

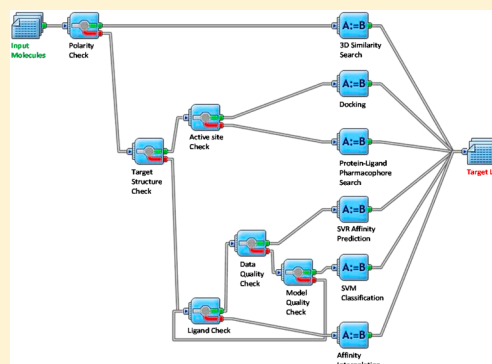
Computational Profiling of Bioactive Compounds Using a Target-Dependent Composite Workflow

Jamel Meslamani,[†] Ricky Bhajun,[‡] Francois Martz, and Didier Rognan*

Laboratory for Therapeutical Innovation, UMR 7200 Université de Strasbourg/CNRS, MEDALIS Drug Discovery Center, F-67400 Illkirch, France

S Supporting Information

ABSTRACT: Computational target fishing is a chemoinformatic method aimed at determining main and secondary targets of bioactive compounds in order to explain their mechanism of action, anticipate potential side effects, or repurpose existing drugs for novel therapeutic indications. Many existing successes in this area have been based on a use of a single computational method to estimate potentially new target–ligand associations. We herewith present an automated workflow using several methods to optimally browse target–ligand space according to existing knowledge on either ligand and target space under investigation. The protocol uses four ligand-based (SVM classification, SVR affinity prediction, nearest neighbors interpolation, shape similarity) and two structure-based approaches (docking, protein–ligand pharmacophore match) in series, according to well-defined ligand and target property checks. The workflow was remarkably accurate (72%) in identifying the main target of 189 clinical candidates and proposed two novel off-targets which could be experimentally validated. Rolofylline, an adenosine A₁ receptor antagonist, was confirmed to inhibit phosphodiesterase 5 with a moderate affinity ($IC_{50} = 13.8 \mu M$). More interestingly, we describe a strong binding ($IC_{50} = 142 \text{ nM}$) of a claimed selective phosphodiesterase 10 A inhibitor (PF-2545920) with the cysteinyl leukotriene type 1 G protein-coupled receptor.



INTRODUCTION

Predicting the most likely macromolecular targets of a bioactive compound (computational bioactivity profiling) is getting increasing interest in early drug discovery for many reasons. First, phenotypic screens are nowadays the primary source of first-in-class low molecular weight drugs.¹ Since phenotype to target identification is only successful in ca. 40% of the cases,² computational bioactivity profiling is a fast and cheap alternative to experimental target identification approaches. Second, the early identification of possible off-targets is an important preclinical safety asset to avoid potential side effects and severe adverse reactions.³ Third, the competitive advantage in targeting several a priori unrelated targets for certain therapeutic indications (e.g., cancers, Alzheimer disease, schizophrenia) has boosted the design of polypharmacological ligands with a controlled selectivity profile.⁴ Last, the easiest route to a novel drug being to start from an old one,⁵ there is an increasing pressure to repurpose clinical candidates or marketed drugs toward novel indications and consequently novel targets, while sparing the long and costly initial clinical trials.

There exists at least three computational methods to predict potential targets from a given ligand.⁶ At first come ligand-based approaches assuming that similar ligands should recognize similar targets. Having a set of target-annotated compounds in hand, any novel ligand similar enough to a database compound is likely to share the same target. The method therefore just requires a chemical descriptor to describe compounds, a

similarity metric to compare them, and a statistical background to rank the corresponding targets. The best representative of ligand-centric methods is the similarity ensemble approach (SEA) developed by Shoichet and co-workers.⁷ In SEA, target-specific ligand sets are iteratively compared to a bait using simple 2D circular fingerprints. Targets are retained if their ligands are statistically more similar to the query than that expected by a random screen of a similar-sized set and ranked by an expectation value. The method has been applied to several prospective cases and proved successful with accuracy close to 50% to identify main and off-targets for several drugs and bioactive compounds.^{8,9} A second group of methods assumes that similar binding sites recognize similar ligands. Any novel binding site close enough to binding sites of known ligands is therefore likely to bind the latter compounds. Structural approaches to binding site similarity detection, notably in the absence of fold conservation, are necessary to decipher subtle relationships among unrelated targets.¹⁰ For example, de Franchi et al. successfully identified, from binding site 3D comparisons, synapsin I as an off-target for some but not all protein kinases inhibitors.¹¹ Interestingly, binding affinity to synapsin I was qualitatively related to the level of binding site similarity for a set of different ATP-binding site inhibitors. Binding site similarity comparisons present the advantage to be insensitive to moderate

Received: May 21, 2013

variations of atomic coordinates, are applicable to orphan targets of known 3D structure, and are therefore quite complementary to ligand-based methods.

The last group of methods relies on existing protein–ligand X-ray structures and uses either molecular docking¹² or receptor–ligand pharmacophore search¹³ as a tool to match the query ligand to target space. Despite the fact that docking is a slow computational procedure amenable to serious errors in estimating binding free energies,¹⁴ it has been successfully used in many cases to discover new targets for existing ligands.^{15–17} Interestingly, success was neither restricted to a particular docking algorithm nor to particular protein classes suggesting that the method can be broadly used. Instead of docking compounds to protein–ligand binding sites, it is faster to just check whether the ligand fulfills pharmacophore features derived from protein–ligand complexes.¹³ Starting from protein–ligand complexes of known X-ray structure, pharmacophore hypotheses are generated by an automated pharmacophore perception algorithm¹⁸ and converted into pharmacophore queries. Receptor–ligand based pharmacophore models may be complemented by classical ligand-based pharmacophores to augment the scope of applicability, notably to protein classes (e.g., membrane proteins) for which few experimental structures but numerous ligands are available.

As seen from the above-cited examples, many methods with distinct applicability ranges have already been used. Beside practical considerations, there is however no rationale for considering a single profiling method across the entire target space. Surprisingly, there has been very few attempts to tackle this issue.¹⁹ On the basis of a comparative evaluation of several ligand-based and target-based methods in profiling 157 diverse ligands on 2556 different targets, we previously shown that (i) ligand-centric methods should be used whenever possible (which means when enough ligands are known for a particular target), (ii) 2D ligand descriptors are usually preferred to 3D descriptors, with the exception of low molecular-weight apolar ligands, (iii) protein–ligand docking should be reserved to polar and buried active sites of known structure for which few ligands are available, and (iv) receptor–ligand pharmacophore search may then be applied to all other protein structures.¹⁹

The current work is aimed at applying the above-described rules to a composite profiling method that uses the best possible screening method according to the current protein–ligand space under consideration. The fully automated profiling protocol covers 4371 unique targets and 4 different ligand-based and 2 different structure-based approaches, each being applied to nonoverlapping target sets. It successfully recovers the main targets of 189 clinical candidates in 72% of the cases and enables the deciphering of previously unknown cross-reactivities of some drug candidates to unrelated targets.

