, a new Ligand-Based Virtual Screening method that uses fingerprints of pharmacophore triplets at atomic resolutions to build very simple and predictive models. Within 3D-Pharma the molecules are described by multiple representations that comprehend several prototropic species and conformations (multiple species, multiple mode approach). All the multiple representations of a compound are concatenated into a unique fingerprint that accounts for most of its chemical and conformational diversity. The biological activity of an ensemble of active molecules are represented by a single modal fingerprint or model, validated through a new exhaustive 10-fold cross-validation scheme, which improves robustness and internal consistency of the models, as well as its predictive power. We benchmark our method with 10 datasets of active compounds and decoys gathered from DUD database and compare its performance against seven state-of-the-art LBVS methods. To generate the models, we used three external and independent datasets of bioactive compounds (Drugs, PDB Ligands and WOMBAT). We concluded that 3D-Pharma overperforms all other state-of-the-art LBVS tools analyzed, in terms of global accuracy as well as scaffold hopping and early recovery capacities. Furthermore, the models produced by 3D-Pharma are simple, robust, consistent and predictive.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Introduction

Since 1996, the pharmaceutical industry has experienced a long period of low profits whilst at same time experiencing continuous increases in expenditures. Productivity, as measured by the number of approved NME's, has fallen to its lowest level since the 80's.[1] As a consequence, there are an impressive number of blockbuster drugs whose patents will be coming to their expiry dates (a phenomena known as "patent cliff").[2–4] Therefore, the major players in the pharmaceutical industry are under pressure to abandon less profitable branches and to concentrate their efforts on those whose financial return is more likely.[5] Within this scenario, *in silico* methods have become even more important in order to speed up the process of discovery at lower aggregated cost.[6]

A major problem in drug discovery is the identification of novel compounds that show binding properties to protein targets of pharmaceutical interest, together with appropriate pharmacokinetic properties. In the lack of previous information regarding target´s structure nor known active compounds, there is no choice but to rely on a brute force hit identification process, such as HTS (High Throughput Screening), over a whole chemical compound library.[7,8] But as data about known ligands or crystallographic structures from protein targets becomes available, such information can be used for selecting a small portion of some virtual compound database that shows a higher likelihood of interaction with a given target than a randomly chosen compound.[9] Such enrichment is the goal of Virtual Screening (VS) applications.[10] VS methods can be broadly classified by the origin of the data used in the screening process. If it uses only data derived from previously known ligands, the application is known as Ligand-Based VS; if the application uses data derived from the structure of the protein target, the method is known as Target-Based VS. There is a plethora of methods available within each class, and a handful of hybrid methods,[11–13] most of them summarized in several recent reviews.[10,14–16]

The methods of discovery and development of new drugs based on knowledge of the structures of biological targets are unavoidably dependent on their validation as potential therapeutic targets, as well as the specific mode of action through which the new candidate drug will exert its putative/predicted effect. It is worth mentioning that approximately 50% of

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

the compounds rejected at Clinical Phase II is due to insufficient efficacy, [17] suggesting deficiencies in target or endpoint biomarkers validations. [18] A TBVS study is typically carried out through molecular docking approaches. However, although they are able to reproduce with reasonable precision the conformations observed in the crystallographic complexes, the scoring functions cannot consistently differentiate active from inactive compounds. [19] Moreover, the methods of drug discovery and development based on a knowledge of the structures of the active ligands, e.g. those substances previously known to cause a biological, pharmacological or therapeutic effect of interest, has a great potential to generate predictive models aiming not only at a specific target, but also for biological effects which can be associated with in vitro or in vivo biological tests. [20–22] Conversely, the ligand-based tools do not depend on information regarding a biological target nor a mechanism of action, hence enabling them to create predictions about effects that should be multifactorial. [23–25] The methods based on the structures of the active ligands, although not necessarily dependent on a specific mode of action, lack a sufficient knowledge of the molecular structure and its role in the interactions with the components of the biological environment. [26] Hence, there is no a consistent set of molecular descriptors capable of foreseeing the biological properties of small molecules with reasonable confidence.

The most popular ligand-based methods are those derived from 2D molecular structures, possibly due to their availability, speed, ease of use and relatively well established protocols.[27] In general, they present a good performance in retrospective studies, [28] but the results of prospective studies generally are disappointing, even though they lead to a relative enrichment of actives.[29] The inclusion of conformational data in 3D LBVS methods increases the complexity of the algorithms and computational costs. Therefore, most of the 3D methods select only one conformation for each compound. The options are the lowest energy conformation, a conformation from a co-crystallized complex, or a conformation based on alignment with other active compounds. However, the binding event could dramatically change the energy of conformers, and the active conformation could not be same of the global minimum calculated in water or in a vacuum, as usually done by most quantum mechanics (QM) or molecular mechanics (MM) tools used in chemoinformatics and medicinal chemistry. In fact, recently

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

Nicklaus and co-workers[30] analyzed the conformations of ligands found in the PDB database at Density Functional level of theory. The difference of energy between the conformations found in PDB structures and the fully optimized conformer remain in the range of 0 to 25 kcal/mol, distributed quite evenly and independently of the crystallographic resolution. The ligands deposited in the Protein Data Bank[31] mostly present only one conformation bound to the active site of the biological target. However, there is an increasing amount of data showing that it is not uncommon for a ligand to exhibit multiple conformations or multiple binding modes.[32,33] In addiction, the fact that crystallographic structures are not obtained under physiological conditions casts doubts over which conformation should be considered as the bioactive.[34] Even the soaking method used to produce the crystal can influence the conformation observed in models obtained from the X-ray diffraction maps.[35]

Some recent studies suggest that 2D methods outperform 3D approaches in terms of accuracy.[28] This could be due to deficiencies in the quality of the 3D descriptors or problems with the conformational sampling. However, some 3D methods based on shape, electrostatics, or pharmacophore features have been shown to perform better than 2D LBVS methods when considering scaffold hopping.[36] There are numerous pharmacophore-based VS approaches. The classical use of pharmacophore elucidation through the alignment of active compounds is the approach taken by the majority of commercially available tools like CATALYST,[37] GALAHAD,[38] GASP,[39] the pharmacophore module of MOE,[40] PHASE[41] and many others non-commercial tools, like PharmaGist.[42,43] Although the traditional pharmacophore mapping approach is well established and performs satisfactorily, it is very sensitive to the size of the dataset to be screened[14] and biased by the conformation selection.[44,45] Pharmacophore keys (or fingerprints) encode the pharmacophoric features in binary vectors. They lack the intuitive nature of the classical pharmacophore elucidation, but they have a considerably higher throughput when dealing with large molecular databases, making them a very popular approach to VS. There are several implementations of pharmacophore searches powered by fingerprints, such as FLAP[46–49] and Pharmer.[50] Other examples found are commercially available suites like Tripos' Tuplets[51] and Accelrys' 3DKeys.[52] Despite the number of available techniques, any comparison between them is

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

difficult due to the lack of benchmarking standards both in databases and the metrics of evaluation.[15]

Currently there is a search for new molecular descriptors that associated with strong information modeling techniques and robust statistical methods could make better predictions than those produced by the traditional methods.[28] Furthermore, considering that chemical and biological systems are dynamic in their very nature, both ligands and biological targets are adaptive flexible molecules and the emergent properties that arise from the binding event can trigger cascade effects that ultimately produce the observed biological effects.[53] A knowledge of the structures of binding partners is not sufficient to clarify such a chain of events and to develop suitable tools for practical application.[26] Thus, we are currently at a crossroads in which the available methods are not up to the challenge of predicting accurately the biological activities of small organic molecules.[54]

Molecular recognition events are strongly dependent on conformational and proton-exchange equilibria but they are often neglected in virtual screening studies.[55,56] Different microspecies of biological macromolecules and their ligands are characterized by different conformations and stereo-electronic characteristics. Therefore, including multiple species and multiple conformational data (MultiSpecies-MultiMode approach, or MS-MM)[57,58] in VS applications could result in a comprehensive approach of biological activity and of inherent dynamics of the process of binding of small molecules to their biological targets. Considering those LBVS methods that use DUD datasets as benchmark, there are a few that have as a feature a full multimode approach.[59–61] Although it is more common to start with multiple conformations, the final results are computed over a single conformation.[46–49,62,63] To the best of our knowledge there is no published LBVS method that uses the MS-MM approach with DUD datasets.

