Pros and cons of virtual screening based on public “Big Data”: In silico mining for new bromodomain inhibitors

Iuri Casciuc, Dragos Horvath, Anastasiia Gryniukova, Kateryna Tolmachova, Oleksandr Vasylchenko, Petro Borysko, Yurii Moroz, Jürgen Bajorath, Alexandre Varnek

Pros and Cons of Virtual Screening Based on Public "Big Data": In Silico Mining for New Bromodomain Inhibitors

Iuri CASCIUC[1], Dragos HORVATH[1], Anastasiia GRYNIUKOVA[2], Kateryna A. TOLMACHOVA[3,4], Oleksandr V. VASYLCHENKO[3], Petro BORYSKO[2], Yurii S. MOROZ[5,6], Jürgen BAJORATH[7], Alexandre VARNEK[1 *]

1) Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 4, Blaise Pascal str., 67081 Strasbourg, France

2) Bienta/Enamine Ltd., Chervonotkatska Street 78, Kyiv 02094, Ukraine

3) Enamine Ltd., Chervonotkatska Street 78, Kyiv 02094, Ukraine

4) Institute of Bioorganic Chemistry & Petrochemistry, NAS of Ukraine, Murmanska Street 1, Kyiv 02660, Ukraine

5) National Taras Shevchenko University of Kyiv, Volodymyrska Street 60, Kyiv 01601, Ukraine

6) Chemspace, ilukstes iela 38-5, Riga, LV-1082, Latvia www.chem-space.com

7) B-IT, Limes, Unit Chem. Biol. & Med. Chem., University of Bonn, Germany

Abstract

The Virtual Screening (VS) study described herein aimed at detecting novel Bromodomain BRD4 binders and relied on knowledge from public databases (ChEMBL, REAXYS) to establish a battery of predictive models of BRD activity for in silico selection of putative ligands. Beyond the actual discovery of new BRD ligands, this represented an opportunity to practically estimate the actual usefulness of public domain "Big Data" for robust predictive model building. Obtained models were used to virtually screen a collection of 2 million compounds from the Enamine company collection. This industrial partner then experimentally screened a subset of 2992 molecules selected by the VS procedure for their high likelihood to be active. Twenty nine confirmed hits were detected after experimental testing, representing 1% of the selected candidates. As a general conclusion, this study emphasizes once more that public structure-activity databases are nowadays key assets in drug discovery. Their usefulness is however limited by the state-of-the-art knowledge harvested so far by

published studies. Target-specific structure-activity information is rarely rich enough, and its heterogeneity makes it extremely difficult to exploit in rational drug design. Furthermore, published affinity measures serving to build models selecting compounds to be experimentally screened may not be well correlated with the experimental hit selection criterion (in practice, often imposed by equipment constraints). In spite of this, a robust 2.6-fold increase in hit rate with respect to an equivalent, random screening campaign showed that machine learning is able to extract some real knowledge in spite of all the noise in structure-activity data.

Keywords: Bromodomain BRD4 binders; Generative Topographic Mapping, Virtual Screening, Classification Models, Ligand-based Pharmacophores, Docking

Introduction

The exponential accumulation of structure-activity data in public databases, representing the advent of Big Data in medicinal chemistry is expected to lead to the development of robust and potent *in Silico* Quantitative Structure-Activity Relationships (QSAR), mathematical models able to serve for Virtual Screening (VS) of compound databases, *i.e.* detect and prioritize novel active structures therein and herewith accelerate drug discovery. Both predictive accuracy and Applicability Domain (AD) of QSAR models are expected to increase significantly with the size and chemical diversity of training sets, while machine learning has already provided Big Data-compatible tools for the fitting of such models. Methods like Support Vector Machines (SVM)[1] and Generative Topographic Mapping (GTM)[2] routinely provide QSAR models based on tens of thousands of compounds. GTM – essentially a fuzzy-logic-based variant of popular Self-Organizing (Kohonen) Maps[3] – has no upper limit on training set size, as GTM-driven predictive models consist of property landscapes that are "colored" (created) by projecting known reference actives and inactives on the map, and attributing to every map point a property value equaling the mean of therein residing compounds. For prediction, candidate compounds are also projected on the map, and are assigned the property value of their residence spot or declared "out of AD" if they fall into blank spots, where no reference compounds are residing. GTM [4]·[5]·[6]·[7] is a multivalent "Swiss-army-knife"-like tool of chemoinformatics, as one of the rare tools competent for both chemical space visualization (with particular interest in library comparison) and predictive modeling, including implicit AD assessment. Albeit it is not primarily designed for VS, the latter abilities nevertheless qualify GTM as a robust VS methodology, whilst its visualization support may be useful to graphically compare relevant compound sets (here, reference "actives" from various sources, *vide infra*). By contrast, SVM is one of the most powerful QSAR predictors. Both of these methods are

fast since they can operate on 2D molecular descriptors avoiding the need of a costly conformational sampling step. Pharmacophore models and, eventually, docking, can be used in conjunction to the fast 2D ligand prioritization tool in a VS funnel, to gradually focus in on candidates with maximum probability to be active.

In practice, however, medicinal chemistry data is rather heterogeneous. The nearly two million compounds in ChEMBL are often associated with reliable IC_{50}/K_i measures, but these concern a plethora of different targets. Thus, compound sets associated with a given target are more often likely to represent classical QSAR sets of hundreds of compounds, rather than Big Data sets. Moreover, dose-response activity measures are typically reported by different groups and may follow distinct protocols, which raises the question whether they are comparable. The key point here is that in “classical” medicinal chemistry, the expert is closely following the work of colleagues/competitors and is familiar with all those distinct protocols, knowing what is comparable. Or, in the Big Data era, the information is allegedly too rich to be trackable by a human expert. Structure-activity sets should be algorithmically extracted, standardized and processed into training sets – with no human intervention. Is this a realistic scenario, or would data heterogeneity eventually outweigh the benefits of information richness provided by on-line public databases? This is a central question addressed in this work, which reports a “Big Data” VS search, followed by experimental validation, for novel BRD4 inhibitors. A database of 2 million available compounds from Enamine (enamine.net) was virtually screened using a hierarchy of 2D QSAR methods coupled to pharmacophore screening and docking and publicly available structure-activity data from REAXYS[8] and ChEMBL[9] databases for automated model training.

Readers of post-translational modifications are structurally diverse proteins than contain one or more effector modules that recognize (that is, read) covalent modifications of proteins and DNA. The recognition of ε-N-acetylation of lysine residues is primarily

initiated by bromodomains, a family of evolutionarily conserved protein interaction modules that were identified in the early 1990s in the brahma gene from Drosophila melanogaster[10]. The human genome encodes 61 bromodomains present in 46 different proteins[11],[12], where differences in the amino acid residues around the acetyl-lysine binding site impart ligand specificity. Proteins that contain bromodomains are involved in the regulation of transcriptional programs and have been identified in oncogenic rearrangements that lead to highly oncogenic fusion proteins, which have a key role in the development of several aggressive types of cancer. They are also implicated in the replication of viral genomes and regulate the transcription of some viral proteins.

Bromodomain modules share a conserved fold that comprises a left-handed bundle of fourα-helices (named αZ,αA, αB and αC) that are linked by diverse loop regions of variable charge and length (known as ZA and BC loops) which surround a central acetylated lysine binding site. Structural data have established that acetylated lysine is recognized in a central hydrophobic pocket, where it is anchored to a conserved asparagine residue. More recently, it has been demonstrated that Brd4 bind to two acetylated lysine histone marks that are simultaneously recognized by the same bromodomain module[13]. This property is shared by all members of the Bromodomain Extra-Terminal (BET) subclass of BRDs. High-resolution crystal structures showed that the first acetylated lysine mark of histone H4 docks directly onto the conserved asparagine (Asn140 in the first bromodomain of BRD4). Simultaneously, a network of hydrogen bonds, formed via conserved water molecules found in the bromodomain cavity, link to the second acetylated lysine mark, thus stabilizing the peptide complex.

BRD inhibitors are reported in several public databases – ChEMBL and REAXYS were the ones exploited here and reported inhibition strength stem from various methods such as DSF or FRET experiments. The only way to cope with data heterogeneity was

to base the VS protocol on categorical models, returning an estimate of the likelihood of a candidate compound to be “active”. Training of categorical models however implies an upstream classification of so-far tested compounds into “actives” and “inactives”. The choice of these examples of actives and inactives used in the machine learning process is empirical, as it implies setting arbitrary thresholds in terms of the available affinity scores. These thresholds might not only be activity score-specific but would also depend on the stage of the hit or lead discovery process. Whilst at primary screening stage a 10μM affinity level might count as “active”, this will no longer be the case in the more ambitious hit-to-lead development stage. As no obvious consensus in designing the “active” BRD training set could be reached, several distinct training sets were employed in parallel, featuring various working hypotheses concerning the “actives”.

A battery of SVM and GTM models, combining above-mentioned training set choices and various methodological strategies were built. In parallel, structure-based pharmacophore models were derived from BRD4-ligand crystal structures. All these were used to screen the 2 million compound library of Enamine, and 12000 structures were selected on the basis of a consensus scheme. Experimental screening of a fixed-size pool of 3000 candidates has been carried out by Enamine, representing a VS-driven alternative to a similar screen done on a randomly picked compound set of same size.[1] Docking with S4MPLE[14],[15] was used to further reduce the primary 12K compounds to the final pool of 3000 molecules submitted to testing. Experimental DSF retrieved 29 hits (1%) in the 3K VS-based library – three times more than the base hit rate in the above-mentioned random screening experiment[16]. This is a significant, yet slightly disappointing enrichment factor. As a consequence, more effort has been allotted to better understand the discrepancies in affinity measurements introduced by different methods. On one hand, some ChEMBL training set compounds with reported IC_{50}/K_i values had their melting temperature shift (ΔT_m) measured by DSF under the

same conditions as the herein retrieved hits. Alternatively, IC_{50} values for some of the newly discovered hits were also experimentally determined. The weak to moderate correlation between the actual hit detection criterion (ΔT_m) and the dose-dependent public data affinity scores (on which model training was based) has significantly and negatively impacted the success rate of this large-scale VS experiment. An *a posteriori* analysis of individual models, aimed to verify how well each one performed in ranking the 29 hits within the 3K selection, showed that hit rates for the most successful individual models could have been as high as 10%. However, for every successful model, alternative models of the same category – differing only with respect to the choice of the "actives" defined in the training set, and the added ChEMBL "decoy" molecules – were found to be low performers. This is clear evidence that the problem stems from data variance or noise and not from rigorously cross-validated models.

Materials and Methods

Training sets

Two sources were used in this project:

- The REAXYS set contains 75 strong actives (compounds having $IC_{50} \leq 100nM$), 404 moderate actives (compounds having IC_{50} between 100nM and 10μM) and 742 inactive ($IC_{50} > 10\mu M$) molecules. In order to remain within the two-class classification strategy, this set was duplicated into two "clone" sets differing only with respect to the assignment of the class labels. The Strict set considers only the 75 strong as "active", and all others are "inactive". The Soft set counts both the 75 strong and the 404 intermediates as "actives".
- The ChEMBL data, where active versus inactive BRD compounds were extracted automatically, as part of a data curation procedure internal to the Laboratory of Chemoinformatics [6] (120 "active" and 554 tested "inactive"). This set is marginally

overlapping with the REAXYS data, sometimes with conflicting activity class assignment.