METHODS

PROFILER is a hybrid ligand profiling PipelinePilot²⁰ protocol (available upon request to the authors) selecting the best possible computational method according to ligand properties and the target space under investigation. We will here describe the data sources, then the different computational methods utilized by PROFILER, and last the criteria used to select the screening method. These choices will be later discussed in the next section, Results and Discussion.

Bioactivity Data Sources (Ligands, Targets, Affinities).

Three databases (ChEMBL,²¹ PubChem BioAssay,²² IUPHAR-DB²³) have been used as public repositories of bioactivity data

and unified in a single archive (from here on called BIOSTGB) as follows. The ChEMBL_12 release²⁴ was first downloaded in SQL format. All target–ligand associations for which an affinity value (K_i , K_d , IC_{50}) was available for a confidence level higher than 7 have been retained. Affinities values will from hereon be labeled as pXC50. Corresponding ligands were annotated by their SMILES strings and targets by their UniProt²⁵ access code. A total of 586 272 assays were downloaded in XML format from the PubChem BioAssay database.²⁶ Corresponding ligands were extracted in sd file format from their PubChem Substance²⁷ accession identifier (SID). Targets were first retrieved from their sequence identification number (GI) further transformed into the UniProt access code thanks to a converting table.²⁸ The IUPHAR-DB²⁹ PostgreSQL relational database was queried by in-house perl scripts to retrieve ligands as SMILES strings. All targets were annotated by their UniProt access code and bioactivity data were expressed in pXC50 values.

Ligand structures were standardized with the Standardizer software,³⁰ and solvent molecules or salts were removed. In case two molecules were present for a single entry, the one with the highest molecular weight was retained. A series of filters was then utilized in the Filter software³¹ to remove unwanted compounds: (i) with no carbon atoms, (ii) macrocyclic structures (cycles with more than 25 atoms), (iii) organo-metallic structures, (iv) peptides with more than 2 peptidic bonds), and (iv) lipids with more than 12 unsubstituted carbon atoms. For each compound, the most abundant tautomeric form at pH 7.4 along with the most likely ionization state was predicted by the Quacpac software.³² Two-dimensional atomic coordinates were then converted into a 3D structure using the Corina program.³³ A maximum of 100 different conformers was last generated for each ligand thanks to default settings of the Omega software.³⁴

To handle data redundancy, each target–ligand–affinity triplet was characterized by three descriptors: the Uniprot target name, the canonical stereoisomeric ligand SMILES string, and the affinity pXC50 value. When two pXC50 values were available for a single target–ligand pair, the most recent data according to the corresponding PubMed identifier from a search in the Batch Entrez database³⁵ was kept. When more than two affinity values exist, the difference between the lowest and the highest value was computed. If this difference is lower than 1, the median value is kept; otherwise, the complete triplet is removed. When more than 25 triplets were available for a human target, all triplets relative to orthologous proteins were removed from the final data set. If less than 25 triplets were available for a human target, triplets from orthologous proteins were considered with the condition that pXC50 values for ligands common to human and orthologous proteins differ by less than 1. In that case, all novel ligands of orthologous proteins never tested on the human variant were added to the pool of ligands for the human target.

Protein–Ligand Structures. The 2011 release of the sc-PBD database³⁶ was used as a source of 3D structures for protein–ligand complexes. It is available online³⁷ and registers 9877 entries (3034 proteins and 5339 ligands) postprocessed from the Protein Data Bank (PDB)³⁸ with curated 3D structures for targets, bound pharmacological ligands, and their corresponding binding sites. All files (MOL2 format) are ready for structure-based docking and pharmacophore search.

Ligand-Based Support Vector Classification. Target-specific support vector machine (SVM) binary classification models were obtained for 1227 targets annotated by at least 25

ligands. For each target, a set of training ligands was assembled by mixing all true actives from the BIOSTBG database and randomly selected decoys from the PubChem Substance repository.²⁷ For each target/SVM model, a constant balance between actives (20%) and decoys (80%) was kept. SVM models were generated with SVM-light³⁹ and ECFP4 fingerprints⁴⁰ as ligand descriptors. A fivefold cross-validation protocol was utilized, as previously described.⁴¹ Briefly, it splits, five times, each of the data sets into a training (4/5th of the data set) and a test set (1/5th of the data set) and analyzes the predictivity of SVM models on the remaining test sets, using the best trade-off C value optimized for each model. Statistical parameters for evaluating the different SVM models were the recall, precision, specificity, and F-measure.

$$\text{recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{F-measure} = 2 \times (\text{recall} \times \text{precision})/(\text{recall} + \text{precision})$$

where TP = true positives, FP = false positives, TN = true negatives, and FN = false negatives

To generate high-quality results, only 667 SVM models with a F-measure ≥ 0.7 (internal and external test sets) are used in the final profiling protocol.

Ligand-Based Support Vector Regression. Using SVM-light, 271 target-specific support vector regression (SVR) models were obtained from ECFP4 ligand fingerprints.³⁹ The 271 targets to which SVR affinity predictions were applied verify two conditions: (i) annotation by at least 25 ligands and (ii) homogeneous distribution of affinity pXC50 values in 3 intervals ([4–6], [6–8], and [8–10]). Each interval should comprise at least 15% of available bioactivity data for a particular target in order to qualify it for this QSAR modeling procedure. SVM models were generated using the above-described fivefold cross-validation protocol. Statistical parameters for evaluating the different SVR models were the crossvalidation correlation coefficient (Q^2), the mean average error (MAE), and the root-mean square error (RMSE).

$$Q^2 = 1 - \frac{\sum_{i=1}^n (Y_{\text{exp},i} - Y_{\text{pred},i})^2}{\sum_{i=1}^n (Y_{\text{exp},i} - \langle Y \rangle_{\text{exp}})^2}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |Y_{\text{exp}} - Y_{\text{pred}}|$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (Y_{\text{exp},i} - Y_{\text{pred},i})^2}{n}}$$

Where, n is the number of compounds, Y_{exp} is the experimental affinity, $\langle Y \rangle_{\text{exp}}$ is the mean of experimental affinity, and Y_{pred} is the predicted affinity.

To generate high-quality results, only 141 SVR models with a $Q^2 \geq 0.6$ and MAE ≤ 1.0 (internal and external test sets) are used in the final profiling protocol.

The diversity of all bioactivity classes (1227 classes with >25 ligands) was measured according to Turner et al.⁴² as follows:

$$D(A) = 1 - \frac{\sum_{j=1}^{N(A)} \sum_{K=1}^{N(A)} \text{SIM}(J, K)}{N(A)^2}$$

where $D(A)$ is the global diversity of database A (normalized between 0 and 1); $N(A)$ the number of compounds in database A; and $\text{SIM}(J, K)$ is the pairwise similarity of molecules J and K from data set A, expressed by the Tanimoto coefficient from ECFP4 fingerprints.