This work aims to introduce 3D-Pharma, a new method for LBVS based on pharmacophore fingerprints. The predictive power of 3D-Pharma is compared with other state-of-the-art LBVS methods available elsewhere that use the Directory of Useful Decoys (DUD) database[64] as a benchmarking data set. The 3D-Pharma method applies a multispecie-multimode (MS-MM) approach for the generation of atom-centered potential pharmacophore triplets. Within 3D-

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

Pharma, all the multiple representations of a compound are concatenated into a unique binary vector or fingerprint that accounts for most of the chemical and conformational diversity of the compound. 3D-Pharma uses a single modal fingerprint [65] to represent the biological activity of an ensemble of active molecules, each one represented by its unique fingerprint. These models were built from three external and independent datasets (Drugs, PDB-Ligands and Wombat) and were validated by a new exhaustive 10-fold cross validation scheme where each external dataset was divided into three subsets comprising training, evaluation and test sets. The final models were selected after extensive internal validation against the evaluation and test sets.

Material and Methods

Data

A retrospective virtual screening study was made of ten protein targets chosen from among the 40 targets available in the Directory of Useful Decoys (DUD) dataset. [64] The DUD Dataset is a benchmarking dataset for docking tools, but is commonly used to assess performance of ligand-based methods and is well suited to do this. [28] The selected targets were: Aldose Reductase (ALR2), Andro- gen Receptor (AR), Cyclin-dependent Kinase 2 (CDK2), Cyclooxygenase-2 (COX-2), Epidermal Growth Factor Receptor Kinase (EGFR), Factor X-α (FX_α), Mitogen activated Protein Kinase 14 (P38), HIV-1 Reverse Transcriptase (HIVRT), Phosphodiesterase V (PDE5) and Peroxisome Proliferator Activated Receptor γ ($PPAR_\gamma$). These targets were selected based on availability of WOMBAT datasets[66,67] in the DUD website.

For each active molecule in the original release of DUD, there are approximately 36 other inactive molecules with similar topological features. Jahn et al [68] performed a lead-like filter [69] over the molecules, as suggested by Good and Oprea, [67] in order to make the benchmark set more suitable for LBVS applications. For each of the aforementioned targets, sets of actives and in-actives molecules were obtained from DUD according to Jahn’s filter. Aiming to have true and reliable external validation of 3D-Pharma approach, three independent datasets were used to build the models for each target. The first dataset, called "Drugs", consists of all available approved

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

and experimental drugs for each target, gathered from the public databases DrugBank,[70,71] KEGG Drugs[72] and Therapeutic Targets Database (TTD).[73] The second dataset, called "PDB-Ligands", contains the ligands bound to any crystallographic structure of the target deposited in the Protein Data Bank (PDB).[31] Finally, the third dataset, as mentioned above, was the WOMBAT dataset available through the DUD website. All datasets used to build models for each target (Drugs, PDB Ligands and WOMBAT) were previously filtered using a 2D comparison within ChemAxon's Instant JChem[74] against the corresponding DUD Actives subset. Any redundancies between them were excluded from the external datasets. A complete list of the molecules gathered for this study is available in the Supporting Information. Table 1 shows the datasets sizes as well as the number of different chemotypes found in DUD Actives datasets for each target. The numbers may vary from the original DUD release and from other publications since some of the molecules generated errors throughout the molecular treatment protocol and were not entered in the final subsets.

Treatment of Molecular Structures

All molecules used in this work were preprocessed following a pre-treatment protocol of manual dessalting, succeeded by standardization and dominant tautomer calculation with the Standardizer program by ChemAxon.[74] These steps ensure that all structures are in the same initial state. All datasets were submitted to the same protocol of molecular treatment, which starts by determining all dominant tautomers between pH 0 and 14, followed by a major microspecie calculation at pH 7 for each tautomer. These steps are important to be sure that a relevant sample of the chemical variability of the compound is taken into account when computing the potential pharmacophore

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

Table 1: Number of unique compounds for each target in the DUD and external datasets (Drugs, PDB-Ligands and WOMBAT), as well as the number of chemotypes for each DUD Active dataset.

Target	Dataset	Number of Compounds	Number of Chemotypes in DUD Actives
ALR2	DUD Actives DUD Decoys Drugs PDB Ligands WOMBAT	26 910 11 26 41	14
AR	DUD Actives DUD Decoys Drugs PDB Ligands WOMBAT	68 2616 45 13 36	10
CDK2	DUD Actives DUD Decoys Drugs PDB Ligands WOMBAT	47 1702 37 132 148	32
COX-2	DUD Actives DUD Decoys Drugs PDB Ligands WOMBAT	212 11577 77 4 66	44
EGFR	DUD Actives DUD Decoys Drugs PDB Ligands WOMBAT	365 14516 10 12 62	40
FX_{α}	DUD Actives DUD Decoys Drugs PDB Ligands WOMBAT	64 1888 5 85 105	19
HIVRT	DUD Actives DUD Decoys Drugs PDB Ligands WOMBAT	34 1370 4 29 97	17
P38	DUD Actives DUD Decoys Drugs PDB Ligands WOMBAT	135 5416 16 76 52	20
PDE5	DUD Actives DUD Decoys Drugs PDB Ligands WOMBAT	26 1561 12 10 85	22
$PPAR_{\gamma}$	DUD Actives DUD Decoys Drugs PDB Ligands WOMBAT	6 38 12 60 27	6

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

points (PPP) (as shown in Figure 1). To accomplish these tasks the ChemAxon's JChem[74] suite was used.

The next step is a conformational sampling along with partial charge calculations for each representation of the initial molecule. As the partial charge distribution affects the conformation and vice-versa, an iterative process would be the most accurate. But this approach is not viable when dealing with large datasets with thousands of molecules, due to its huge computational cost. Hence, a better approach was devised, trying to optimize CPU time without significant loss of precision.[75–77] Using OpenEye's QuacPac and Omega2[78] suites, the lowest-energy conformation is computed using the MMFF94s[79] force field, and its partial charges are determined using the semi-empirical AM1BCC[80] method. The last step is a conformational sampling, limited up to 200 conformers (for the maximum of 25 rotatable bonds), within an energy window of 5 − 10 kcal/mol and an RMSD between 0.5-1.0 Å.

Pharmacophore Fingerprint Build

After the molecular treatment, each compound is represented by a large ensemble of several conformations, associated to a small number of tautomers and protonation states. The next step is a pharmacophoric mapping, performed by ChemAxon's PMapper.[74] This process is done atom-wise and assigns at least one out of the following six PPP types to all heavy atoms in a molecule. The Aromatic (R) feature is assigned to any atom that is part of an aromatic ring. Hydrogen Bond Donor (D) and/or Hydrogen Bond Acceptor (A) are assigned to atoms able to establish hydrogen bonds with a potential target. Positively Charged (P) and Negatively Charged (N) are assigned to atoms with partial charges above +0.4 or below −0.4, respectively. Any other heavy atom that fails to fit in the classes above is assigned as Hydrophobic(H). All triplets formed by PPP's in 3D space are generated for each conformer. The Euclidean distance between each pair of points is discretized in ten distance bins (in Å): 0–3, 3–4.5, 4.5–6, 6–8, 8–10, 10–12.5, 12.5–15, 15–18, 18–21 and 21–∞. Each triplet is a putative pharmacophore formed by three heavy atoms with a PPP type assigned and a defined distance bin for each edge. This triplet is represented by a 6-character string (3 characters for the feature on vertices and 3 for the distance bins on the edges) that

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

identifies it unequivocally. Figure 1 illustrates how the pharmacophore triplet formed by a negatively charged carboxyl group, a positively charged amine group and the hydrogen donor nitrogen atom in the structure of histidine is represented within 3D-Pharma.

Each conformation of a single molecule could have hundreds of three-point potential pharmacophores, so 3D-Pharma transforms the strings in indexes of a binary fingerprint using a hash function provided by the CMPH [81] library. These numeric fingerprints are analogous to standard bi- nary fingerprints, but instead of storing a very sparse array of bits, only the indexes of lit bits are stored. This decision was made due to the non-scalability of the binary vector size when increasing the number of nodes used in each tuple (for example, when using tetrahedrons instead of triangles). Therefore, new operations are needed to substitute the binary operators AND and OR, since they are not applicable to a non-binary representation. Using Set Theory, the analogs to the two binary operations can be redefined: using Intersection (∩) for AND and Union (∪) for OR. The Tanimoto coefficient was used for similarity computation between two vectors using Set Theory operations. Given the fingerprints of two molecules A and B, the Tanimoto coefficient is given by:

$$T = \frac{|A \cap B|}{|A \cup B|}$$

.