STRICT, SOFT and ChEMBL were thus considered as three independent training sets, and used for model calibration and/or class landscape coloring, in conjunction with random decoy compounds, assumed as inactive and randomly picked among the non-BRD molecules in ChEMBL.

Screening set

The provided screening set contained 2 million compounds (synthesized at Enamine) encoded in SMILES format.

Virtual screening protocol

In this project the VS protocol included following steps:

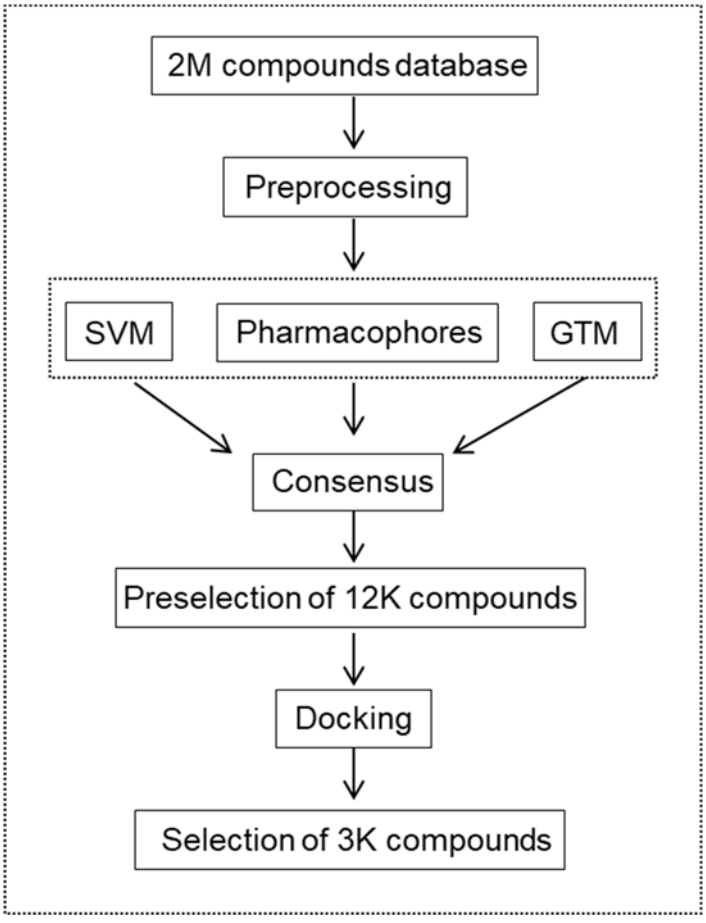

Figure 1: Applied Virtual Screening protocol.

For some of these steps, a dedicated section is presented below.

Compound Standardization and Description

In this project compound standardization followed the default protocol installed on our public web server (infochim.u-strasbg.fr/webserv/VSEngine.html), powered by ChemAxon[17] tools. It includes:

- Dearomatization and final re-aromatization according to the “basic” setup of the ChemAxon procedure (heterocycles like pyridone are not aromatized)
- Removal of salts and mixtures
- Neutralization of all species, except nitrogen (IV)
- Generation of the major tautomer according to ChemAxon

The descriptors used here were ISIDA descriptors computed by ISIDA Fragmentor[18]·[19]. More than 100 different types of descriptors sets were generated. They include sequences, atom pairs, circular fragments and triplet counts of different length, colored by formal charges, pharmacophore features or force field types.

Modeling Methodology

Support Vector Machine

The Support Vector Machine (SVM) is a machine learning method developed by Vapnik[1]. The input variables are mapped into a higher dimensional feature space using a kernel function, and then a linear model is built on this new feature space. The most common kernel functions include linear, polynomial and radial basis functions. The performance of this method depends on type of kernel and a number of parameters. SVM does classification by finding the hyperplane that maximizes the margin between the two classes.

SVM models were validated in 3 fold Cross-Validation (CV) procedure repeated 12 times. These were built on the basis of STRICT and respectively SOFT training sets, each randomly completed with a number of 1221 decoys (i.e. as many decoys as total BRD compounds). Model fitting was performed using the libSVM model optimizer[20], and resulting consensus were posted on the Strasbourg web server. There were thus two SVM consensus models serving for selection.

Table 1: Detailed description of individual SVM models included in the "SOFT", respectively "STRICT" consensus predictors. The table reports the characteristics of descriptor types involved and the model performances (Balanced Accuracy) in cross-validation

Map number	Fragments topology	Informational content	Min/max number of atoms	Atoms labeling	BA_{CV}
SOFT					
1	Sequences	Atoms only	2/8	Force Field(FF)	0.87
2	Sequence	Atoms only	2/7	FF and Formal Charge (FC)	0.87
3	Sequences	Atoms and bonds	2/6	FF	0.87
4	Sequences	Atoms and bonds	2/5	FF and FC	0.86
STRICT					
1	Sequences	Atoms only	2/7	FF and FC	0.73
2	Sequences	Atoms and bonds	2/4	FC	0.75
3	Sequences	Atoms and	2/6	FF	0.74

		bonds			

Generative Topographic Mapping

Generative Topographic Mapping (GTM) is a non-linear mapping method used for data visualization originally described by Bishop. In GTM (Figure 2), a 2D latent space (called manifold) is embedded into the descriptor space. The points which are close in the latent space remain neighbors in the data space. The manifold represents a grid of k x k nodes; each node is mapped in the initial descriptor space using the mapping function y(x, W). The mapping function is given as a grid of m x m radial basis functions (RBF). In order to build a GTM-based QSAR model, the weighted average of properties of all molecules associated with any particular node is used to "color" the manifold according to that property. Here, the projected property is activity class membership, resulting into a fuzzy activity landscape. Molecule "responsibilities" are used as weights. Red and blue zones are only populated by active and inactive compounds, respectively; all colors in between correspond to the regions occupied by compounds of both classes in different proportions. White zones represent unpopulated areas.

GTM activity class landscapes are obtained after the "transfer" of the knowledge about the most likely class to be encountered in a given chemical space neighborhood onto the latent grid nodes that represent this neighborhood. The prediction implies locating the candidate into one of these neighborhoods represented by the population of the nodes, therefrom learning the class to which it should be assigned. GTM-driven predictors typically behave like Nearest-Neighbors-based predictors, which includes the support of identification of candidates outside of its applicability domain, i.e. compounds which do not sufficiently resemble to any of the reference compounds in order to allow an extrapolation of their properties in virtue of the similarity principle.

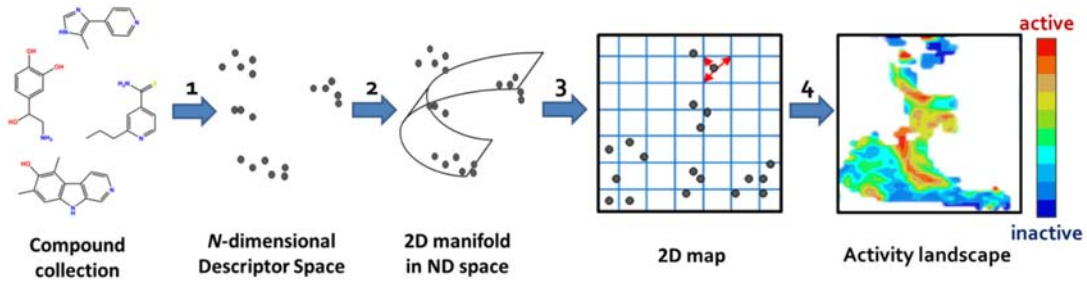

Figure 2: Generative Topographic Mapping

Eight (4 universal[6] and 4 local) maps based on eight distinct ISIDA fragment descriptor spaces were used in this study, as described below:

Table 2: Description of eight maps (4 universal and 4 local), their descriptor types and cross-validated predictive propensity (Balanced Accuracy BA_{CV}) of the two-class classification landscapes colored according to the CHEMBL (universal maps) and SOFT (local maps) activity labels. Some typical fuzzy classification landscapes are illustrated in Figure 3 and Figure 4, respectively.

Map number	Fragments topology	Informational content	Min/max number of atoms	Atoms labeling	BA_{CV}
Universal maps					
1	Sequences	Atoms only	2/3	FF and FC	0.83
2	Atom-centered	Atoms and bonds	1/2	FF	0.82
3	Sequences	Atoms and bonds	2/4	Ph and FC	0.80
4	Sequences	Atoms only	2/7	None	0.84
Local maps					
1	Sequences	Atoms only	2/4	FF	0.87
2	Atom centered	Atoms and bonds	1/3	FC	0.88

3	Normalized Atom centered	Atoms and bonds	1/3	FC	0.86
4	Sequences	Atoms and bonds	2/3	FF and FC	0.84

Note that some of the maps are built on hand of descriptors (detailed atom-centered fragments) capturing connectivity information, whilst other rely on fuzzier atom pair counts, while still others rely on topological pharmacophore descriptors. If the virtually screened library contains compounds from the same chemical series of reference actives (featuring a roughly same scaffold and/or pharmacophore pattern), these will be consensually selected by all the maps and models, irrespective of underlying descriptor space. However, virtual hits may be only partially related to reference actives, so that only the maps able to recognize the specific underlying similarity will be able to retrieve these compounds At one extreme, candidates may be scaffold-hopping analogues of reference compounds, typically not perceived as similar by the human eye. In this case, maps focusing on connectivity-based similarity criteria might exclude such candidates from their AD. Pharmacophore descriptor-based maps will, by contrast, successfully recognize their “matching” pharmacophore patterns. Last but not least, it is important to highlight that similar activity of two compounds does not imply any underlying structural similarity: two actives may have both distinct topologies and distinct pharmacophores, because they bind to different (sub)pockets of the active site. No machine learning technique could infer the activity of the one based on the example of the other – only docking could in principle predict that both are interacting favorably with the site.