Ligand-Based Affinity Interpolation. This protocol was applied to 1014 targets (less than 25 annotated compounds; unsuitable for either SVR or SVM modeling) and interpolates the affinity F of a compound for one target from that of its n nearest neighbors (binding to the same target) using the Shepard weight function:⁴³

$$F = \sum_{i=1}^n w_i f_i \quad \text{with } w_i = \frac{d_i^{-2}}{\sum_{j=1}^n d_j^{-2}} \text{ and } d_i = 1 - \text{Tc}$$

where F is the interpolated pXC50, n is the number of neighbors within a distance threshold, f_i is the pXC50 value of the neighbor i , w_i is its assigned weight function, d_i the distance in chemical space between the neighbor i and the point to interpolate, and Tc is the Tanimoto similarity to ligand i , expressed from ECFP4 fingerprints.

Ligand-Based 3D Similarity Search. This protocol is only used if the compound to profile has less than three hydrogen-bond donors and/or acceptors. OMEGA conformers of the ligands are compared to OMEGA conformers of 144029 bioactive ligands of the BIOSTBG database ($\text{pXC50} \geq 5$), as well as to the X-ray structure of 2978 sc-PPDB ligands using ROCS.⁴⁴ Bioactive compounds are then ranked by decreasing *TanimotoCombo* score, after keeping the best score for every query–hit pair. Hits are finally retained if the following two conditions are verified:

$$\text{TanimotoCombo} \geq 1.4$$

$$\frac{\text{TanimotoCombo}}{2} - 0.2 \leq \text{ColorTanimoto} \leq \frac{\text{TanimotoCombo}}{2} + 0.2$$

Protein–Ligand Docking. This protocol was only applied to a specific set of 998 targets for which (i) less than 10 ligands are available in the BIOSTBG bioactivity database, (ii) a high-resolution X-ray structure is available in the sc-PDB archive, and (iii) the ligand-binding site verifies the following conditions:

$$\% \text{ hydrophobic} \leq 30$$

$$\text{OR } \% \text{ polar} \geq 30 \text{ AND } \% \text{ aromatic} \geq 15$$

$$\text{OR } \% \text{ accessible} \leq 3.6$$

Binding site properties (% hydrophobic, % polar, % aromatic, % accessible) are computed from pharmacophore-annotated cavity shapes, as implemented in the recently described VolSite algorithm.⁴⁵ VolSite places the binding site on 20-Å length grid lattice of resolution 1.5 Å and assigns pharmacophoric properties to each lattice point according to its nearest binding site heavy atom. The distribution of seven pharmacophoric properties (hydrophobic, aromatic, h-bond donor, h-bond acceptor, positive ionizable, negative ionizable, null) describes particularly well the binding site polarity and accessibility.⁴⁵

Default settings of the Surflex-Dock algorithm⁴⁶ were used for docking, starting from the Corina 3D mol2 file (ligand) and already prepared sc-PDB input files (target, protomol). For each of the 10 solutions, a molecular interaction fingerprint was calculated using the IFP program⁴⁷ and the similarity of the

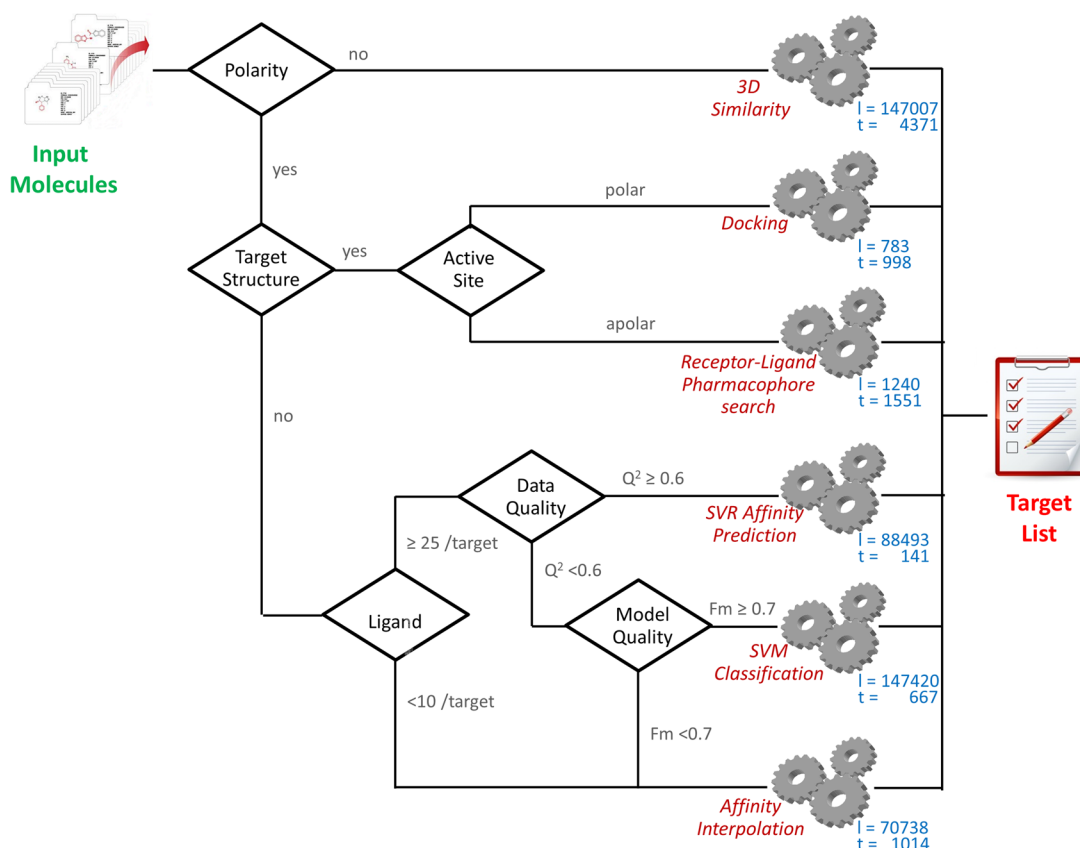


Figure 1. PROFILER workflow. **Polarity check:** number of hydrogen bond donors/acceptors < 3. **Target structure check:** ≤10 bioactive ligands and 3D structure available in the sc-PDB. **Active site check:** ≤30% hydrophobic site points OR ≥30% polar site points AND ≥15% aromatic site points OR ≤3.6% null site points (according to VolSite⁴⁵ cavity site point mapping). **Ligand check:** ≥25 bioactive ligands. **Data quality check:** $Q^2 \geq 0.6$ AND MAE ≤ 1.0 pK unit in SVR models. **Model quality check:** F-measure ≥ 0.7 in SVM models. The letters *l* and *t* indicate the number of ligands and targets from which each method is derived.

interaction fingerprint to that of the cocrystallized ligand was computed. Two fingerprints were outputted, one full-length bit string (7 bits/residue) registering all possible interactions (hydrophobic, aromatic face-to-face, aromatic edge-to-face, h-bond, ionic bond), and one short-length fingerprint (4 bits/residue) registering polar interactions only. Only poses with a similarity above 0.6 (on full-length fingerprint) and 0.5 (short-length fingerprint) were selected. The highest predicted affinity pose was finally retained if the predicted affinity score was above 3.0 and the crash score above −2.0.