Model Construction

In some LBVS applications, it is necessary to generate comprehensive queries that represent a pursued activity profile. This query should hold enough information to search and retrieve molecules with a potential activity from a large compound database. In 3D-Pharma, the query is a model built from a set of molecules previously known to be actives. A model (M) is formally defined by a set of pharmacophore triplets (x) that are present in the molecules that form the training set (T) in a frequency above a given threshold (τ). It can be formally defined as:

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

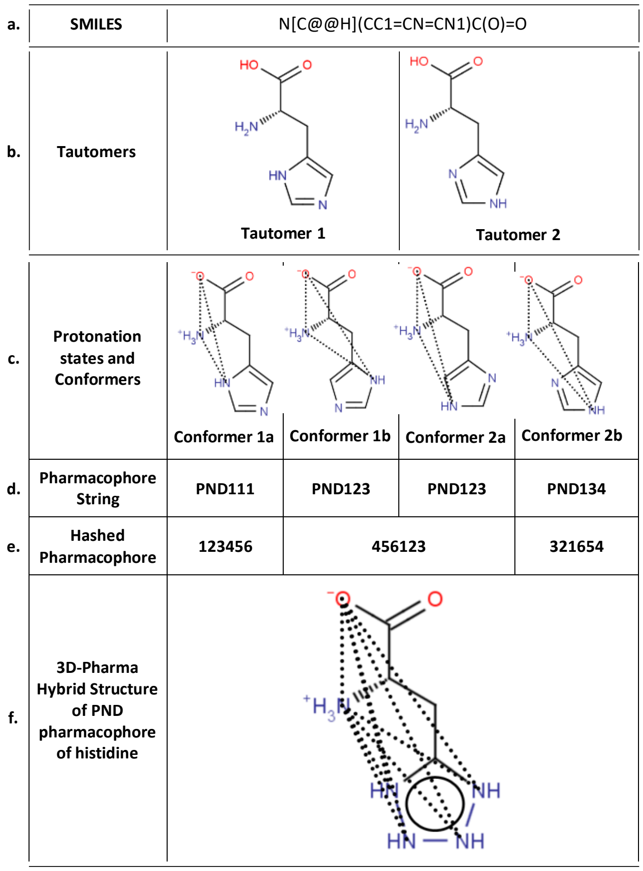

Figure 1: The effects of the proposed molecular treatment and generation of the 3D-Pharma pharmacophore triplets fingerprint. The amino acid histidine wase selected as an example and all structures are depicted in 2D for clarity. a) The SMILES representation of the neutral form of histidine. b) The structures of the two dominant tautomers of histidine (pH between 0 and 14), showing the hydrogen exchange between the two atoms of nitrogen in the imidazole ring. c) The structures of the major microspecies (protomers) of each tautomer of histidine at pH 7. For each protomer two hypothetical conformations are presented (only the imidazole ring flip are considered). The dotted lines represent the pharmacophore triplet formed by one of the negative charged (N) oxygen atom of the carboxyl group, the positive charged (P) alpha-nitrogen atom, and the hydrogen-bond donor (D) nitrogen atom in the imidazole ring. The same pharmacophore triplet is monitored over all conformations d) the pharmacophore triplets of each conformation are converted to a string. PND stands for a triplet formed by a positive, a negative and hydrogen-bond donor pharmacophores. The numbers after the alphabetic string are indicatives of the distance between the atoms (see text). e) The hashed pharmacophore form derived from alphanumeric string. f) The hypothetical hybrid representation of the PND pharmacophore of histidine encoded by 3D-Pharma fingerprint. All pharmacophore triplets detected over all conformations, represented by the dotted lines, are equally considered.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

$$x \in M \Leftrightarrow \frac{\sum_{i=1}^{m} f(x, T_i)}{m} \geq \tau,\ f(x, T_i) = \begin{cases} 1 & \text{if } x \in T_i \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

m being the size of the training set T, and T_i is the i^{th} molecule in the set. We tried to optimize the performance by incrementing the value of τ by 0.1 ($0 \leq \tau \leq 1$). The gathered data (not shown here) suggested that a value of 0.7 is a generally optimal cutoff when treating single molecular target datasets.

In order to perform a virtual screening study, it is necessary to have at least two datasets of active molecules: a training set and a test set. The former is used for model construction, which should be able to retrieve the latter among a set of inactive molecules. Although this is a widely used approach, it presents a major problem: how to split the active data between these two groups? The query construction is strongly affected by the training set selection. Besides this, test and training groups should not be too similar to turn the classification problem into a trivial one, nor be too different, so that molecules could show different profiles of activity or action mode.[82,83]

To build and validate the models, 3D-Pharma uses a new protocol inspired by the work of Tropsha[84,85] on validation problems of QSAR models. In his work, Tropsha argues that the model must be first exhaustively tested and validated internally before being used in external comparisons. In order to accomplish this, one should generate multiple training and test groups, and only the most internally consistent models should be considered in an external validation. 3D-Pharma splits the active molecules among ten groups, using the average 3D similarity between them to create homogeneous groups. These groups are used to build models in a stratified 10-fold cross-validation scheme where each group plays the role of test group once, and the remaining nine groups are recursively split between six training groups and three evaluation groups. Each training group is used to build a model, which is compared with the molecules in its correspondent evaluation group. Since 84 combinations of training/evaluation sets can be formed from nine groups, 84 models are generated. Of these, only the ten models with the highest average similarity to the molecules of its respective evaluation group are selected. These models are then used as a query to recover the

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

test group among a set of inactive molecules, and only the one which has the best recovery rate, measured by the area under the Receiver Operating Characteristic (ROC) curve is retained (Figure 2). Subsequently, the next group assumes the role of test group and the process is repeated. At the end, the full protocol generated a total of 840 models and produced 10 final models, one per test group. Henceforth, all results from 3D-Pharma presented here were averaged over these 10 models. When a dataset contains less than 10 molecules, it is not possible to do a stratified 10-fold cross-validation and, in this case, a simple model is built without internal validation, considering all active molecules to be part of the Training Set.

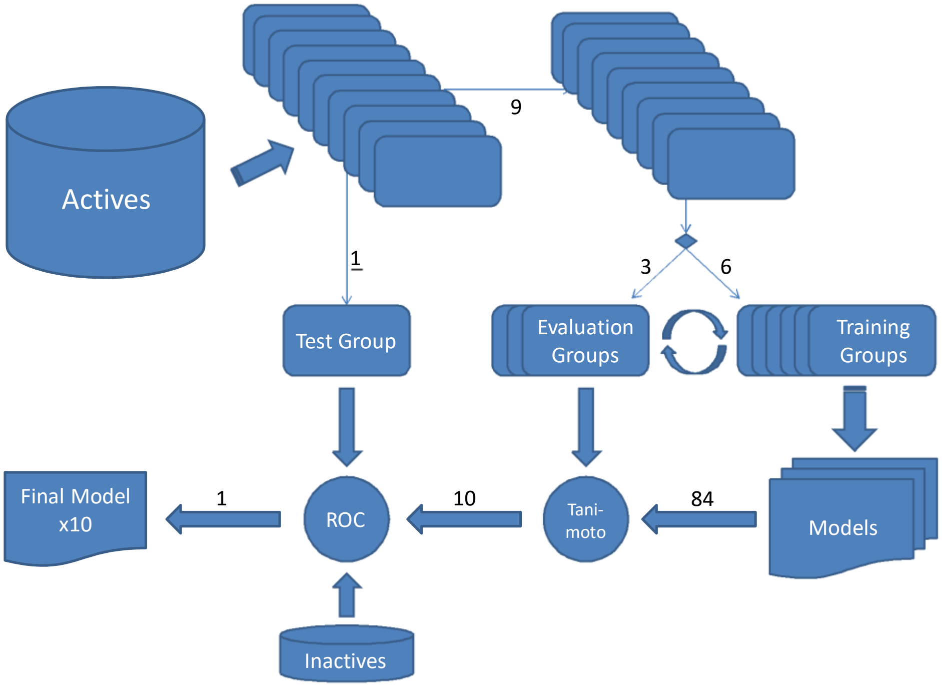

Figure 2: The 3D-Pharma exhaustive 10-fold cross validation scheme used to model construction and validation. Each group out of ten assumes the role of test group once, and the remaining groups are systematically split between training (six groups) and evaluation (three groups) sets. Each possible distribution generates a model from its training set to be compared against its corresponding evaluation set. There are 84 possible distributions and 84 models are generated. The 10 models most similar to its evaluation set are selected to a final validation against the test set. Only the highest predictive model for each test set is selected, resulting in 10 final models built by 3D-Pharma.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Metrics of Evaluation

To assess the performance of 3D-Pharma and also compare it with other available tools, a set of metrics were defined for measuring the performance of 3D-Pharma in VS applications. Three key features should be addressed when designing a VS method, and each one has its own set of metrics:

- Accuracy - The overall performance of a VS method. Can be easily quantified using the area under the Receiver-Operating Characteristic curve (AUC_{ROC}), which is a statistically relevant and unbiased metric for classification performance assessment.[86]

- Early Recognition - The capacity of the VS method to recover active compounds at early cuts. The AUC_{ROC}'s capacity to assess the "early recognition" has been criticized,[87] since whenever a true positive is found, it's contribution to the final score is proportionally the same regardless of ranking position. Usually results published elsewhere rely on Enrichment Factor (EF) of selected cuts to assess the "Early Recognition" problem. However, this metric is not suitable as it depends on the size of the database and on the actives/inactives ratio. Even in standardized databases, unbiased metrics are preferable. In a search for a better metric, Truchon and Bayly [87] generalized the Receiver-Operator Characteristic and designed the parametric Boltzmann-Enhanced Discrimination ROC curve ($BEDROC_{\alpha}$). The α parameter is used to specify the range of the ranked list that would contribute the most to the overall score. In their work, Truchon and Bayly formalized this relation and suggested some α values. Within these suggestions, values of $\alpha = 160.9$, 32.2 and 20, which corresponds respectively to an EF at 1%, 5% and 8% of the selection, were chosen.