Class Landscapes based on Universal Generative Topographic Maps

Universal GTMs were built independently of this work, as “best compromise” maps, able to properly accommodate a maximum of classification landscapes for very diverse biological properties. These GTMs were proven to successfully serve as hosts for 618 classification landscapes associated to the respective target-specific structure-activity

ChEMBL compound series and providing significant separation of actives from inactives. Note that the herein employed, automatically extracted CHEMBL BRD4 training set is one of the above-mentioned 618 targets. More specifically, it did not serve at map building stage, but was one of the external validation sets in that study. Each of the four maps was used to “color” a BRD class landscape according to each of the 3 sets (STRICT, SOFT, CHEMBL) which were supplemented with ChEMBL decoy compounds. For each set, two distinct landscapes were obtained by toggling the Bayesian normalization option on/off (this latter serves to “enhance” the impact of rare actives in the “ocean” of inactives on the landscape). With Bayesian normalization on, 5% of non-BRD ChEMBL molecules were randomly added as decoys (note – different 5% being used for each set). Without normalization, only 1% of the non-BRD ChEMBL compounds were added as decoys. Thus, the combination of 4 maps x 3 sets x 2 normalization options produced 24 distinct “Universal” BRD class landscapes. In order to keep track of individual models, we propose, for each such landscape, the nomenclature scheme UGTM(map number, 1-4)-(BRD training set: SOFT, STRICT, CHEMBL)-DEC(decoy set ID)-BN(Bayesian normalization toggle on or off)”. For example, UGTM2-SOFT-DEC0-BNon is the landscape based on universal manifold #2 (as labeled in the article describing it), considering the SOFT BRD training set completed with the pool DEC0 of random ChEMBL decoy compounds, and using Bayesian normalization.

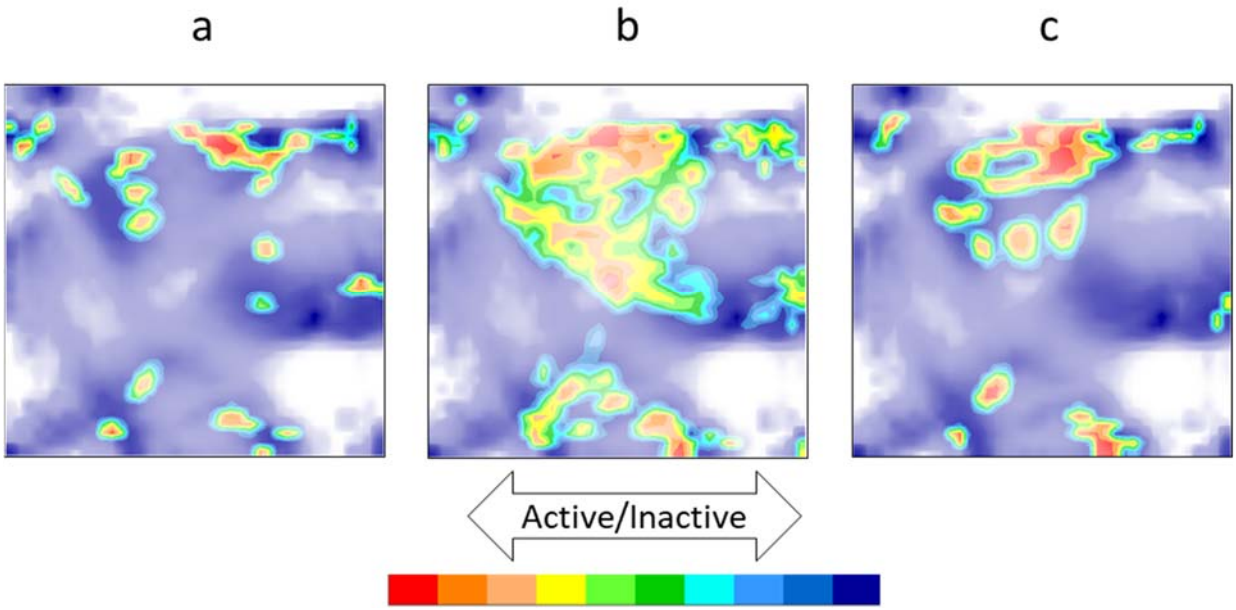

Figure 3: Fuzzy classification landscapes on UGTM1, highlighting zones populates by the actives of the – a) CHEMBL, b) SOFT and c) STRICT training sets, against a common background of the ~1.5M ChEMBL v.23 compounds that were not tested on BRD. These plots apply Bayesian normalization in order to compensate for the extreme imbalance between active and inactive set sizes.

Class Landscapes Based on Dedicated Generative Topographic Maps

Dedicated (or local) GTMs were built with the goal to specifically achieve optimal separation of BRD actives from inactives. For this purpose, the evolutionary map builder procedure used for universal map generation was employed with the key restriction of using as “selection sets” the decoy-enhanced STRICT, SOFT and CHEMBL training sets. Note that while the four universal GTMs work each in a given ISIDA descriptor space – selected independently of this BRD4-related project, the evolutionary optimizer of dedicated maps is free to pick, out of the 100 distinct fragmentation schemes considered, the ISIDA descriptor space(s) that specifically maximize separation of BRD4 actives and inactives. The BRD data sets were each “triplicated” (STRICT1, STRICT2, STRICT3, SOFT1, etc.) by addition of different pools of ~5000 decoys. Four DGTMs with top separating propensities for BRD compounds were retained. For each

of the 4 DGTMs, BRD landscapes were created by coloring with each of the decoy-enhanced triplicates of the three sets, again with and without Bayesian normalization – this gives 4 x 3 x 2 = 72 landscapes based on the dedicated GTMs. The same model nomenclature introduced for universal maps will be used, however using the "DGTM" label for these BRD-dedicated maps.

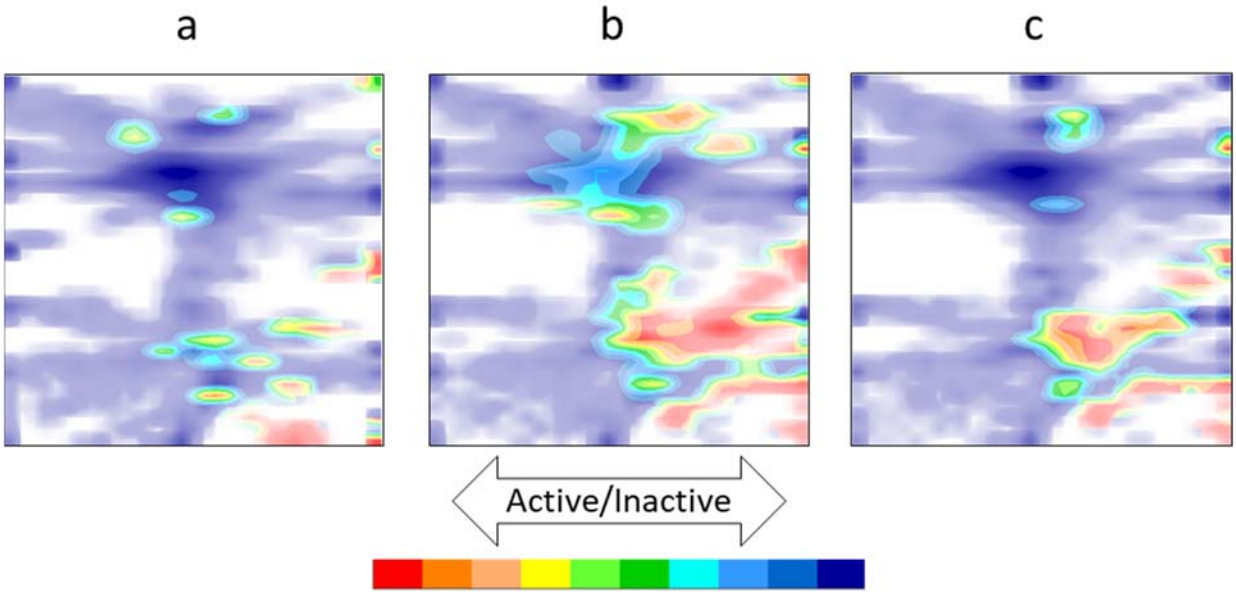

Figure 4: Fuzzy classification landscapes on DGTM2, highlighting zones populates by the actives of the – a) CHEMBL, b) SOFT and c) STRICT training sets, against a common background all the inactive BRD compounds.

Ligand-based pharmacophores

LigandScout[21] was used in the current work.

The models were obtained on the basis of STRICT dataset. The procedure is the following:

1. Generation of conformers of each molecule, with an RMS threshold of 0.5 and energy window 15 kcal/mol, having as maximal number of possible conformers set to 25.

2. Clustering the ligand sets according to the geometry of the 3D pharmacophoric features. Here Pharmacophore radial distribution function was used for similarity calculations. Cluster distance was set to 0.45.
3. Five different pharmacophore hypotheses capable to accommodate actives and discard inactives were considered (see Supporting Information).

Virtual Screening Using QSAR and Pharmacophore models

The Enamine collection of 2M compounds was first submitted to standardization, according to the internal procedure of the Strasbourg web server. The molecular descriptors (ISIDA fragment counts) required for the predictive models were generated. Alternatively, stable conformers were enumerated for the compounds, and submitted to the pharmacophore matching procedure of LigandScout, which allowed ranking of all the 2M candidates by their quality of fit into each of the five pharmacophore models.

Each GTM landscape is a predictive model, since projecting a candidate compound onto it allows to “read” its propensity to be active. Furthermore, GTM projection may explicitly assess the pertinence of each prediction, which is trustworthy if (a) the projected candidate compound is close to the GTM manifold in original descriptor space (it has a “LogLikelihood” criterion similar to the frame compounds used to build the manifold), and (b) if it resides in an area of the map which hosts many compounds from the set used to color the landscape. Both aspects (a) and (b) were used, for each landscape-based prediction, to discard candidates not fulfilling the conditions (technically, they were “ranked” at the bottom of the preference list). The ranking of the other candidates was done according to the propensity to belong to the active class, as read from the landscape.

Similarly, the consensus SVM models also predict the propensity to belong to the active class, and also provide various measures for assessing the applicability of the model to

each candidate. Likewise, candidates within the AD were ranked, for each model, according to predicted propensity to be active, whilst the ones out of AD were ranked as lowest priority.

Thus, each of the 2 SVM models + 24 UGTM landscapes + 72 DGTM landscapes + 5 pharmacophore models proposed their own ranked list of candidates, for the 2 million Enamine compounds. No single molecule was systematically ranked number one by all the approaches. Therefore, a “frequency@TopN” (f@N) empirical scale was established for the final selection: selected compounds are asked to achieve some empirically established minimal frequency of presence within the TopN of some methods, where N was varied. f@N represents the number of models that have simultaneously ranked compound C among the most promising top N. The lower the chosen N, the lower will be f@N. In other words, the notation f@50, for example, means how many models have ranked the compound C in top50. At low N – in particular, for N=1, the event of retrieving the same molecule ranked #1 by many independent models is quite rare. Being ranked #1 by only a few of the different models is a good enough reason to be kept for the final selection. By contrast, being a member of the much broader Top1000 is less “prestigious” – as compensation, membership in Top1000 must be achieved with a significantly higher frequency in order to justify the selection of the compounds.

Selection follows thus a “Pareto” philosophy – some compounds are if some few models give them an excellent ranking, while other are coopted because very many models give them an acceptable ranking. By empirically choosing minimally required thresholds for frequency@TopN values, a pool of 12K compounds was preselected.

This 12K preselection was submitted to docking into the BRD receptor, using the S4MPLE program. This provided an estimation of their binding energy as the final selection criterion. The 3000 best dockers (with lowest calculated binding energies)

were communicated to the Enamine team, in view of experimental assessment of their BRD4 affinity.