Protein–Ligand Based Pharmacophore Search. This protocol was only applied to 1551 targets for which (i) less than 10 ligands are available in the BIOSTBG bioactivity database, (ii) a high-resolution X-ray structure is available in the sc-PDB archive, (iii) the ligand-binding site does not verify the properties required for docking (see above). The corresponding protein–ligand-based pharmacophores were extracted from the recently described PharmaDB database.¹⁹ These archive stores for every sc-PDB entry up to 10 pharmacophore queries (in CATALYST chm fileformat) with 3–6 features. With respect to standard queries, PharmaDB pharmacophores are directly derived from protein–ligand structures and places pharmacophoric features only on protein-interacting ligand atoms. A maximum of 100 conformers (FAST settings) of the ligand to profile were matched to PharmaDB pharmacophores (rigid fit) using otherwise default settings of the “Ligand Profiler” protocol in Discovery Studio.⁴⁸ Matches were scored by decreasing adapted fit (AF) value:

$$AF = \frac{(FM)}{T}$$

where *F* is the fitness value, *M* the number of matched features, and *T* the total number of pharmacophore features.

Data were merged by ligand conformers and target names, and a single AF value was retained for every target–ligand pair. Only matches with $AF > 2.6$ were retained.

PROFILER Workflow. A completely automated profiling workflow (Figure 1) was implemented in Pipeline Pilot.²⁰ Ligands are read in a 2D sd file format. A first check (*Polarity check*) computes the cumulated count of hydrogen bond donors/acceptors (HBAD). Ligands for which HBAD < 3 are screened only by 3D ROCS similarity to bioactive and scPDB ligands. The corresponding 4371 targets are kept in the target list only if the ROCS scores (TanimotoCombo and Color-Tanimoto) are within the above-defined thresholds (Ligand-Based 3D Similarity Search section). The best score for every protein is retained, and potential targets are saved in the final target list.

Remaining compounds are then submitted to a second control (*Target structure-check*) assigning specific screening methods to specific targets. For targets annotated by less than 10 ligands but present in the sc-PDB database, a structure-based screening check is done (*Active site check*). For polar and buried active sites, the ligands are docked to our collection of sc-PDB binding sites with Surflex-Dock (see the Protein–Ligand Docking section). For other binding sites, ligands are

matched to the corresponding protein–ligand PharmaDB pharmacophores (see the Protein–Ligand Based Pharmacophore Search section). Targets are ranked by decreasing docking and pharmacophore AF fitness scores and sent to the final target list if the scores are higher than the above-described thresholds.

For targets annotated by more than 25 ligands, SVM models are used to either predict the affinity for the corresponding targets (Ligand-Based Support Vector Regression section) or just predict the probability of target–ligand association (Ligand-Based Support Vector Classification section). The corresponding targets are sent to the final target list if an affinity value is predicted (SVR models) or if the association probability is positive (SVM models).

Targets for which neither SVM nor SVR models are of good quality, as well as those annotated by less than 25 ligands but for which no X-ray structure is available, are pooled to the set of targets investigated by the affinity interpolation protocol. If the compound to profile is close enough in chemical space (using ECFP4 fingerprints) to BIOSTBG compounds, its affinity is inferred from that of these neighbors (Ligand-Based Affinity Interpolation section). All methods utilized in the workflow are summarized in a look-up table (Supporting Information Table S1) along with the target space to which it applies and the diverse thresholds used for their applicability domain.

All targets selected by the different computational protocols are finally stored in a compound-specific final report (Supporting Information Table S2) with the following data: profiling method, target UniProt access codes, identifiers, names, and PDB code; predicted affinity (interpolation, SVR, docking), fitness value (pharmacophore search).

External Ligand Set. The PROFILER workflow was evaluated on a set of clinical candidates which were retrieved from the Integrity database⁴⁹ and verify the following constraints: (i) no identity to any of the BIOSTBG ligands and (ii) annotation by a target present in the BIOSTBG database. All structures were downloaded in sd file format and treated as explained above (see the Bioactivity Data Sources section). After filtering, a total of 189 compounds (Supporting Information Table S3) were retained for an external validation of the profiling protocol. For each compound, the main target (highest affinity target) was stored.

In Vitro Binding Assays. Compounds were obtained from commercial sources (Axon MedChem, Selleck Bio), with guaranteed purity >97%. All in vitro binding (adenosine A_{2A} receptor, bromodomain-containing protein 4), functional (cannabinoid CB₂ receptor antagonist, CysLT1 receptor antagonist, CysLT2 receptor antagonist, A_{2a} receptor antagonist) and enzyme (phosphodiesterase 5, CamK1 delta, MST4 kinase, TIE2 kinase, B-raf kinase) assays were realized at Cerep (<http://www.cerep.fr>) using standard conditions of the following catalogue assay numbers (A_{2A} agonist binding no.

0004, BRD4 inhibitor no. 3628, CB₂ antagonist no. 1747, CysLT1 receptor antagonist no. 1607, CysLT2 receptor antagonist no. 2052, A_{2a} receptor antagonist no. 2056, phosphodiesterase 5 inhibition no. 0204, CamK1 delta inhibition no. 2922, TIE2 kinase inhibition no. 2736, B-raf inhibition no. 3294). All compounds were tested in duplicates at a single concentration of 10 μ M.

Dose–response curves (eight concentrations in duplicates) were obtained for PDE5 inhibition (assay no. 0204) by rollofylline and CysLT1 functional antagonism (assay no. 1607) by PF-2545920. The IC₅₀ values (concentration causing a half-maximal inhibition of control specific activity) and Hill coefficients (nH) were determined by nonlinear regression analysis of the inhibition curves generated with mean replicate values using Hill equation curve fitting. This analysis was performed using a software developed at Cerep (Hill software) and validated by comparison with data generated by the commercial software SigmaPlot 4.0 for Windows (1997 by SPSS Inc.).

RESULTS AND DISCUSSION

The current study aims at designing a hybrid workflow for predicting the most likely targets of bioactive compounds. Contrarily to all existing approaches, it does not rely on a single virtual screening method (e.g., ligand-based 2D similarity search or protein–ligand docking) but capitalizes on known strengths and limitations of a panel of computational profiling technologies (ligand-based and structure-based approaches);¹⁹ then, it selects the presumably best method according to the target of interest. We therefore decided to use 2D similarity search whenever possible (all targets annotated by more than 25 ligands) and only use other methods in the remaining cases. 3D shape similarity was therefore reserved to hydrophobic molecules, protein–ligand docking was limited to targets of known X-ray structure but polar active site, and receptor–ligand pharmacophore matches was applied to all other X-ray structures from the sc-PDB database.