- Scaffold Hoping - The ability of the VS method to find novel (or diverse) molecular scaffolds. To account for scaffold hopping in retrospective studies, one would need to cluster the actives into groups of similar molecular structures. Since the DUD database has already clustered its active molecules, it is straightforward to apply the metrics. The arithmetic weighting of the ROC curve (awROC)[88] was used to assess scaffold hopping capabilities, which weights

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

the ROC curve to take into account cluster information. A true positive influence in the score is inversely proportional to the size of the cluster that it is inserted into, so early recognition of low represented clusters contributes more to the final score.

Methods in LBVS used for comparison

The performance of 3D-Pharma was compared with those of other LBVS methods that also used DUD as a benchmarking dataset and its authors supplied enough data to support the full compari- son.

Optimal Assignment methods

Optimal Assignment is a graph theory solution to the optimization problem. Given a bipartite graph where each node is linked to another node by a weighted edge, an optimal assignment is a graph-matching where the sum of the edges is maximized. This was first applied in molecular similarity by Fröhlich and co-workers,[89,90] when they created the OAK (Optimal Assignment Kernel) method. $\mathrm{OAK}_{\mathrm{FLEX}}$ was a modification of OAK made by Fechner et al[91] that included conformational space similarity into the calculations. Other implementations of algorithms which mapped the optimal assignment problem into molecular similarity measures include 2SHA (Two-Step Hierarchical Assignment)[68] and OAAP (Optimal Local Atom Pair Environment Assignment).[68]

4D $\mathrm{FAP}_{\mathrm{OA}}$

4D $\mathrm{FAP}_{\mathrm{OA}}$[59] generates a very large ensemble of conformations whose atom-pair distance profiles are encoded in a series of Gaussian Mixed Models (GMM) generating a single probabilistic model. The energy of conformations is used as a weight factor of each measured atom-pair distance in the GMM generation. The complete information of the conformational space of a molecule is encoded into a list of Gaussian mixture models that could be used to compare different molecules without the need for original conformational ensemble. The final similarity value is computed through an optimal assignment algorithm over atom-pairs in a distance matrix.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

FLAP

FLAP (Fingerprints for Ligands and Proteins)[46–49] is a well-known 3D fingerprint tool that utilizes molecular interaction fields (MIFs) from both ligand and target structures, generated by the program GRID.[92] Each grid point with a local maximal value of the MIF generates a pharmacophoric point of the type of the probe used to generate the MIF and all tetrahedrons formed by these points are stored in a fingerprint. A similarity search is made considering the fingerprints and the alignment of the query molecule to the template. In the recent work by Cross et al,[49] data fusion techniques were used over several data sources to improve recall rates. Of these, LBtParetoR and LBOpt were the best ligand-based techniques reported. LBtParetoR uses a recursive Pareto sum ranking of the alignments, using the DUD cluster representatives (DUD-Parents) as templates. The LBOpt mode uses information of inactive compounds to choose the best template among the DUD-Parents.

FieldScreen

FieldScreen[93] computes molecular field points around a "relevant" conformation of the query molecule and searches a multiconformer database for matching patterns, using maximal colored cliques of field points for the alignment.

Results and discussion

All compounds included in this study (DUD Actives, DUD Decoys, Drugs, PDB Ligands and WOMBAT entries) were submitted to the same protocol of molecular treatment. A set of models was built from the molecules of each active dataset (except COX-2 PDB Ligands, FX_α Drugs and HIVRT Drugs, which had less than 10 molecules in the dataset). Each model was used as query in a similarity search against a pool of molecules formed by DUD Actives and DUD Decoys. The resulting ranking was evaluated through the metrics aforementioned and compared to the LBVS methods mentioned previously.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

As shown in Figure 4, only 3D-Pharma and 4D FAP_{OA} had data comprising all 10 targets included in this study. The other techniques mentioned above only made data available for 13 targets with high chemotype diversity (greater than 15 classes). Within these, there are seven targets in common with our selection : CDK2, COX-2, EGFR, FX_{α} , HIVRT, P38 and PDE5. Hence, all averaged data on the LBVS methods depicted in Figure 5 and Figure 6 (except for 3D-Pharma and 4D FAP_{OA}) are averaged across these seven targets.

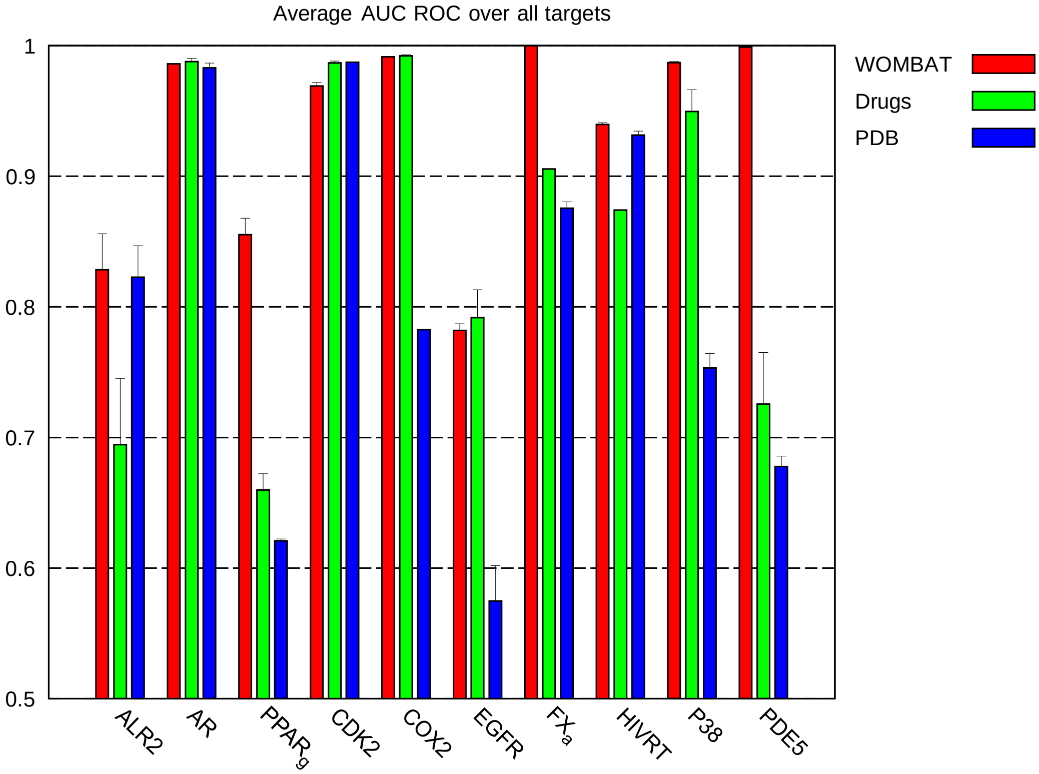

Figure 3: Average AUC_{ROC} produced by 3D-Pharma for all targets. For each dataset, the AUC_{ROC} value was averaged over the 10 models derived from the 10-fold Cross-validation protocol. The error bars correspond to the calculated standard deviation for each dataset.