Docking

Docking part of the project was done using S4MPLE, (Sampler For Multiple Protein or Ligand Entities) a conformational sampling tool[14]·[15] based on a hybrid genetic algorithm, which allows the simulation of one molecule (conformer generation) or many molecules (docking). Energy calculations were carried out using AMBER force field for biological macromolecules and its generalized version - GAFF for ligands. Here, S4MPLE was used for standard, rigid docking into the active site of the BRD4 structure (PDB code 3MXF), which was assigned standard protonation states for amino acid side chains and then truncated to a sphere of residues with at least one atom within 12 Å from the co-crystallized ligand. Site atoms directly interacting with the ligand were set as “hot spots” for the initial position of ligands by S4MPLE. Ligand processing and docking followed the standard S4MPLE procedure previously described[22] and, like in the cited protocol, the binding energy difference served as final docking score. Before applying S4MPLE to select the 3000 best docking candidates of the 12K pool preselected by the SVM/GTM QSAR models, it was first challenged to dock the REAXYS training set of 1221 actives and inactives, completed with 1221 randomly picked ChEMBL decoy compounds, assumed BRD4-inactives. Figure 5 shows that the BRD4 cavity is mainly hydrophobic formed at one end of the BRD α-helix and the residues of the αZ-αA and αB-αC loops, thus leading to the fact that the nature of binding pocket allows various possible interactions with the ligands.

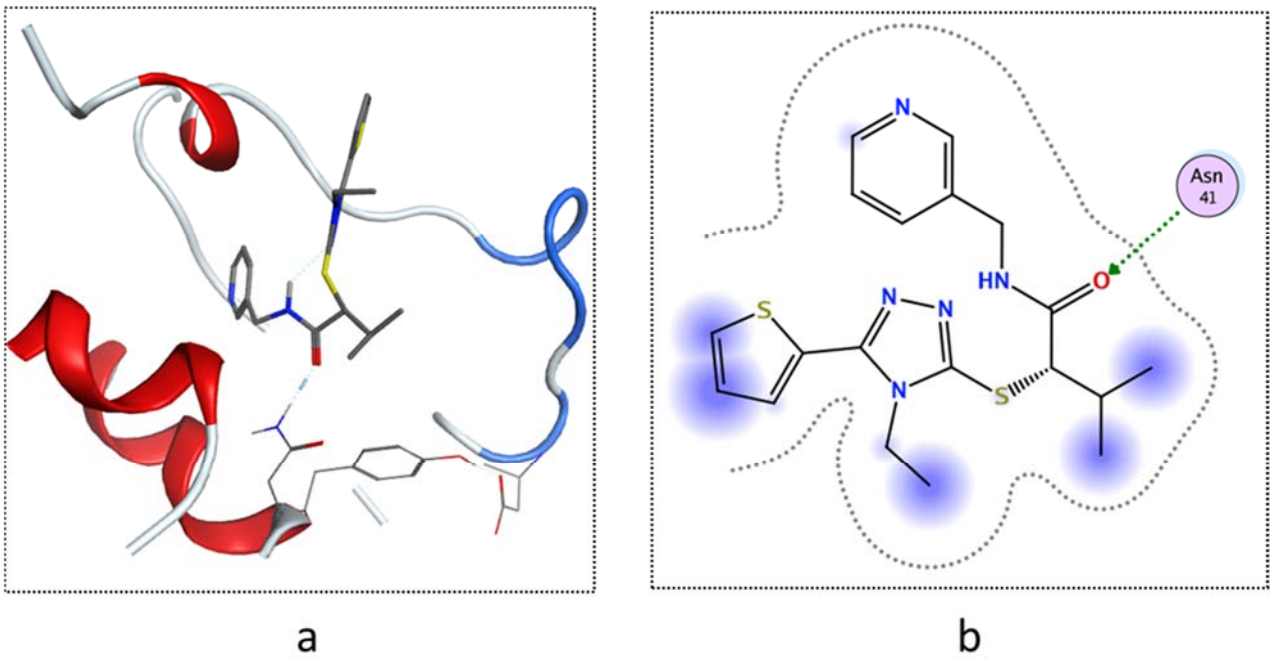

Figure 5: Docking pose (a) of the hit with the lowest IC_{50} value, and associated 2D interaction map according to the MOE[23] software (b).

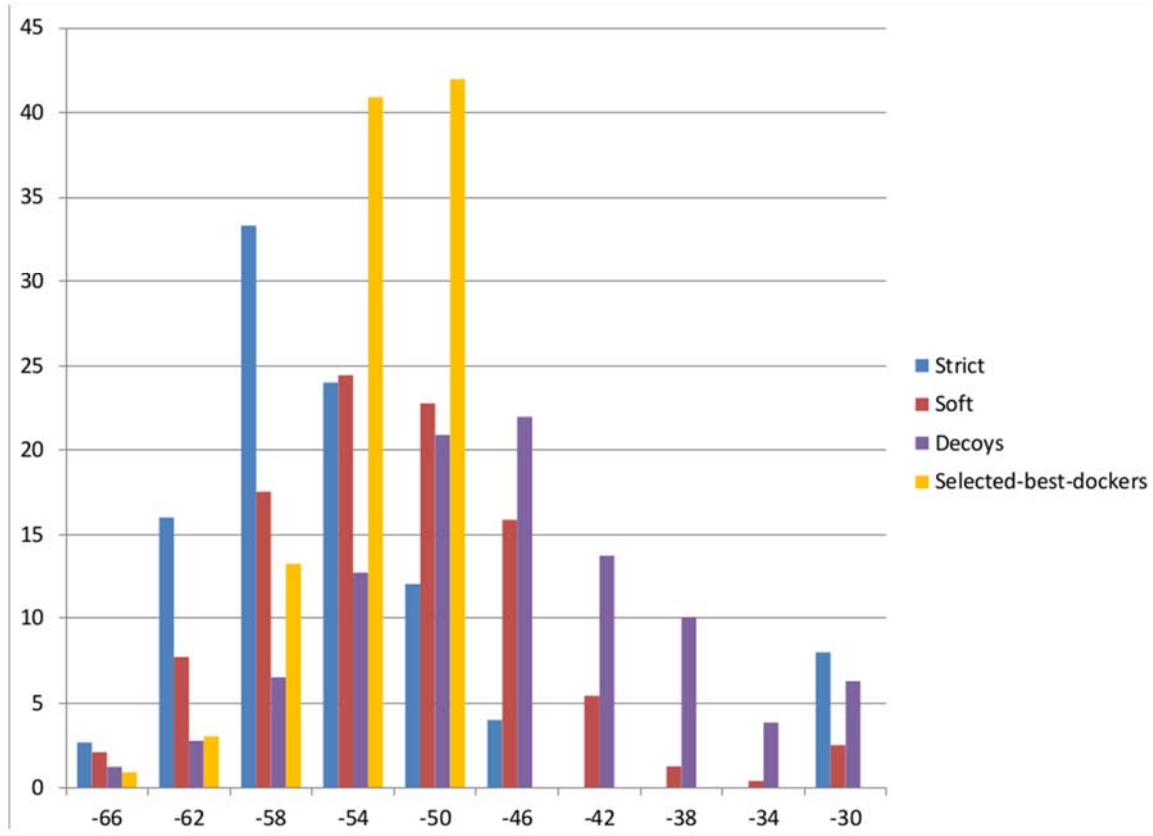

Figure 6: Distribution of docked compounds by binding energy. X-axis: S4MPLE binding energy bins, Y-axis: the percentage of compounds of a given set found to score the given energy.

Experimental testing protocol

Compounds were experimentally tested using DSF, which detects the shift in protein denaturation temperature upon ligand binding as reported by fluorescent dye interacting with protein core exposed by heat denaturation. DSF is a simple, label-free HTS technology applicable to most soluble proteins, irrespectively of their functions and activities. BRD4 sequence fragment corresponds to sequence entry O60885.1 in UniProtKB Database[24]. Represents domain 1 (44-168 AA), contains N-terminus His6-tag and 16-amino acid linker. For details, please refer to previous publications already reporting the use of this experimental protocol[16].

Results

Out of the 3000 selected compounds, 2992 were actually tested and 29 were found to be active. A recently screened random selection of 3200 compounds tested with the same technique gave 0.375% of hits[16], *e.g.* achieved a 2.6 times lower hit rate. VS has thus clearly enhanced the hit rate, albeit a higher enrichment score was expected.

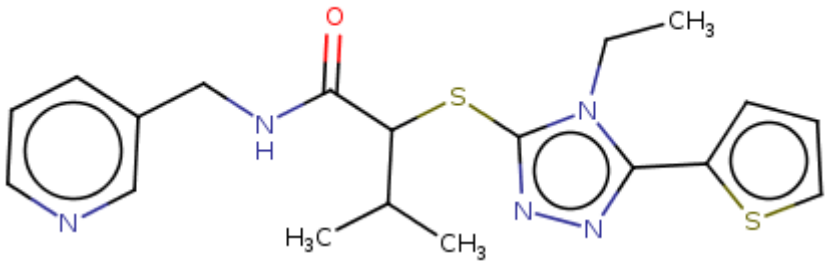

Figure 7: Structure of the hit having the lowest IC_{50} value. Structures and activities of all 29 hits are given in Supplementary Materials.

Structural novelty of discovered hits

In order to assess the originality of the novel hits, they were encoded as ISIDA fragment descriptors, using the three fragmentation schemes that were selected by the DGTM models (see **Erreur ! Source du renvoi introuvable.**). Pairwise Soergel distances (1-Tanimoto similarity) were calculated, in each descriptor space, between the 29 hits and

all the active BRD4 compound present both in SOFT and CHEMBL training sets. Considering the lowest distance in either of the descriptor spaces, 18 of the 29 hits were found to display at least one of the training actives within a neighborhood radius of 0.2. Figure 8 displays the hits closest to training active in the 4th DGTM map descriptor space systematically returning lowest Soergel distances.