Bioactivity Data Extraction. Of primary importance to generate high-quality QSAR models is the selection of appropriate bioactivity data. Three databases (ChEMBL, PubChem BioAssay, IUPHAR-DB) were used as main sources and merged while removing redundancy at the target–ligand–data triplet level. To achieve this processing, a homogeneous annotation of targets (by Uniprot access code) and ligands (by stereoisomeric SMILES strings) was adopted. To guarantee the highest possible quality, only binding data that vary by less than one pXC50 unit for a particular target–ligand pair were kept. This drastic selection protocol removes ca. 75% of the ligands and targets and more than 90% of all data points in ChEMBL (Table 1) but produces very high quality data to train QSAR models. The main criteria for discarding data are the

Table 1. Postprocessing of Bioactivity Data Sources

database	ligands		targets		data ^a	
	original	filtered	original	filtered	original	filtered
ChEMBL	1076486	243625	8703	2337	6186504	454611
PubChem	223942	61393	33636	322	659497	75496
IUPHAR-DB	2761	1633	624	325	6734	3169
sc-PDB	5339	1537	3034	2549	9877	3820
merged (STBG)		308188		4371		537096

^apXC50 values (ChEMBL, PubChem, IUPHAR-DB) or 3D structures (sc-PDB).

following: (i) high variation of the pXC50 value for a given protein–ligand pair (more than 1 unit between the lowest and the highest value), (ii) low ChEMBL confidence level (<7), (iii) undefined target, and (iv) high redundancy (several similar pXC50 values for a single protein–ligand pair).

Altogether, the merged database (BIOSTBG) archives 533 276 binding data and 3820 protein–ligand structures, from 308 188 unique ligands and 4371 unique targets (Table 1). The target space covered by the database spans the major target families, namely G protein-coupled receptors, enzymes, protein kinases, proteases, and ion channels (Figure 2). The

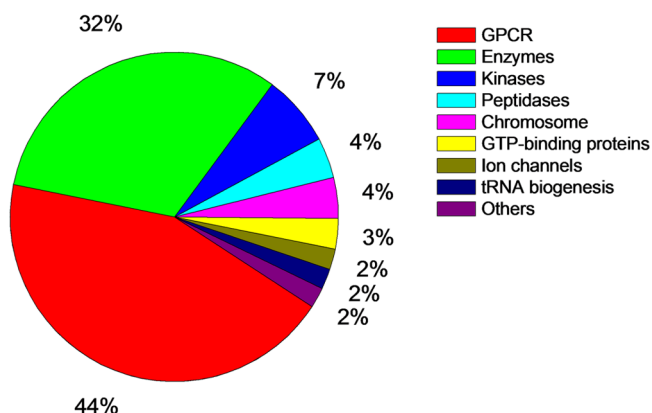


Figure 2. Target space coverage by the BIOSTBG database, as a function of the KEGG-BRITE functional annotation.⁵⁰

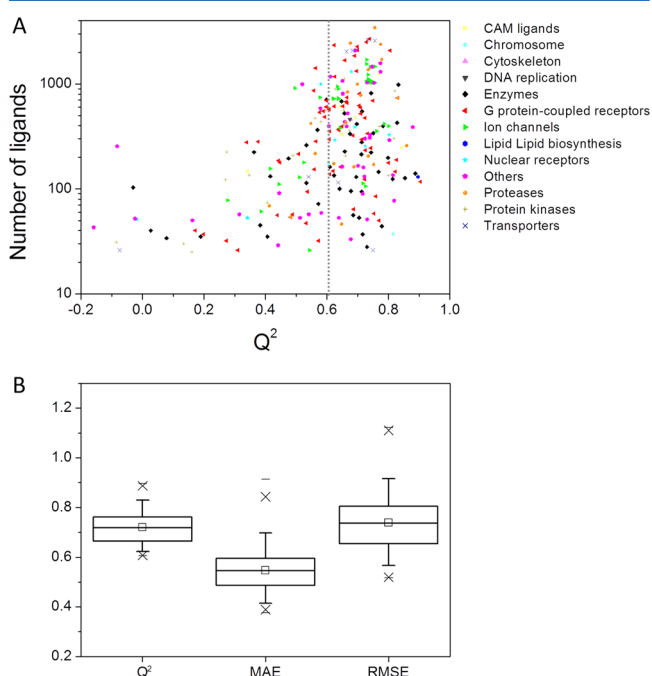


Figure 3. Performance of support vector regression (SVR) models. (A) Cross-validated coefficient Q^2 as a function of the number of ligands by target. A dotted line indicates the lowest acceptable Q^2 value for incorporating the corresponding target-specific model into the final profiling workflow. (B) Box-and-whisker plot of statistical parameter distributions (cross-validated correlation coefficient Q^2 , mean average error MAE, root-mean square error MSE) for selected SVR models. The box delimits the 25th and 75th percentiles, the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1st and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

KEGG-BRITE⁵⁰ functional annotation, used here to classify targets, is an ontology database representing functional hierarchies of various biological objects, including molecules, cells, and organisms. Kinases and peptidases were separated from the enzymes set just because these proteins are overrepresented in the sc-PDB data set from which we source 3D structures. It should be pointed that our data set stores fewer ion channels ($n = 49$) and transporters ($n = 70$) than the corresponding proportion recently reported for marketed drugs,⁵¹ a discrepancy likely to be due to the nature of targets annotated in the herein investigated public domain databases.

Quality of SVR and SVM Models. Among the 1227 targets from the BIOSTBG data set which are annotated by at least 25 compounds, 271 verify a relatively homogeneous distribution of binding data over three pXC50 intervals ([4–6], [6–8], [8–10]) and therefore constitute a very good data set for calibrating the accuracy of SVR models. Following previous recommendations,⁵² only models with a crossvalidated Q^2 higher than 0.6 and a mean average error (MAE) lower than

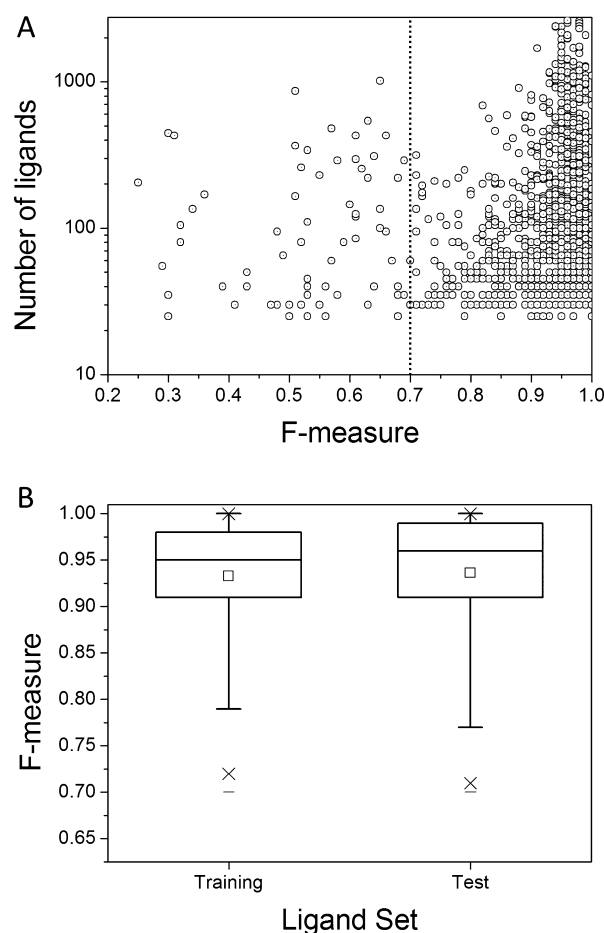


Figure 4. Performance of support vector classification (SVM) models. (A) F-measure as a function of the number of ligands by target. A dotted line indicates the lowest acceptable F-measure for incorporating the corresponding target-specific model into the final profiling workflow. (B) Box-and-whisker plot of the F-measure distribution for acceptable SVM models applied to the training and external test sets. The box delimits the 25th and 75th percentiles, and the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1st and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