As seen in Figure 3, 3D-Pharma had an excellent overall accuracy, with 20 out of 30 models with AUC_{ROC} above 0.8. Of these, 14 had AUC_{ROC} above 0.9. As for the targets, seven out of ten had at least one dataset with AUC_{ROC} above 0.9, with nine out of ten targets with at least one model with AUC_{ROC} above 0.8. The models constructed from the WOMBAT database presented the best accuracy, with average AUC_{ROC} of 0.93 ± 0.08, compared to the other datasets (Drugs AUC_{ROC} = 0.85 ± 0.12 and PDB AUC_{ROC} = 0.80 ± 0.14). This might be due to the fact that

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

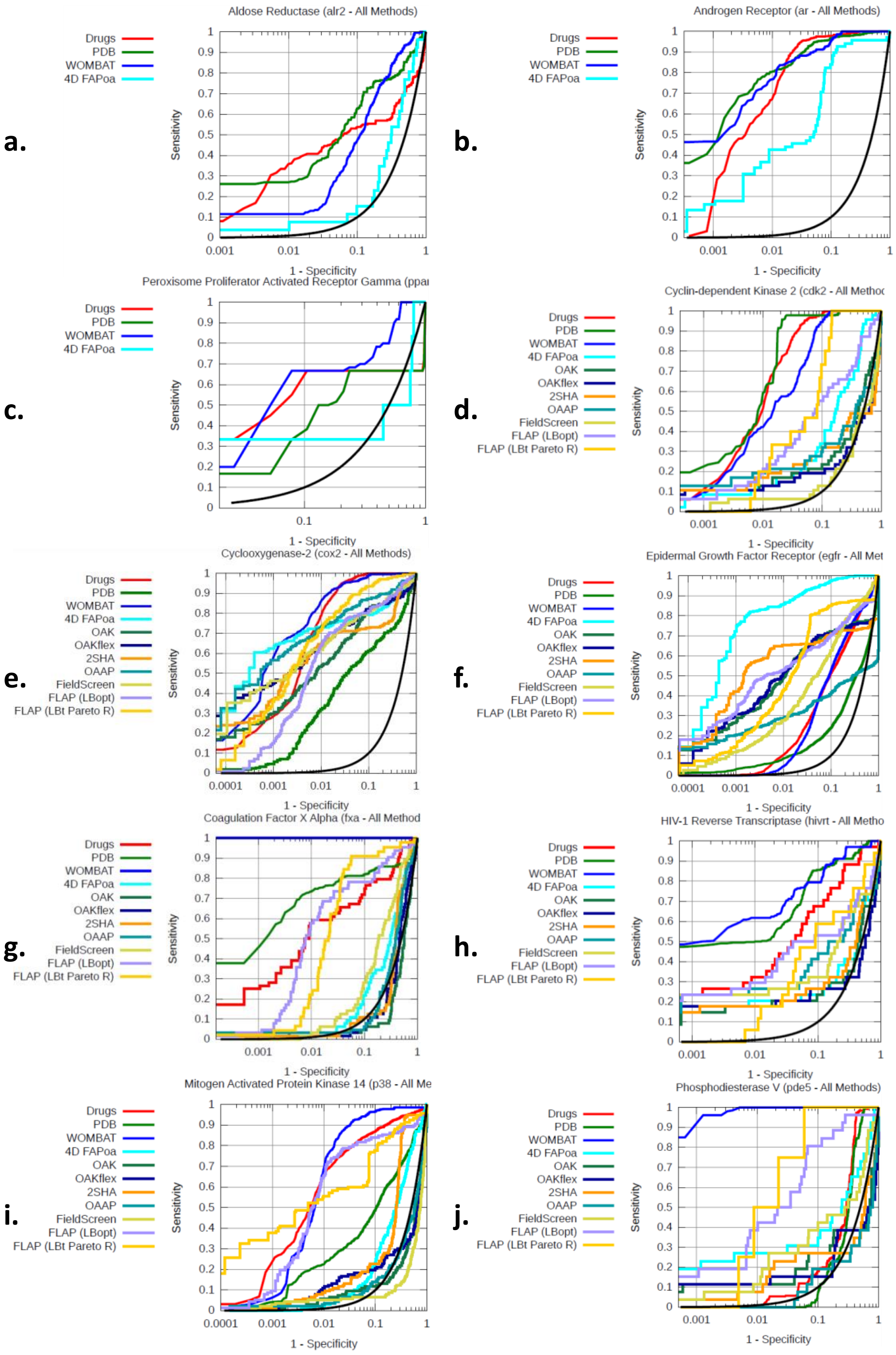

Figure 4: Logarithmic ROC plots of all analysed Ligand-Based Virtual Screening tools against the DUD datasets: a) AR b) ALR2 c) PPAR$_\gamma$ d) CDK2 e) COX2 f) EGFR g) FX$_\alpha$ h) HIVRT i) P38 j) PDE5

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

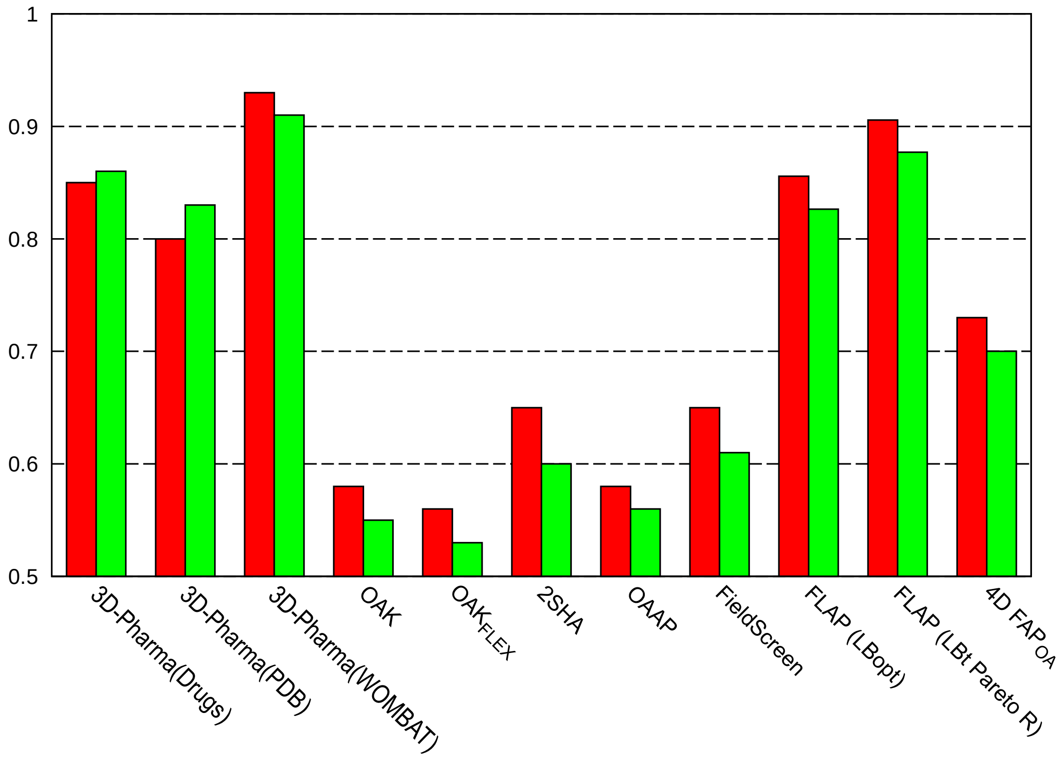

Figure 5: Average AUC_{ROC} and AUC_{awROC} over 10 selected targets for 3D-Pharma (datasets WOMBAT, Drugs and PDB Ligands) and 4D FAP_{OA}. The results of the remaining six Ligand-Based Virtual Screening tools were averaged over seven targets. It is worth noting that all methods, except 3D-Pharma with PDB ligands and Drugs datasets, show a decrease in the area under the curve (AUC) from ROC to awROC.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

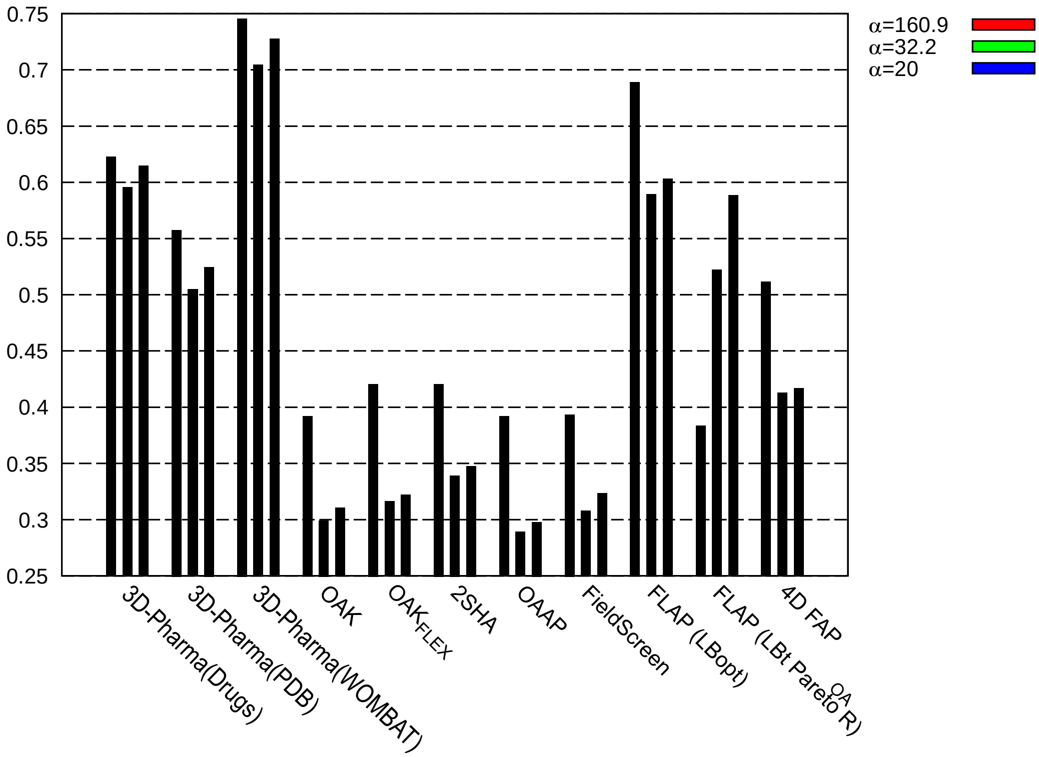

Figure 6: Average BEDROC$_{\alpha}$ scores over 10 selected targets for 3D-Pharma (datasets WOMBAT, Drugs and PDB Ligands) and 4D FAP$_{OA}$. The results of the remaining six Ligand-Based Virtual Screening tools were averaged over seven targets. The values for the α parameter were 160.9, 32.2 and 20, which correspond to Enrichment Factors (EF) at 1%, 5% and 8% of selection, respectively.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