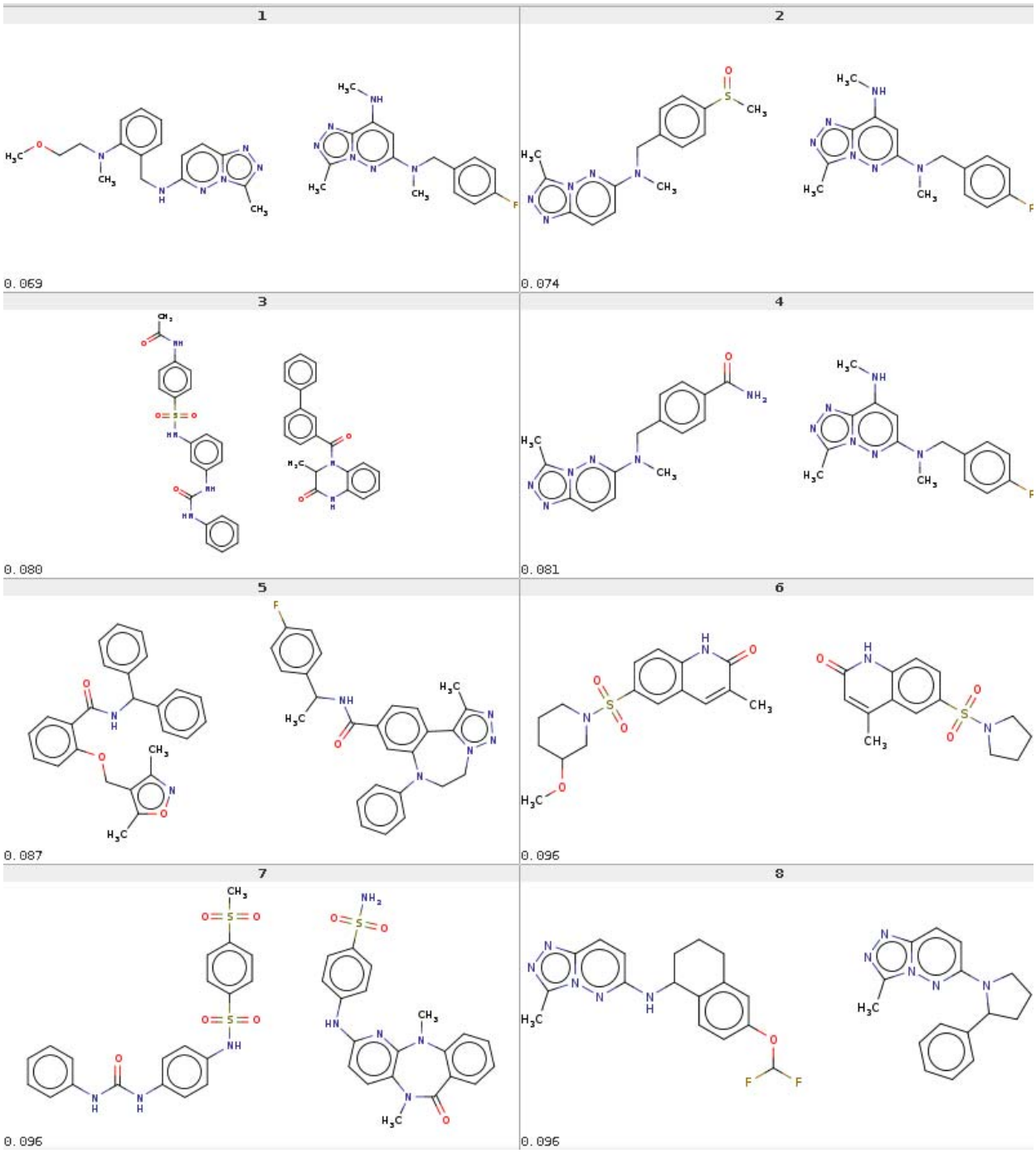

Figure 8: The eight hits (left) closest to training set actives (right), with Soergel distance in the descriptor space of 4th DGTM given below each pair.

The close relationship to known actives is visible – scaffold being often shared, but not always. Even within the eight closest pairs, three examples of “scaffold hopping” are present.

Understanding the reasons of this modest success is a direct opportunity to investigate the strength and pitfalls of this VS strategy based on public data for model training. The two following paragraphs address two key questions:

- How do public-data affinity values that served to build the model relate to the experimental hit detection criterion ΔT_m?
- Which of the specific models were better at prioritizing the 29 discovered hits, and why?

Is DSF-ΔT_m correlated with dose-response affinity measures?

Due to objective constraints, the experimental testing of the selected 3K compounds followed a protocol other than the ones used to characterize the affinity of the training set compounds from the public databases. This requires a better understanding of the degree of correlation. In the absence of strong correlation the training data used may not be relevant with respect to the measured property used to select hits. To this purpose, 39 BRD4-associated compounds in ChEMBL and are furthermore among the compounds in stock at Enamine were also subjected to DSF measurement of ΔT_m at three different concentrations (10, 20 and 40 μM, respectively). 22 of these compounds were present in the CHEMBL training set and have reported IC_{50} or K_i values (the negative log of ChEMBL dose-response affinity value will further on generically be referred to as “pX”). However, none of them qualified for the “active” class as assigned by the automated procedure used to extract ChEMBL structure-activity class sets. For the remaining 17, the ChEMBL records could not be interpreted by the algorithm, so they were not included in the CHEMBL training set at all. Seven of the 22 were also

present in the REAXYS set – five of which were assigned as inactives, and two as moderate actives by the human expert.

For twelve compounds, ChEMBL actually reports *both* IC_{50} and ΔT_m from measurements by their initial discoverers were reported. The magnitudes are weakly correlated, at R^2=0.5 (see Figure 9 a).

If the analysis is extended to the herein measured ΔT_m values *versus* ChEMBL-reported pX data for 22 compounds (Figure 9 b), the strongest correlation is obtained with the ΔT_m values at 10 μM concentration. Note, furthermore, that the average melting temp shift for the 22 compounds with reported pX values was of 0.37 ± 0.33 degrees, whereas the 17 ChEMBL compounds which had no associated pX values were all inactive with respect to ΔT_m: their average shift was of 0.08 ± 0.15 degrees. They would have represented valuable true negatives but were not considered due to the intrinsic limitations of the ChEMBL activity series extraction protocol.

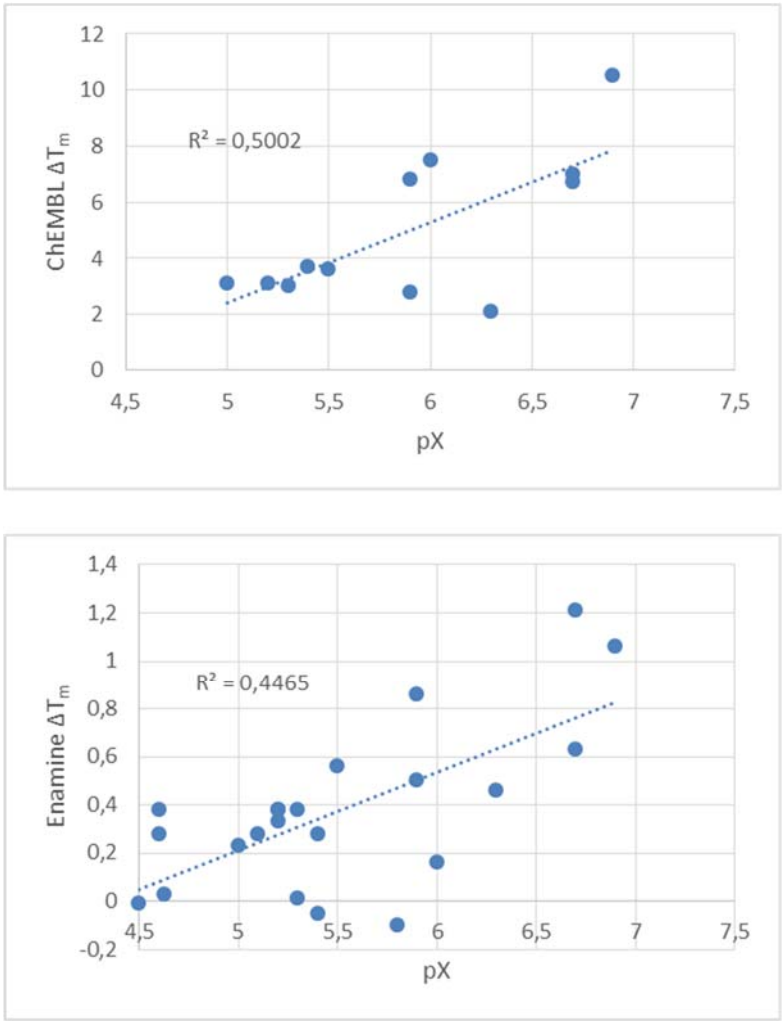

Figure 9: a) Correlation between ChEMBL-reported ΔT_m and ChEMBL-reported dose-response-based affinity measure (pX) for 12 BRD4 inhibitors. b) Correlation between the ΔT_m values measured according to the current experimental screening protocol at Enamine, and ChEMBL-reported dose-response affinity measure (pX) for 22 BRD4 inhibitors from ChEMBL.

Last but not least, the correlation (Figure 10 a) between ChEMBL ΔT_m and Enamine ΔT_m has been determined for 13 compounds and it turns out that the Enamine measure at 40 µM is the one best correlating with the ChEMBL data. The Enamine measure at 10 µM, the one that best correlated the ChEMBL-pX, is significantly less well related to ChEMBL ΔT_m values (R^2~0.4).

Eventually, for nine of the herein obtained hits, a FRET-based estimation of their IC_{50} values was experimentally undertaken. As seen in Figure 10 b, these results are completely uncorrelated with the reported ΔT_m values.

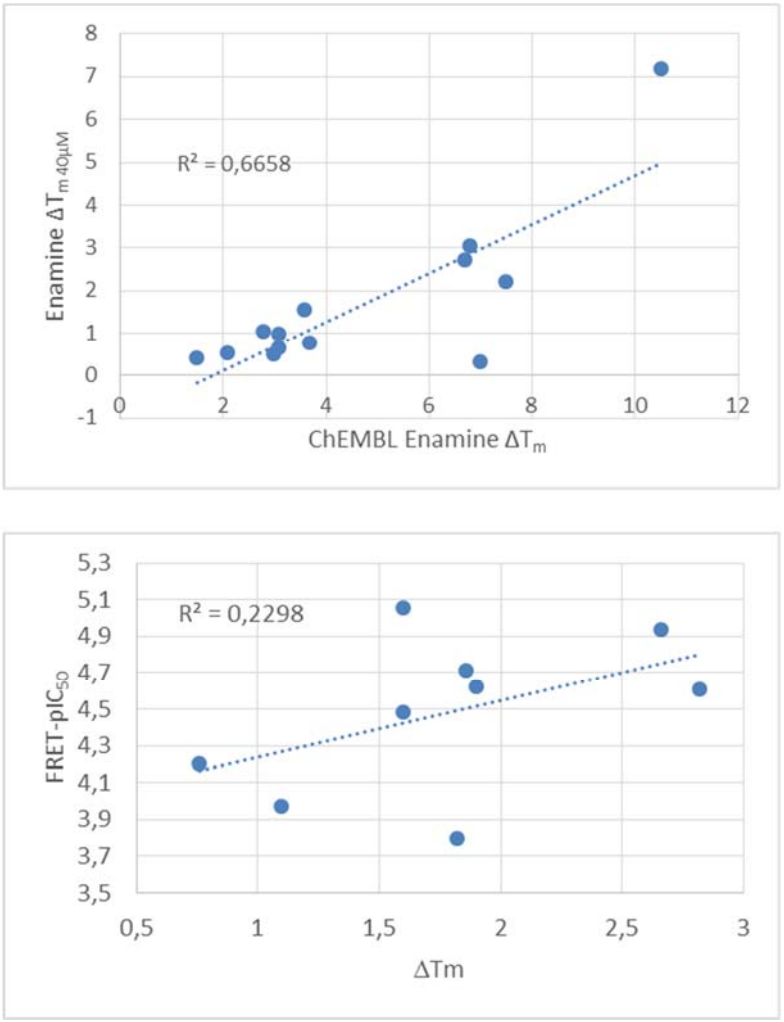

Figure 10: a) Correlation between ChEMBL-reported and Enamine re-measured ΔT_m values, for ChEMBL BRD4-compounds. b) Correlation between FRET-based IC_{50} values estimated at Enamine for 7 of the newly discovered hits and their ΔT_m values in primary screening.