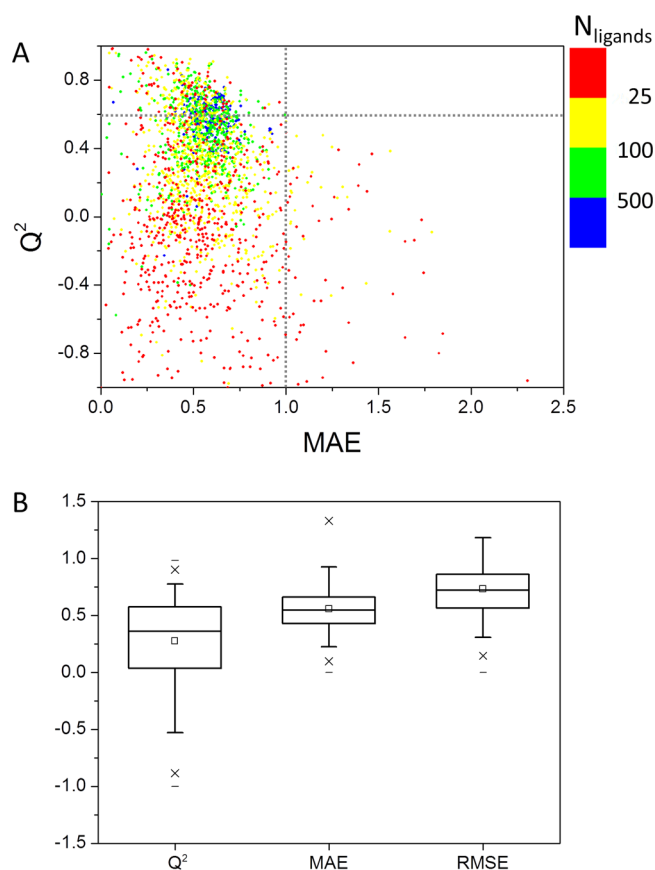


Figure 5. Performance of interpolation models. (A) Cross-validated correlation coefficient (Q^2) and mean average error (MAE) as a function of the number of ligands (N_{ligands}) by target. (B) Box-and-whisker plot of the distribution of statistical parameters (cross-validated correlation coefficient, Q^2 ; mean average error, MAE; root-mean square error, RMSE). The box delimits the 25th and 75th percentiles, and the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box. Crosses delimit the 1st and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

one pXC50 unit were selected for incorporation in the final profiling workflow. 141 out of the 271 models (52%) satisfy these conditions (Figure 3). Most but not all of the corresponding targets are described by more than 100 ligands in our database. They originate from the major target families of pharmaceutical interest with 41 enzymes, 40 GPCRs, 13 proteases, 11 protein kinases, and 5 nuclear receptors.

For the remaining 1086 targets, lower resolution SVM binary classification models have been used. In 61% of the cases (667 targets), the SVM models are accurate enough to discriminate true actives from decoys and exhibit an F-measure above 0.7 (Figure 4). Interestingly, the accuracy of these models was

equivalent on training and external test sets (Figure 4). For the corresponding 667 proteins, our profiling protocol will thus only output the likelihood of target–ligand association according to the corresponding SVM classification models. The average precision of the 667 models (92%) is higher than that reported by a decision tree (82%) on a data set with a comparable target coverage (340 targets) albeit with much less training ligands (2721).⁵³ Failure to distinguish true actives from decoys in some target classes was essentially due to the very high chemical diversity of ligand sets for certain promiscuous targets (e.g., cytochromes) or the presence of multiple binding sites for a single target (e.g., proteases, targets involved in protein–protein interactions). However, low diversity cannot be invoked to explain the high predictivity of either SVR or SVM models since ca. 95% of bioactivity classes present a high diversity according to the method of Turner et al.⁴² (Supporting Information Figure S1).

Quality of the Affinity Interpolation Method. A simple alternative to SVM regression to predict binding affinities is to interpolate the affinity of a compound from that of its nearest neighbors sharing the same target. Using an inverse distance weighting (IDW) approach, a weight is assigned to every scatter point (reference) that decreases with increasing distance in chemical space to the interpolation point (test compound).⁵⁴ In our application, the chemical space is described by ECFP4 circular fingerprints and neighbors are selected by Tanimoto similarity of the corresponding fingerprints to that of the test compound. After preliminary trials, a Tc threshold of 0.37 was chosen to select neighbors in chemical space since it afforded the best compromise between high Q^2 values, low MAE, and low percentage of data points outside the applicability domain (Supporting Information Figure S2).

For 2087 targets of the BIOSB database (annotated by at least five ligands, no requirement for equidistribution of pXC50 values), the interpolation method could be applied with a fivefold crossvalidation scheme to predict the pXC50 value of 431 000 ligands (Figure 5, Table 2). Altogether, pK_i values could be correctly predicted ($Q^2 \geq 0.6$ and $\text{MAE} \leq 1.0$) for 461 out of the 2087 target classes (22% of the targets). The main reason for failure was the paucity of ligand information for some targets (Figure 5). The method is clearly sensitive to the number of compounds annotated for every target with a gradual increase of the mean Q^2 from 0.28 (at least five ligands/target) to 0.51 (at least 100 ligands/target; Table 2). Interestingly, the MAE remains constant whatever the target class definition, illustrating the fact that pK_i values are distributed in a narrow range (6–7) for many targets. Comparing interpolation to SVM regression was possible on the list of 271 targets annotated by at least 25 compounds and for which at least 15% of available data are distributed in three pXC50 intervals ([4–6], [6–8], and [8–10]). The interpolation method logically performs better when data are better distributed (mean Q^2 increasing from 0.40 to 0.57, Table 2) with 47% of the models being now

Table 2. Prediction of pXC50 Values Per Interpolation

ligands ^a	targets ^b	equidistribution ^c	Q^2 ^d	MAE ^e	correct models, % ^f
≥ 5	2087	no	0.28 ± 0.40	0.56 ± 0.23	22
≥ 25	1473	no	0.40 ± 0.28	0.56 ± 0.17	27
≥ 25	271	yes	0.57 ± 0.19	0.64 ± 0.13	47
≥ 100	767	no	0.51 ± 0.19	0.56 ± 0.13	34

^aMinimum number of ligands per target. ^bNumber of unique targets. ^cHomogeneous distribution of pXC50 values in three intervals ([4–6], [6–8], and [8–10]; at least 15% of all data in each interval). ^dFivefold cross-validated correlation coefficient. ^eMean-average error (in pXC50 unit). ^fModels with $Q^2 \geq 0.6$ and $\text{MAE} \leq 1$.