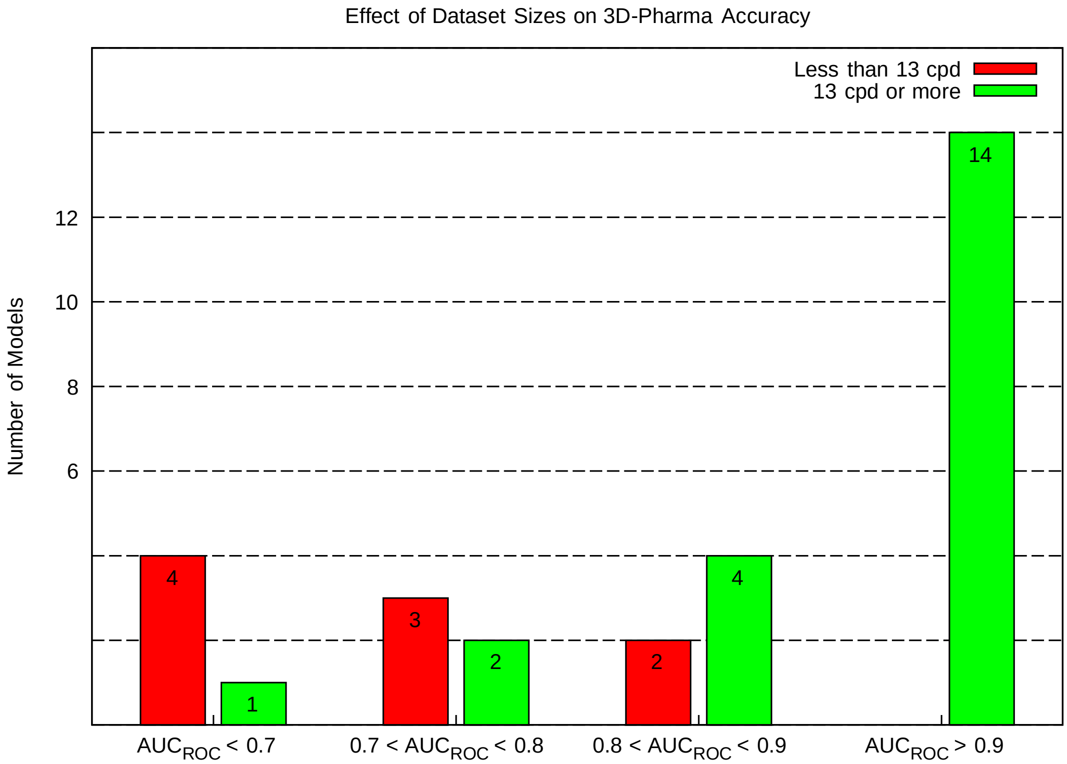

Figure 7: The effect of dataset sizes on 3D-Pharma accuracy. For datasets with 13 or more compounds most datasets (14 out of 21) produced high quality models (AUC> 0.9). For those datasets with 12 or less compounds, two out of nine datasets produced good models (AUC> 0.8).

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

WOMBAT is a solid well stablished database, whereas the Drugs and PDB are still a growing set of gathered data. Besides the nature and quality of datasets used to build the models, it seems that the size of the datasets shows the major impact on the performance of 3D-Pharma (Figure 7). For those datasets with at least 13 compounds (21 datasets), not less than 14 (67%) produced high quality models with AUC_{ROC} above 0.9, and 18 (86%) produced very good models with AUC_{ROC} above 0.8. On the other hand, for those datasets with less than 13 compounds, none was able to

generate models with AUC_{ROC} above 0.9 and only two out of nine (22%) generated models with AUC_{ROC} above 0.8.

In general, 3D-Pharma outperforms the other methods regardless of the dataset chosen to build the models, with FLAP as a close second. However, despite the results achieved, both FLAP ligand- based approaches are subject to one drawback: the use of DUD cluster-parent actives as query or templates. Therefore, the analysis is subject to analogue bias[67] that could potentially increment artificially the results. Another interesting observation arises when BEDROC scores are analyzed. When one looks at the bigger cuts ($\alpha = 32.2$ and 20), all methods analyzed have an abrupt fall in the early recovery rate, as seen in Figure 6, but 3D-Pharma fairly maintains its scores as α diminishes. It seems that the higher $BEDROC_{160.9}$ score with a subsequent substantial decrease on $BEDOC_{32.2}$ is an indication of some kind of analog bias that puts a few active compounds very high in the ranking (before 1% of selection) and leaves many other active molecules spread over the rank positions. In 3D-Pharma we used three really external public datasets, used "as is", that is, without any kind of filter, except for redundant molecules between the DUD actives and external datasets. As a consequence, the BEDROC score is sustained over all α values used in the benchmark.

The LBOpt and LBParetoR FLAP scenarios yielded very good overall results and performed better than the other methods surveyed, except for 3D-Pharma. LBParetoR uses all DUD chemotype cluster parents (DUD-own dataset) for each target as query in an ensemble approach with con- sensus analysis of the individual template similarity results. Consequently, the number of DUD actives in the ROC analysis is smaller than is all other methods under comparison in this paper, and

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

the conclusions must be considered carefully. When we look at the ROC and awROC from LBPare- toR, the results are impressive, with average AUC above 0.9 for the seven targets under scrutiny. However, the analysis of BEDROC results discloses a disappointing "early recovery", mainly for $\text{BEDROC}_{160.9}$ and $\text{BEDROC}_{32.2}$. The LBOpt approach produces smaller AUC in the ROC analysis than LBParetoR, but the "early recovery" is much better. LBopt also uses the DUD-Parents as query, but they are subject to a previous optimal template selection that optimizes proportions of false positives and false negatives in order to select the single template to be used as query.

When analyzing scaffold hopping capabilities, one can note that all techniques, except 3D-Pharma, have a significant drop in their average AUCs when considering awROC over the standard ROC (Figure 5). On the contrary, 3D-Pharma using the Drugs and PDB-Ligands datasets gave an increase in the score. The other techniques tend to rank higher the most populated scaffolds, hence lowering $\text{AUC}_{\text{awROC}}$ scores in relation to AUC_{ROC}.

Conclusions

The main characteristics of 3D-Pharma are the use of pharmacophore triplets fingerprint based on atom-centered potential pharmacophores, the use of several representations of the compound that include tautomers, protonation states and conformers and the ensemble template approach producing a single modal fingerprint based on frequency of pharmacophore triplets over the active compounds. This dynamic pharmacophore fingerprint encodes all chemical and conformational variability of the compound in a single fingerprint representation. Thus, the 3D-Pharma approach adopts a paradigm where the full ensemble of conformers is taken into account at the same time in a single modal fingerprint, similar to the approach implemented by Ranu and Singh.[60] In our study, 30 sets of models were generated for 10 selected targets from the DUD database, using three external and independent datasets as reference. Is seems that the size of the modeling dataset exerts a major impact on the model quality. For those datasets with at least 13 compounds (21 datasets) not less than 18 (or 86%) were able to produce very good models with AUC_{ROC} above 0.8, and 14

https://doi.org/10.26434/chemrxiv-2024-dkvf8 ORCID: https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

datasets (67%) produced very high-quality models with an AUC_{ROC} above 0.9.

The analysis of the scaffold hopping and early recovery capabilities of 3D-Pharma has shown two distinguishing behaviours. The AUC_{awROC}, used to estimate scaffold hopping capabilities, shows evidence that 3D-Pharma with Drugs and PDB Ligand datasets has a better performance in detecting rarer scaffolds than the other LBVS tools analyzed. The second 3D-Pharma discerning behaviour can be seen in the BEDROC plot (Figure 6) where 3D-Pharma datasets sustain high scores over the three values chosen for the α parameter. All other LBVS tools (except FLAP LBParetoR) presented a higher score at lower cuts ($BEDROC_{160.9}$) than those at higher cuts, with a significant decrease in the $BEDROC_{32.2}$ and $BEDROC_{20}$ scores.