The above discussion shows clearly that the exploitation of public databases obliges the user to face a wide spectrum of heterogeneous activity indices, which may or may not be “compatible” with the setups of the in-house experimental protocols for hit discovery. Clearly, the ChEMBL text mining protocol that assigned active/inactive labels to the BRD-associated compounds was successfully used for hundreds of other targets and returned modelable structure-activity sets. However, it specifically focused on molecules with reported dose-response activity measures (all while ignoring their exact nature – no distinction between K_i and IC_{50} values was made). It was not considering DSF-ΔT_m values. Activity classification based on such values is heavily target-specific, thus there is no simple threshold to be provided to a general data mining protocol. Text mining

algorithms cannot cope with the subtleties of biological testing. The alternative of considering separate compound sets published by a same source using a same testing protocol does not solve the problem, but simply produces many disjoint small series, of no use in QSAR training. Also note that falling back to binary classification models – be it either by automated, algorithmic, or by expert hand-made choice of the activity threshold – is *per se* a source of information loss. The original dose-response activity scores from public sources were shown to be *per se* rather poorly correlated to the hit selection criterion DSF-ΔT_m. The fact that they were not used as such for model training, but first underwent conversion into a categorical variable has most likely had a negative impact on model performance as well.

A posteriori analysis of the ability of individual models to prioritize discovered hits

Seventeen of the 29 hits have been ranked #1 by at least one of the dedicated GTM landscapes (notably DGTM2, but also DGTM1 and DGTM4) while two were ranked first by SVM. Note that “ranked #1” practically means that these compounds were given the maximum likelihood to be active, *ex aequo* with (often numerous) other candidates. None of the hits was seen to clearly outperform all the other 2M candidates of the Enamine library in terms of predicted likelihood of activity. More precisely, no model selected a single compound ranked #1. The other twelve hits entered selection because of consensual ranking within top IM by several models.

In order to gain a better insight of the individual models that would have preferentially top ranked the 29 confirmed hits within the pool of selected 3K compounds, the plot (Figure 11 a) was realized by monitoring, for each model, the minimal number of better-or-equally ranked compounds it would have had to select in order to “discover” *H* hits. Let *M(H)* denote the minimal size of the subset of top-ranked compounds by the model *M* that include *H*=1…29 of the confirmed hits. The percentage of hits in such selection,

H/M(H)×100 has been plotted against *H*, for all considered QSAR, pharmacophore and docking models. For each *H* value (*H≥10* shown in Figure 11), there will be one "winning" model which managed to regroup *H* hits within the smallest *M(H)*, *i.e.* provided the best possible ranking for the *H* hits. A Pareto front of "dominating" configurations (hit number, hit percentage) can be drawn, all methods confounded. This Pareto front is mainly contributed by two methods: DGTM2-SOFT-DEC1-BNoff (Dedicated GTM2 landscape "colored" by the SOFT training set completed with decoy pool 1, without Bayesian normalization) and the SVM model trained on SOFT. Out of the tested 3K library, eleven of the 29 hits are found within the 106 top-ranked compounds DGTM2-SOFT-DEC1-BNoff, which represents the highest hit density (10.38%) that was achieved by any of the methods, all while retrieving a significant number of discovered hits. This represents the ten-fold of the hit rate of the current experiment, and the 26-fold of the one achieved in random screening[16]. The SVM model is an equivalent potent solution, favoring however the retrieval of more hits at a lesser hit rate.

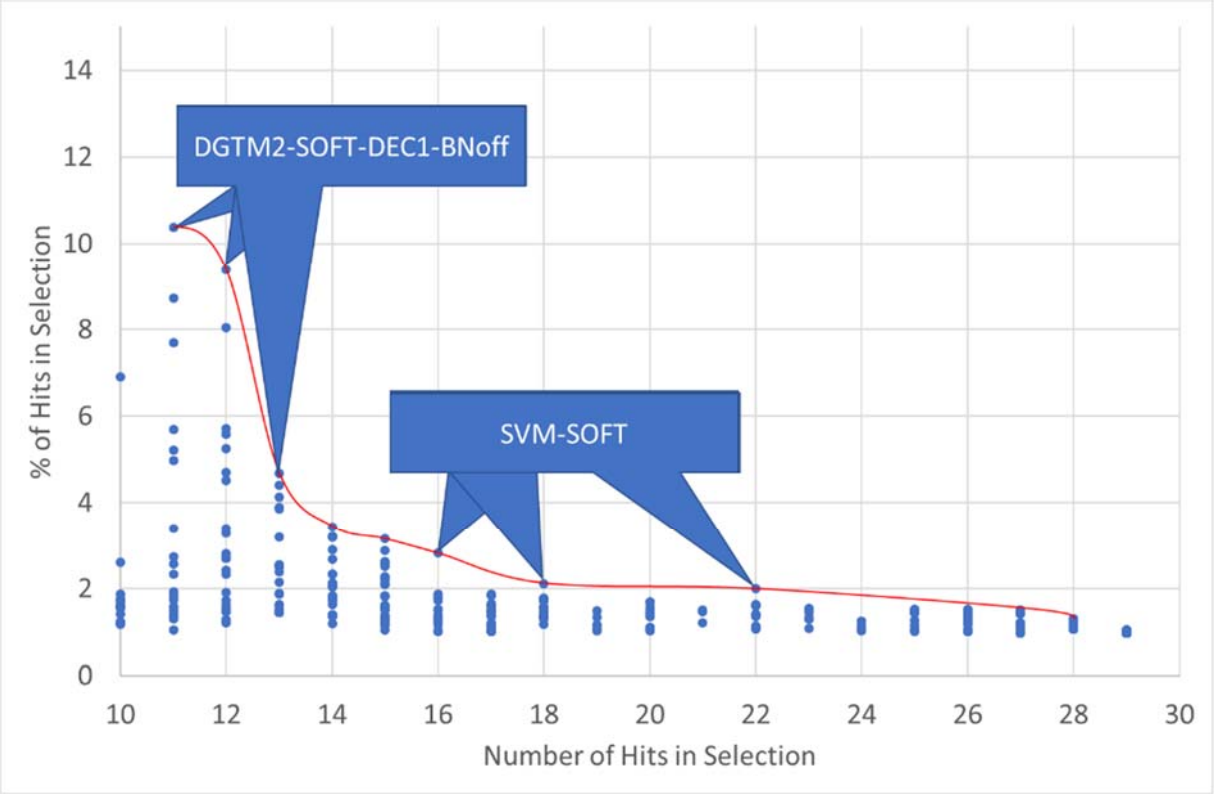

Figure 11: The number of hits H (out of the 29 discovered) found within a minimal subset of M(H) compounds top-ranked by a model M, *versus* the percentage they represent within this subset. The red "Pareto front" regroups models returning the most hit-rich subsets containing H of the 29 hits.

It is instructive to construct such Pareto fronts not only for the entire battery of models, but also for specific subsets of models. The Pareto front of all the DGTM2-based models would consist of the best (hit number, hit rate) combinations scored by either of the landscapes based on the DGTM2-manifold, irrespective of training set, added decoys or Bayesian normalization strategy. The following plot illustrates the fronts constructed for the best performers amongst the QSAR models, which were able to reach or exceed 2% of hit rate (2-fold enrichment over the global hit density, and 5.2-fold enrichment over random VS).

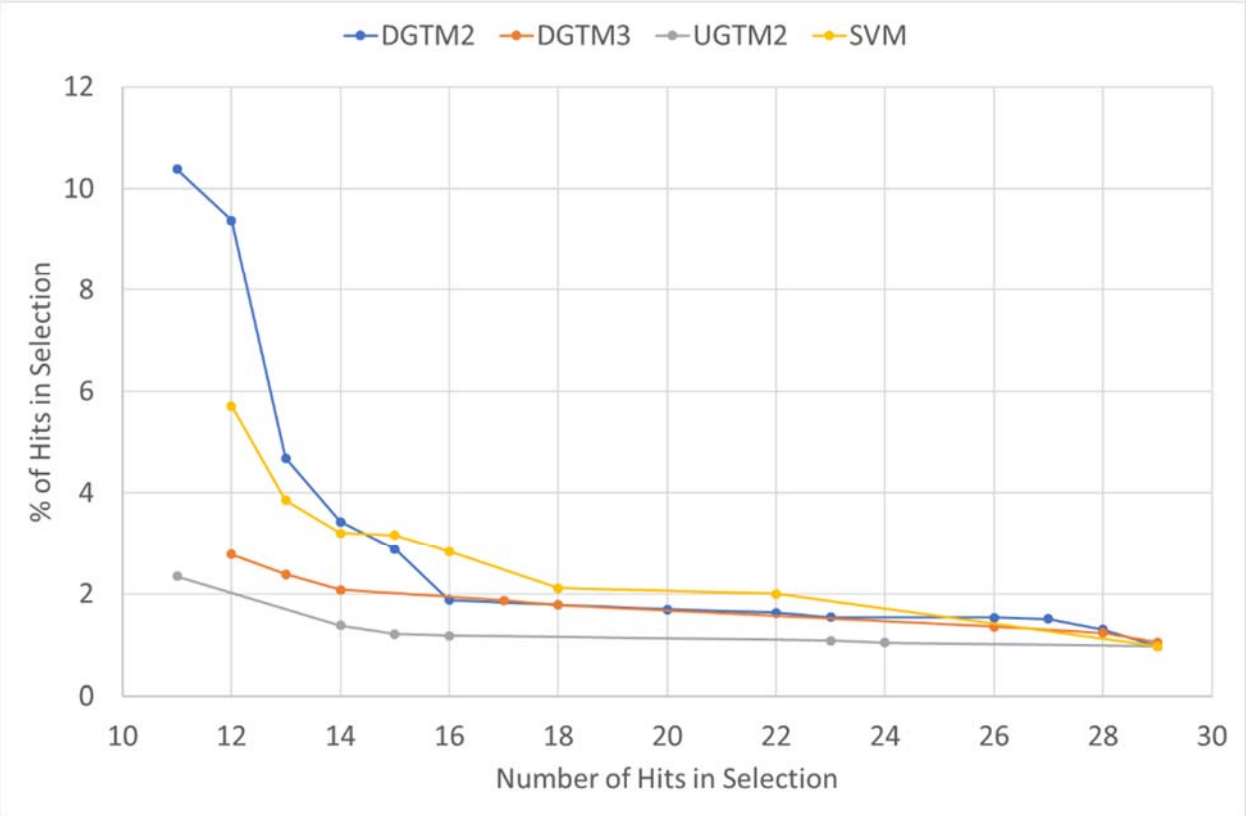

Figure 12: The most significant (hit number, hit rate) Pareto fronts associated with the four QSAR approaches DGTM2, SVM, DGTM3, UGTM2.