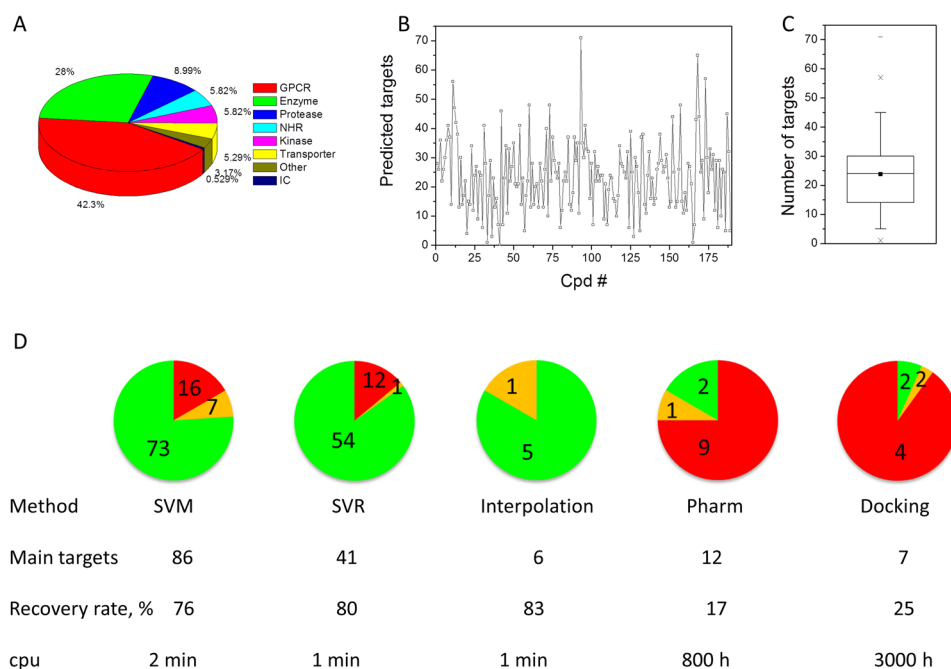


Figure 6. Profiling of 189 clinical candidates with the PROFILER protocol. (A) Target space covered by the compounds. (B) Number of predicted targets for each compound (see Supporting Information Table 2 for compound numbering). (C) Box-and-whisker plot of the distribution of predicted targets. The box delimits the 25th and 75th percentiles, and the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and a filled square in the box. Crosses delimit the 1st and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash. (D) Profiling successes and failures according to the profiling method. A success (green pie chart section) indicates that the main target is enclosed in the target list. A near-success (orange chart section) indicated that only a target very similar to the true one is in the target list. A failure (red pie chart section) is reported elsewhere. Numbers in each pie chart section refers to the number of profiling cases. The CPU time is given for profiling all 189 ligands on a 3.16 GHz Intel Core Duo E 8500 processor with 4 GB RAM. Ligand-based SVM, SVR, and interpolation profiling were realized on single processing units whereas structure-based methods (receptor–ligand pharmacophore search and protein–ligand docking) were done on 100 processors of the CC-IN2P3 Linux farm (IN2P3 calculation center, Villeurbanne, France).

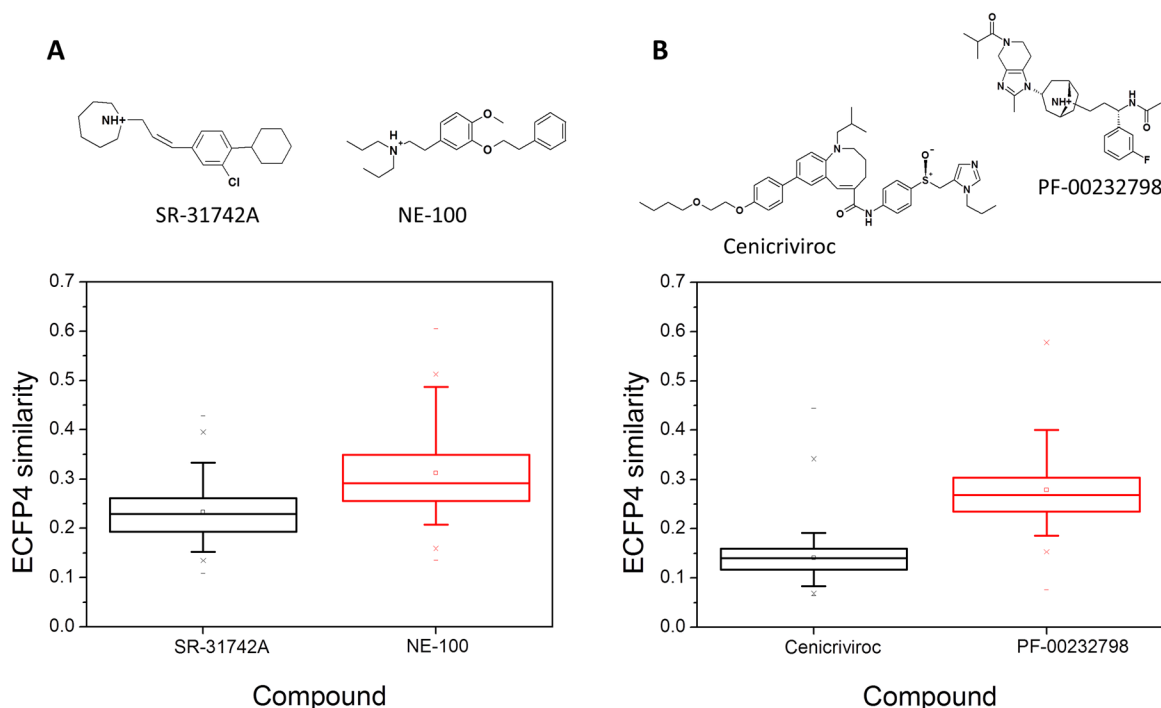
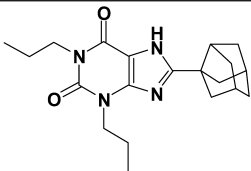
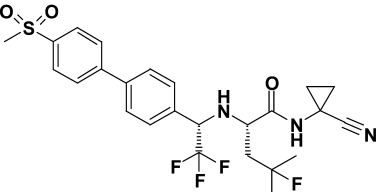
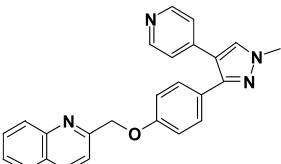
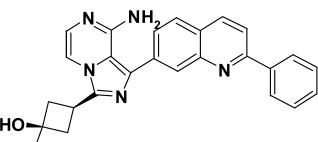


Figure 7. Distribution of the similarity (Tanimoto coefficient from ECFP4 fingerprints) of the query compound to all BIOSTBG ligands of its main target. The box delimits the 25th and 75th percentiles, and the whiskers delimit the 5th and 95th percentiles. The median and mean values are indicated by a horizontal line and a filled square in the box. Crosses delimit the 1st and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash. (A) Sigma-1 receptor ligands. (B) Chemokine CCR5 ligands. In both cases, the first ligand of the pair was not correctly profiled whereas profiling the second ligand of the pair was successful. Sigma-1 and CCR5 targets are assigned by SVM and SVR profiling, respectively. The chemical structure of all ligands is shown above the distribution plots.