Thus, the data shown here leads us to strongly believe that 3D-Pharma outperforms all other state-of-the-art LBVS tools analyzed, in terms of global accuracy as well as scaffold hopping and early recovery capacities. The fact that three external datasets were used to generate the models that are at the same time simple, robust, consistent and predictive should be highlighted. Its predictive power in prospective virtual screening cases remains to be seen, but as far as the results shown here can assess, 3D-Pharma is a promising method that can effectively contribute to the success of any drug discovery process.

Acknowledgement

We would like to acknowledge the following agencies for their support: CNPq, CAPES and FAPEMIG. We would also like to thank ChemAxon and OpenEye Scientific Software for providing us with the academic licenses to their products.

Supporting Information Available

Supporting Information contains the ROC plots and complete tables comparing all methods for each target. It also contains the molecular files (in SMILES) for Drugs and PDB-Ligands datasets for each target, and the WOMBAT IDs considered in this study. The full ranking od DUD molecules from each model is also available within the files. Please refer to the README.txt

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

file for more details.

References

(1) Paul, S.; Mytelka, D.; Dunwiddie, C.; Persinger, C.; Munos, B.; Lindborg, S.; Schacht, A. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery 2010, 9, 203–214.

(2) IMAP's Pharma & Biotech Industry Global Report 2011. http://www.imap.com/imap/media/resources/IMAP_PharmaReport_8_272B8752E0FB3.pdf, 2011; Accessed 05/01/2012.

(3) Koenig, J. Does process excellence handcuff drug development? Drug Discovery Today 2011, 16, 377 – 381.

(4) Harrison, C. Patent Watch: The patent cliff steepens. Nature Reviews Drug Discovery 2011, 10, 12–13.

(5) Arrowsmith, J. A decade of change. Nat Rev Drug Discov 2012, 11, 17–18.

(6) Mucsi, Z.; Csizmadia, I. The Future of the Drug Discovery Process and the Fate of the Pharmaceutical Industry: An economical and scientific study. Philosophic Nature 2009, 1, 61-75.

(7) Mayr, L. M.; Fuerst, P. The Future of High-Throughput Screening. Journal of Biomolecular Screening 2008, 13, 443–448.

(8) Kümmel, A.; Parker, C. N. In Chemoinformatics and Computational Chemical Biology; Bajorath, J., Ed.; Springer Science+Business Media, LLC 2011, 2011; Chapter 17, pp 435–457.

(9) Bajorath, J. Integration of virtual and high-throughput screening. Nature Reviews Drug Discovery 2002, 1, 882–894.

(10) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-art in ligand-based virtual screening. Drug Discovery Today 2011, 16, 372 – 376.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(11) Cortés-Cabrera, Á.; Gago, F.; Morreale, A. A reverse combination of structure-based and ligand-based strategies for virtual screening. Journal of Computer-Aided Molecular Design 2012, 26, 319–327.

(12) Svensson, F.; Karlén, A.; Sköld, C. Virtual Screening Data Fusion Using Both Structure- and Ligand-Based Methods. Journal of Chemical Information and Modeling 2012, 52, 225–232.

(13) Swann, S. L.; Brown, S. P.; Muchmore, S. W.; Patel, H.; Merta, P.; Locklear, J.; Hajduk, P. J. A Unified, Probabilistic Framework for Structure- and Ligand-Based Virtual Screening. Journal of Medicinal Chemistry 2011, 54, 1223–1232.

(14) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. Journal of Medicinal Chemistry 2010, 53, 539–558.

(15) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. Journal of Chemical Information and Modeling 2010, 50, 205–216.

(16) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo Vadis, Virtual Screening? A Comprehensive Survey of Prospective Applications. Journal of Medicinal Chemistry 2010, 53, 8461–8467.

(17) Arrowsmith, J. Trial watch: Phase II failures: 2008-2010. 2011, 10, 328–329.

(18) Colburn, W. Biomarkers in drug discovery and development: from target identification through drug marketing. The Journal of Clinical Pharmacology 2003, 43, 329–341.

(19) Virtanen, S.; Pentikäinen, O. Efficient virtual screening using multiple protein conformations described as negative images of the ligand-binding site. Journal of Chemical Information and Modeling 2010, 50, 1005–1011.

(20) Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P.; Whitmore, A.; Zimmer, S.; Young, M.; Jenkins, J.; Glick, M.; Glen, R.; Bender, A. From in silico target prediction to multi-target

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

drug design: Current databases, methods and applications. Journal of Proteomics 2011, 74, 2554–2574.

(21) Marrero-Ponce, Y.; Siverio-Mota, D.; Gálvez-Llompart, M.; Recio, M. C.; Giner, R. M.; García-Domènech, R.; Torrens, F.; Arán, V. J.; Cordero-Maldonado, M. L.; Esguera, C. V.; de Witte, P. A.; Crawford, A. D. Discovery of novel anti-inflammatory drug-like compounds by aligning in silico and inăvivo screening: The nitroindazolinone chemotype. European Journal of Medicinal Chemistry 2011, 46, 5736 – 5753.

(22) Bottegoni, G.; Favia, A. D.; Recanatini, M.; Cavalli, A. The role of fragment-based and computational methods in polypharmacology. Drug Discovery Today 2012, 17, 23 – 34.

(23) Ekins, S.; Mestres, J.; Testa, B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. British Journal of Pharmacology 2007, 152, 9–20.

(24) Bajorath, J. Computational analysis of ligand relationships within target families. Current Opinion in Chemical Biology 2008, 12, 352–358.

(25) Luis G. Valerio, J. In silico toxicology for the pharmaceutical sciences. Toxicology and applied pharmacology 2009, 241, 356–370.

(26) Maggiora, G. The reductionist paradox: are the laws of chemistry and physics sufficient for the discovery of new drugs? Journal of Computer-Aided Molecular Design 2011, 25, 699–708.

(27) Willett, P. Similarity-based virtual screening using 2D fingerprints. Drug discovery today 2006, 11, 1046–1053.

(28) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods. Journal of Chemical Information and Modeling 2010, 50, 2079–2093.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(29) Stumpfe, D.; Bajorath, J. In Virtual Screening; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA, 2011; Chapter 11, pp 291–318.

(30) Sitzmann, M.; Weidlich, I. E.; Filippov, I. V.; Liao, C.; Peach, M. L.; Ihlenfeldt, W.-D.; Karki, R. G.; Borodina, Y. V.; Cachau, R. E.; Nicklaus, M. C. PDB Ligand Conformational Energies Calculated Quantum-Mechanically. Journal of Chemical Information and Modeling 2012, 52, 739–756.

(31) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. i. E. The Protein Data Bank. Nucl. Acids Res. 2000, 28, 235–242.

(32) Mattos, C.; Ringe, D. In 3D QSAR in Drug Design: Theory, Methods and Applications; Kubinyi, H., Ed.; Escom, 1993; pp 226–256.

(33) Lewi PJ, de Jonge M, Daeyaert F, Koymans L, Vinkers M, Heeres J, Janssen PA, Arnold E, Das K, Clark AD Jr, Hughes SH, Boyer PL, de Béthune MP, Pauwels R, Andries K, Kukla M, Ludovici D, De Corte B, Kavash R, Ho C. On the detection of multiple-binding modes of ligands to proteins, from bio- logical, structural, and modeling data. Journal of Computer-Aided Molecular Design 2003, 17, 129–134.

(34) DePristo, M. A.; de Bakker, P. I. W.; Blundell, T. L. Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. Structure 2004, 12, 831 – 838.

(35) Steuber, H.; Zentgraf, M.; Gerlach, C.; Sotriffer, C.; Heine, A.; Klebe, G. Expect the unexpected or caveat for drug designers: multiple structure determinations using aldose reductase crystals treated under varying soaking and co-crystallisation conditions. Journal of Molecular Biology 2006, 363, 174–187.

(36) Tresadern, G.; Bemporad, D.; Howe, T. A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor. Journal of Molecular Graphics and Modelling 2009, 27, 860–870.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(37) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations Among Molecules. Journal of Chemical Information and Computer Sciences 1996, 36, 563–571.

(38) Richmond, N.; Abrams, C.; Wolohan, P.; Abrahamian, E.; Willett, P.; Clark, R. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. Journal of Computer-Aided Molecular Design 2006, 20, 567–587.

(39) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. Journal of Computer-Aided Molecular Design 1995, 9, 532–549.

(40) Chemical Computing Group, Molecular Operating Environment. www.chemcomp.com.

(41) Dixon, S.; Smondyrev, A.; Knoll, E.; Rao, S.; Shaw, D.; Friesner, R. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. Journal of Computer-Aided Molecular Design 2006, 20, 647–671.