These also include, in addition to already highlighted DGTM2 and SVM-based models, another dedicated GTM (DGTM3) and a Universal map (UGTM2). All classes of 2D-QSAR models have at least one representative that would have been in principle able to significantly enhance the hit rate beyond the achieved 1% – but not pharmacophore models, nor docking. In terms of docking scores these 29 compounds do not stand out in any way, compared to the rest of the tested 3K library – the mean of their docking scores perfectly matches the mean over the 3K set. Nevertheless, S4MPLE was proven to effortlessly discriminate between the BRD actives *versus* inactives and decoys of the STRICT set, with a ROC AUC of 0.77. When the compounds of intermediate potency are counted as active, in the SOFT training set, the S4MPLE ROC AUC value decreases to 0.66, *i.e.* remains well above random selection level. The clear separation of strong, medium actives and respectively inactives in terms of S4MPLE binding energy differences is visible in the histogram (Figure 6).

Independently of this, S4MPLE has been successfully used in fragment-based drug design of novel BRD4 inhibitors[25]. However, the weak correlation between DSF-ΔT_m and actual IC_{50} values is certainly a significant reason for which the 29 selected hits are not “special” in terms of docking scores: the seven hits for which FRET-based IC_{50} values were actually measured are μM at best and all would have qualified as “inactive” according to the STRICT class assignment criteria, while some would have been qualified “inactive” even by the more lenient SOFT criteria. Clearly, the DSF-ΔT_m measurement protocol is useful for primary screening and is meant to select hits that are just potent enough to serve as a departure point in hit to lead optimization. It would not specifically single out very potent (but very rare) nM binders, which are unlikely to be discovered as such in primary screens. Or, discrimination by docking between weak binders and non-binders is notoriously difficult. Using the STRICT set for machine learning in general turned out to be a poor working hypothesis: the remarkable SVM model Pareto front is completely contributed by the SOFT set-trained SVM model, while the STRICT alternative provides no prioritization at all for the 29 hits. The same applies for the CHEMBL set, which also featured mostly sub-μM compounds as actives, its

usage in both DGTM and UGTM landscape coloring would not have led to preferential selection of the 29 hits. By contrast to the highly diverse set of medium-potency inhibitors, the highly active BRD4 compounds in STRICT are fundamentally based on three key scaffolds: a seven-membered N heterocycle fused to phenyl, a pyrrole/imidazole ring fused to a saturated 5-membered lactam ring, and respectively benzimidazoles and derivatives featuring additional heterocyclic N atoms (Figure 13 a, b and c respectively). Other molecules (e.g., a macrocycle, spiro derivatives, or a diazo derivative) are basically singletons, *i.e.* too rare to allow machine-learning of their underlying structural patterns. Apparently, the dominating patterns are not well represented in the Enamine candidate collection submitted to this VS experiment.

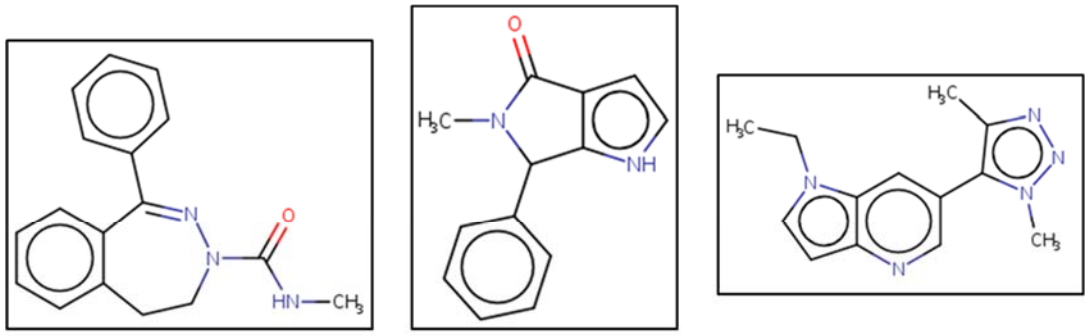

Figure 13: Three key scaffolds of STRICT dataset.

Figure 14 displays the specific behavior of DGTM2 landscapes based on the SOFT training set and without Bayesian normalization, as a function of the added pools of decoys. Interestingly, the maximal hit enrichment is seen to substantially depend on the randomly included decoys (here, 1% of ChEMBL compounds, excluding the BRD-tested structures). Decoy pools are thus rather large compound collections, of sizes around 15K. However, the obtained class landscapes might behave significantly different with respect to the relative ranking of the 29 hits (the size of the compound set needed to encompass 11 hits actually doubles when the decoy pool 0 is used instead of pool 1).

"Noise" from the decoy pools notwithstanding, these landscapes remain amongst the models that are significantly prioritizing the discovered hits.

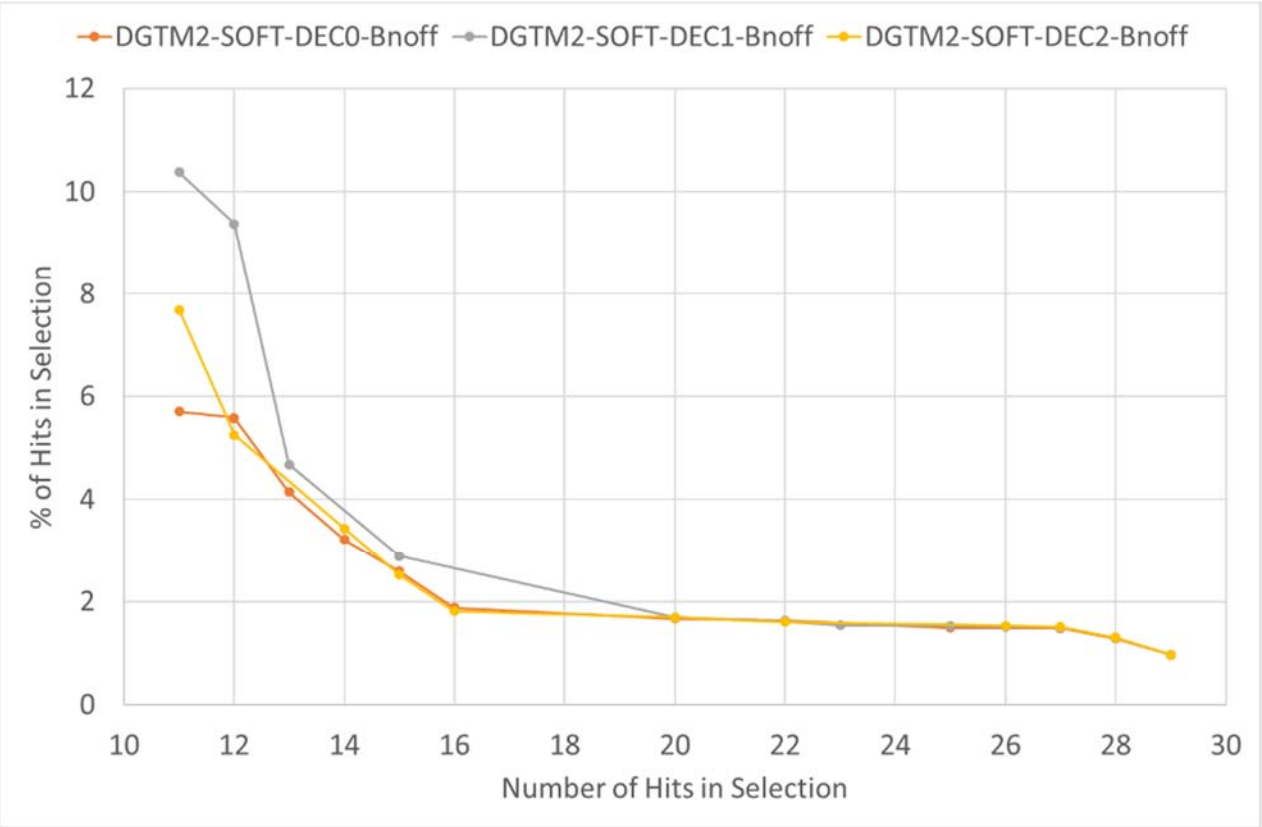

Figure 14: (Hit number, Hit percentage) Pareto fronts of DTGTM2 landscapes as a function of the used random ChEMBL decoy set.

No meaningful selection of 3K compounds could be achieved, making the addition of random ChEMBL decoys a necessity. In absence of decoys both GTM and SVM models displayed unexpectedly high propensities to rank the Enamine candidates as "active" – a tendency most marked with the SOFT training set, and also with the slightly more specific CHEMBL. The underlying reasons can be understood by inspecting fuzzy class landscapes – Figure 15 shows the separation of BRD inactives (red) from actives (blue) in the (decoy-free) SOFT set, on UGTM1 and UGTM2, respectively. The (moderately) actives reported are – as expected from public databases compiling many sources – structurally quite diverse. Irrespective of the underlying descriptor space (ISIDA Force-field-colored atom sequence counts for UGTM1 *versus* force-field-colored

circular atom counts for UGTM2), the SOFT “actives” cover a very large area of the SOFT-populated GTM landscapes (the majority of it, in case of UGTM1).

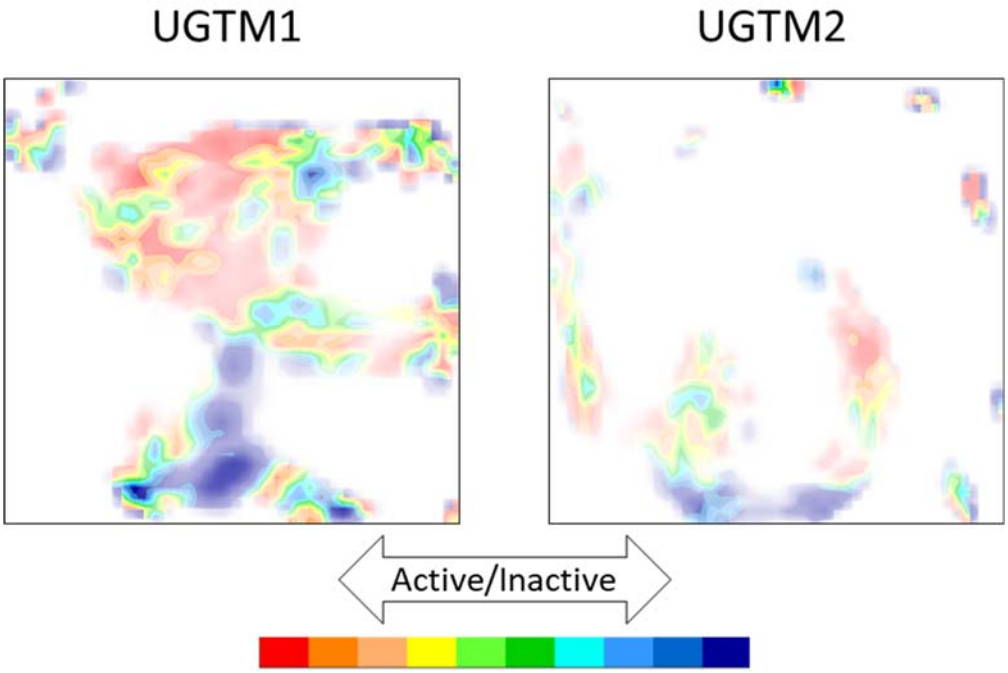

Figure 15: Fuzzy classification landscapes showing the separation of actives and inactives as defined in the SOFT training set, on two Universal maps. These landscapes do not use Bayesian normalization – a red color means that SOFT actives are the absolute majority of residents in those areas.