(42) Schneidman-Duhovny, D.; Dror, O.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. Deterministic Pharmacophore Detection via Multiple Flexible Alignment of Drug-Like Molecules. Journal of Computational Biology 2008, 15, 737–754.

(43) Dror, O.; Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. Novel Approach for Efficient Pharmacophore-Based Virtual Screening: Method and Applications. Journal of Chemical Information and Modeling 2009, 49, 2333–2343.

(44) Cottrell, S. J.; Gillet, V. J.; Taylor, R.; Wilton, D. J. Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques. Journal of Computer-Aided Molecular Design 2004, 18, 665–682.

(45) Kristam, R.; Gillet, V. J.; Lewis, R. A.; Thorner, D. Comparison of Conformational Analysis Techniques To Generate Pharmacophore Hypotheses Using Catalyst. Journal of Chemical Information and Modeling 2005, 45, 461–476.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(46) Perruccio, F.; Mason, J. S.; Sciabola, S.; Baroni, M. In Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction; Cruciani, G., Ed.; Wiley-VCH: Weinheim, Germany, 2006; Chapter 4.

(47) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. J Chem Info Model 2007, 47, 279–294.

(48) Carosati, E.; Sforna, G.; Pippi, M.; Marverti, G.; Ligabue, A.; Guerrieri, D.; Piras, S.; Guaitoli, G.; Luciani, R.; Costi, M. P.; Cruciani, G. Ligand-based virtual screening and ADME-tox guided approach to identify triazolo-quinoxalines as folate cycle inhibitors. Bioorganic & Medicinal Chemistry 2010, 18, 7773 – 7785.

(49) Cross, S.; Baroni, M.; Carosati, E.; Benedetti, P.; Clementi, S. FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation using the DUD Data Set. Journal of Chemical Information and Modeling 2010, 50, 1442–1450.

(50) Koes, D. R.; Camacho, C. J. Pharmer: Efficient and Exact Pharmacophore Search. Journal of Chemical Information and Modeling 2011, 51, 1307–1314.

(51) Tripos, L.P., Tuplets. www.tripos.com.

(52) Accelrys Software, Cerius^2. www.accelrys.com.

(53) de Benedetti, P.; Fanelli, F. Computational quantum chemistry and adaptive ligand modeling in mechanistic QSAR. Drug Discovery Today 2010, 15, 859–866.

(54) Schneider, G. Virtual screening: an endless staircase? Nat Rev Drug Discov 2010, 9, 273–276.

(55) Kubinyi, H. Drug research: myths, hype and reality. Nature Reviews Drug Discovery 2003, 2, 665–667.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(56) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. Journal of Chemical Information and Modeling 2012, 52, 867–881.

(57) Natesan, S.; Wang, T.; Lukacova, V.; Bartus, V.; Khandelwal, A.; Balaz, S. Rigorous Treatment of Multispecies Multimode Ligand-Receptor Interactions in 3D-QSAR: CoMFA Analysis of Thyroxine Analogs Binding to Transthyretin. Journal of Chemical Information and Modeling 2011, 51, 1132–1150.

(58) Natesan, S.; Subramaniam, R.; Bergeron, C.; Balaz, S. Binding Affinity Prediction for Ligands and Receptors Forming Tautomers and Ionization Species: Inhibition of Mitogen-Activated Protein Kinase-Activated Protein Kinase 2 (MK2). Journal of Medicinal Chemistry 2012, 55, 2035–2047.

(59) Jahn, A.; Rosenbaum, L.; Hinselmann, G.; Zell, A. 4D Flexible Atom-Pairs: An efficient probabilistic conformational space comparison for ligand-based virtual screening. Journal of Cheminformatics 2011, 3, 23.

(60) Ranu, S.; Singh, A. K. Novel Method for Pharmacophore Analysis by Examining the Joint Pharmacophore Space. Journal of Chemical Information and Modeling 2011, 51, 1106–1121.

(61) Pérez-Nueno, V. I.; Ritchie, D. W. Using Consensus-Shape Clustering To Identify Promiscuous Ligands and Protein Targets and To Choose the Right Query for Shape-Based Virtual Screening. Journal of Chemical Information and Modeling 2011, 51, 1233–1248.

(62) Sastry, G. M.; Dixon, S. L.; Sherman, W. Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlap Scoring. Journal of Chemical Information and Modeling 2011, 51, 2455–2466.

(63) Cai, C.; Gong, J.; Liu, X.; Jiang, H.; Gao, D.; Li, H. A novel, customizable and optimizable parameter method using spherical harmonics for molecular shape similarity comparisons. Journal of Molecular Modeling 2012, 18, 1597–1610.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

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

(65) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. Journal of Chemical Information and Computer Sciences 2004, 44, 1177–1185.

(66) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. In Chemoinformatics in Drug Discovery; Oprea, T. I., Ed.; Wiley-VCH: New York, 2004; Chapter 9, pp 223–239.

(67) Good, A.; Oprea, T. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? Journal of Computer-Aided Molecular Design 2008, 22, 169–178.

(68) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal assignment methods for ligand-based virtual screening. Journal of Cheminformatics 2009, 1, 14.

(69) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. Journal of Chemical Information and Computer Sciences 2001, 41, 1308–1315.

(70) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Research 2006, 34, D668–D672.

(71) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledge base for drugs, drug actions and drug targets. Nucl. Acids Res. 2008, 36, D901–906.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(72) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucl. Acids Res. 2000, 28, 27–30.

(73) Chen, X.; Ji, Z. L.; Chen, Y. Z. TTD: Therapeutic Target Database. Nucl. Acids Res. 2002, 30, 412–415.

(74) ChemAxon, www.chemaxon.com, 1999–2012.

(75) Mobley, D. L.; Dumont, É.; Chodera, J. D.; Dill, K. A. Comparison of Charge Models for Fixed-Charge Force Fields: Small-Molecule Hydration Free Energies in Explicit Solvent. The Journal of Physical Chemistry B 2007, 111, 2242–2254.

(76) Tsai, K.-C.; Wang, S.-H.; Hsiao, N.-W.; Li, M.; Wang, B. The effect of different electrostatic potentials on docking accuracy: A case study using DOCK5.4. Bioorganic & Medicinal Chemistry Letters 2008, 18, 3509 – 3512.

(77) Wang, J.-C.; Lin, J.-H.; Chen, C.-M.; Perryman, A. L.; Olson, A. J. Robust Scoring Functions for Protein-Ligand Interactions with Quantum Chemical Charge Models. Journal of Chemical Information and Modeling 2011, 51, 2528–2537.

(78) OpenEye Scientific Software, OMEGA. www.eyesopen.com, 1997–2012.

(79) Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. Journal of Computational Chemistry 1999, 20, 720–729.

(80) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. Journal of Computational Chemistry 2002, 23, 1623–1641.

(81) Botelho, F. C.; Ziviani, N. External perfect hashing for very large key sets. CIKM ’07: Proceedings of the sixteenth ACM Conference on information and knowledge management. Lisbon, Portugal, 2007; pp 653–662.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(82) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Journal of Computer-Aided Molecular Design 2002, 16, 357–369.

(83) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. Journal of Computer-Aided Molecular Design 2003, 17, 241–253.

(84) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR & Combinatorial Science 2003, 22, 69–77.

(85) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics 2010, 29, 476–488.

(86) Nicholls, A. In Chemoinformatics and Computational Chemical Biology; Bajorath, J., Ed.; Springer Science+Business Media, LLC 2011, 2011; Chapter 22, pp 531–581.

(87) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. Journal of Chemical Information and Modeling 2007, 47, 488–508.

(88) Clark, R.; Webster-Clark, D. Managing bias in ROC curves. Journal of Computer-Aided Molecular Design 2008, 22, 141–146.

(89) Fröhlich, H.; Wegner, J. K.; Sieker, F.; Zell, A. Optimal assignment kernels for attributed molecular graphs. ICML. 2005; pp 225–232.

(90) Fröhlich, H.; Wegner, J. K.; Sieker, F.; Zell, A. Kernel Functions for Attributed Molecular Graphs - A New Similarity-Based Approach to ADME Prediction in Classification and Regression. QSAR & Combinatorial Science 2006, 25, 317–326.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(91) Fechner, N.; Jahn, A.; Hinselmann, G.; Zell, A. Atomic Local Neighborhood Flexibility Incorporation into a Structured Similarity Measure for QSAR. Journal of Chemical Information and Modeling 2009, 49, 549–560.

(92) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985, 28, 849–857.

(93) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. Journal of Chemical Information and Modeling 2008, 48, 2108–2117.

This material is available free of charge via the Internet at http://pubs.acs.org/.

https://doi.org/10.26434/chemrxiv-2024-dkvf8 **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0