According to the perception of chemical diversity supported by UGTM1 descriptors, the SOFT subset of actives is actually more diverse than the inactive – and this perception of chemical diversity cannot be dismissed as irrelevant, because it supports robust separation of actives from inactives for > 600 target-specific compound series, including BRD4 (for the SOFT set, the cross-validated balanced accuracy of separation is of 0.78).

Note that “visual” monitoring of diversity by the areas covered on the map, as illustrated above, may appear less rigorous than some quantitative measure – like the count of “clusters” that may be obtained by a classical algorithm. In practice, this is not the case – first, because the outcome of such a clustering algorithm may widely fluctuate in

response of chosen descriptor set, dissimilarity metric, clustering algorithm and (algorithm-dependent) clustering thresholds, *etc.* Again, such hyperparameters were, in case of herein used GTMs, chosen as a result of a quantitative optimization of some predictive power propensity of the model. The density patterns seen on the map are thus representative of chemical diversity of a compound library and may even be quantitively expressed – by the entropy score. Such values are however of relevance for large library comparison and will not be reported here.

The addition of decoys is indeed arguable – some of these decoys might actually be yet untested actives – but is nevertheless needed to "reclaim" some of the chemical space dominated by SOFT actives due to the fact that the training set as such fails to include such examples.

Conclusions

While the "Big Data" label may apply to public structure-activity databases as a whole, the specific target-related data needed for predictive model building in view of virtual screening of electronic compound databases is unfortunately rather sparse and heterogeneous. The VS study presented herein aimed at detecting novel BRD4 binders and relied on knowledge from public databases (ChEMBL, REAXYS) to establish a battery of predictive models used to virtually screen a collection of 2M compounds from Enamine. This industrial partner then experimentally screened a subset of 3K (2992) molecules selected by the VS procedure for their high likelihood to be active. Previous work at Enamine – random selection and screening, by the strictly identical Differential Scanning Fluorimetry protocol – of an equal-sized library drawn out of the same initial collection[16] – presented an excellent reference to estimate the benefit of VS in terms of hit rate enhancement. Twenty nine confirmed hits were detected alter experimental testing, representing 1% of the 3K selected candidates. While this hit rate is a robust 2.6

times superior to the hit rate found in random screening under identical conditions, it is, on the absolute, rather disappointing for a “Big Data”-driven VS experiment, in which every single model (including docking) has been thoroughly (cross-)validated with respect to public BRD4 data. This prompted us to an in-depth investigation of the quality of training data, and its compatibility with the in-house hit selection criterion, DSF-ΔT_m. On one hand, it was shown that the heterogeneous public data cannot be fused into a single, rigorously defined training set. Specific dose-response-based activity values reported by authors contributing to the public databases cannot be merged and are only weakly correlated with DSF-ΔT_m assays. Therefore, active/inactive classification models were the only option, and attribution of the “active” label to public database ligands is a highly empirical, arguable undertaking. On one hand, ChEMBL compounds were extracted and classified by a simple text mining procedure developed for previous studies. By contrast, REAXYS-extracted compounds were classified according to IC_{50} thresholds and, as it was impossible to know beforehand what threshold will lead to the most predictive models, two distinct hypotheses were pursued in parallel, leading to the alternative STRICT and SOFT training sets.

Retrospectively, the SOFT training set produced models that provided the best rankings for the discovered hits. The highly active compounds exclusively labeled as “active” in the STRICT set mostly represent congeneric series based on common scaffolds completed with several singletons that cannot be exploited by machine learning. In the SOFT set, adding the moderately actives to the “active” class leads to the opposite scenario where “actives” seem to dominate a significant (even majority) of the training chemical space. In absence of decoys – random ChEMBL compounds that were never associated to the BRD receptors – SOFT-based models tend to overestimate the likelihood to be active. Adding 1% to 5% of ChEMBL compounds as decoys counterbalances the artificial dominance of SOFT actives in the chemical space. Given

the overall low to moderate correlation between dose-response activity values that are at the basis of training set definition and the DSF-ΔT_m criterion used to select hits, the decoy-enhanced SOFT training set worked better in conjunction with a screening method focused on discovery of moderate actives, *i.e.* typical primary hits. ChEMBL is a good source of decoys – within the intrinsic limitations of this approach. The random-drawn decoy subsets were seen to have a visible impact on the rankings returned by the GTM landscapes. STRICT and CHEMBL sets put more weight on the high potency of active examples, herewith limiting the number and diversity of actives to be "learned" by models.

VS selection was based on ranking each of the 2M candidates by their likelihood to be active, according to each model. These included SOFT, STRICT and CHEMBL-trained SVM, GTM (featuring both Universal and BRD4-Dedicated manifolds) and LigandScout pharmacophore models. A consensus scheme was employed to select 12K candidates – either top-ranked by at least one model, or ranked within the top N candidates by at least M models (several empirical N,M pairs were used – with more models M required as the required top rank N is relaxed). Docking with S4MPLE, which showed robust ROC AUC separation of training set actives, and independently served for successful fragment-based design of BRD inhibitors – was used to pick the 3K best-docked of the 12K selected candidates. These were experimentally screened by the Enamine partner, with already mentioned results.

As a general conclusion, this study emphasizes that public structure-activity databases are nowadays key players in drug discovery. Their limits are set by the state-of-the-art knowledge harvested so far by published studies, and these limits can be very stringent. Data heterogeneity makes it extremely difficult to exploit in rational drug design. Furthermore, published activity values may not be easily comparable or may not be correlated with values from other assays. In spite of this, a robust 2.6-fold increase of

the hit rate with respect to an equivalent, random screening campaign showed that machine learning is able to enrich active compounds in selection sets.

Supporting Information

Five used pharmacophore models and list of structures of hits with associated IDs, experimental ΔT_m and IC_{50} are provided

Acknowledgment

Iuri Casciuc thanks the Région Grand Est for a PhD fellowship.

Bibliography

[1] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.

[2] C.M. Bishop, M. Svensén, C.K.I. Williams, GTM: The generative topographic mapping, Neural Comput. 10 (1998) 215–234.

[3] T. Kohonen, The self-organizing map, Proc. IEEE. 78 (1990) 1464–1480.

[4] N. Kireeva, I.I. Baskin, H.A. Gaspar, D. Horvath, G. Marcou, A. Varnek, Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison, Mol. Inform. 31 (2012) 301–312.

[5] H.A. Gaspar, I.I. Baskin, G. Marcou, D. Horvath, A. Varnek, GTM-Based QSAR Models and Their Applicability Domains, Mol. Inform. 34 (2015) 348–356.

[6] P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath, Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds, J. Comput. Aided. Mol. Des. 29 (2015) 1087–1108.

[7] P. Sidorov, E. Davioud-Charvet, G. Marcou, D. Horvath, A. Varnek, AntiMalarial Mode of Action (AMMA) Database: Data Selection, Verification and Chemical Space Analysis, Mol. Inform. 37 (2018) 1800021.

[8] Reaxys database, (2017). https://www.reaxys.com.

[9] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2011) D1100--D1107.

[10] J.W. Tamkun, R. Deuring, M.P. Scott, M. Kissinger, A.M. Pattatucci, T.C. Kaufman, J.A. Kennison, brahma: a regulator of Drosophila homeotic genes structurally related to the yeast transcriptional activator SNF2SWI2, Cell. 68 (1992) 561–572.

[11] S.-Y. Wu, C.-M. Chiang, The double bromodomain-containing chromatin adaptor Brd4 and transcriptional regulation, J. Biol. Chem. 282 (2007) 13141–13145.

[12] V. Brès, S.M. Yoh, K.A. Jones, The multi-tasking P-TEFb complex, Curr. Opin. Cell Biol. 20 (2008) 334–340.

[13] J. Morinière, S. Rousseaux, U. Steuerwald, M. Soler-López, S. Curtet, A.-L. Vitte, J. Govin, J. Gaucher, K. Sadoul, D.J. Hart, others, Cooperative binding of two acetylation marks on a histone tail by a single bromodomain, Nature. 461 (2009) 664.

[14] L. Hoffer, D. Horvath, S4MPLE--Sampler For Multiple Protein--Ligand Entities: simultaneous docking of several entities, J. Chem. Inf. Model. 53 (2012) 88–102.

[15] L. Hoffer, C. Chira, G. Marcou, A. Varnek, D. Horvath, S4MPLE-sampler for multiple protein-ligand entities: methodology and rigid-site docking benchmarking, Molecules. 20 (2015) 8997–9028.

[16] P. Borysko, Y.S. Moroz, O. V Vasylchenko, V. V Hurmach, A. Starodubtseva, N. Stefanishena, K. Nesteruk, S. Zozulya, I.S. Kondratov, O.O. Grygorenko, Straightforward hit identification approach in fragment-based discovery of

bromodomain-containing protein 4 (BRD4) inhibitors, Bioorg. Med. Chem. (2018).

[17] ChemAxon, Standardizer, C version 5.12, (2012).

[18] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, ISIDA Property-Labelled Fragment Descriptors, Mol. Inform. 29 (2010) 855–868.

[19] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. V Tetko, G. Marcou, ISIDA-Platform for virtual screening based on fragment and pharmacophoric descriptors, Curr. Comput. Aided. Drug Des. 4 (2008) 191.

[20] D. Horvath, J.B. Brown, G. Marcou, A. Varnek, An evolutionary optimizer of libsvm models, Challenges. 5 (2014) 450–472.

[21] Gerhard Wolber and Inte:Ligand GmbH, LigandScout 4.1, (2017). http://www.inteligand.com/ligandscout/.

[22] M. Zhenin, M.S. Bahia, G. Marcou, A. Varnek, H. Senderowitz, D. Horvath, Rescoring of docking poses under Occam's Razor: are there simpler solutions?, J. Comput. Aided. Mol. Des. (2018) 1–12.

[23] H. 2R7 Chemical Computing Group Inc., 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, Molecular Operating Environment (MOE), 2016.08, (2016).

[24] T.U. Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45 (2017) D158–D169. doi:10.1093/nar/gkw1099.

[25] L. Hoffer, C. Muller, P. Roche, X. Morelli, Chemistry-Driven Hit-To-Lead Optimization Guided by Structure-Based Approaches, Mol. Inform. (2018).