Table 3. Experimental Validation of PROFILER Results for Four Clinical Candidates

Drug	Main target	Predicted off- targets	Predicting method	% inhibition (10 μ M)
 Rolofylline	Adenosine A ₁	PDE5	SVR ^a	28
		CamK1 delta	Pharm ^b	n.d. ^h
		MST4 kinase	Pharm ^c	n.d.
		BRD4	Pharm ^d	8
 Odanacatib	Cathepsin K	Cannabinoid	SVR ^e	n.d.
		CB2 receptor		
 PF-2545920	PDE10A	CLT1 receptor	SVM ^f	127
		CLT2 receptor	SVM	n.d.
		TIE2 kinase	SVM	n.d.
		B-Raf kinase	SVM	7
 Linsitinib	Insulin-like growth factor 1 receptor	Adenosine A _{2A}	SVR ^g	16

^aPredicted pXC50 = 6.39. ^bTc1 = 0.71, Tc2 = 0.67, AF = 3.27. ^cTc1 = 0.65, Tc2 = 0.67, AF = 3.25. ^dTc1 = 0.62, Tc2 = 1.00, AF = 3.48. ^ePredicted pXC50 = 7.10. ^fNo affinity predicted. ^gPredicted pXC50 = 6.78. ^hNo detectable binding.

correct according to our criteria ($Q^2 \geq 0.6$ and $MAE \leq 1.0$). It is slightly inferior to results obtained by SVR (52% of correct models, see above). However, given the simplicity and speed of the method, these results are quite encouraging and in agreement with previous predictions on a much narrower target–ligand space.⁵⁴ Since the interpolation method is potentially useful for any novel compound close enough to existing ligands, we decided to keep it in our global profiling protocol but to apply it to 1014 targets for which less than 25 ligands are available and for which neither SVR nor SVM models were satisfactory.

Computational Profiling of 189 Clinical Candidates.

Here, 189 compounds absent from our training sets and under clinical investigation were extracted from the Integrity database (Supporting Information Table S3). They were annotated to bind to 152 unique targets covering a relevant target space (Figure 6A) with no major bias with respect to all targets stored in the BIOSTBG data set (Figure 2). All compounds were profiled, starting from input 2D sd files, with the PROFILER Pipeline Pilot protocol in a fully automated manner to yield on average to ca. 25 putative targets for every compound (Figures 6B and C). With respect to the total number of targets

to which the profiling protocol applies (4371), this small number (ca. 0.5%) is very interesting since it enables a fast experimental validation by in vitro binding or functional assays at a quite reasonable cost.

Most of the targets to recover have to be found by a ligand-centric approach (Figure 6D) which illustrates a well-known bias in modern drug discovery: most compounds under clinical evaluation are targeting a reduced number of molecular targets (GPCRs, kinases, proteases) for which considerable ligand information is already available. It is therefore logical that the profiling protocol utilizes very frequently a ligand-based method (SVM or SVR) and rarely target-centric methods (Figure 6D). Altogether, the protocol could recover the main target of the 189 compounds in 72% of the cases (136/189). In 12 cases (6%), the main target could not be found but a very similar protein (e.g., adenosine A_{2A} receptor instead of adenosine A₁ receptor for compound BAY-68-4986, Supporting Information Table S3) was saved in the target list. A very large majority of successes originate from ligand-based profiling methods (ca. 80%) with no preference for one method with respect to another, thereby justifying our hybrid profiling workflow. Disappointingly, structure-based approaches (docking, protein–ligand pharmacophore

search) were only occasionally satisfactory with 17–25% of success rate (Figure 6D). This observed discrepancy should not however be analyzed by simply stating that ligand-based profiling is superior to target-centric methods. There are two major reasons explaining the lower success rates of structure-based approaches: (i) “easy” profiling cases (e.g., aminergic GPCRs, serine/threonine protein kinases) are already handled by ligand-centric methods, and (ii) failure to correctly assign a target to a compound by docking or pharmacophore match may simply be due to the fact that the compound binds to a cavity not stored in the sc-PDB archive. For example, shikimate kinase was absent from the target list of the shikimate kinase inhibitor 446170, since the latter compound is a ATP noncompetitive inhibitor⁵⁵ and only ATP-binding sites (PDB entries 2iyv, 2iyw, 3baf) have been screened by structure-based approaches. Likewise, ligand-based approaches fail when the test compound is too dissimilar to known target ligands. As example, we computed the pairwise similarity of two pairs of compounds to all ligands of their corresponding main targets (Figure 7). To remove any bias, we chose two pairs of compounds sharing the same target (sigma-1 receptor, chemokine CCR5 receptor), one being successfully profiled and one for which the profiling failed. In both cases, the profiling fails whatever the method (SVM, Figure 7A; SVR, Figure 7B) when the test compound is chemically dissimilar to true actives of the corresponding target.

Experimental Validation of Predicted Off-Targets for Four Clinical Candidates. No conclusions on the selectivity and specificity of the current profiling protocol can be drawn since all potential targets have not been challenged experimentally. However, the predicted target–ligand matrix (189 ligands \times 4371 targets) offers us the opportunity to experimentally validate some of our predictions. Ten novel target predictions (Table 3) were chosen on the basis of commercial availability for the ligands, and possible *in vitro* evaluation using binding, enzyme, or functional assays. Out of the 10 predictions, detectable binding/inhibition could be confirmed in 5 cases, at a single ligand concentration of 10 μ M (Table 3). More precise dose–response curves were obtained in two cases (Figure 8). A moderate affinity ($IC_{50} = 13.8 \mu$ M) to phosphodiesterase 5 (PDE5) was confirmed for the adenosine A1 receptor antagonist rolofylline (Figure 8A), whereas the known phosphodiesterase 10A inhibitor PF-2545920 was confirmed to potentially inhibit the cysteinyl leukotriene receptor 1 (CysLT1) with an IC_{50} of 142 nM (Figure 8B). If binding of rolofylline to adenosine (A_1 receptor) and cAMP (PDE5) binding sites could not be considered as surprising, the second example is particularly interesting since it illustrates an unprecedented cross-reactivity of a single compound against totally unrelated targets; an intracellular enzyme mainly expressed in the striatum (PDE10A) and a GPCR (CysLT1) expressed in smooth muscle lung cells. PF-2545920 has been reported to be a potent ($IC_{50} = 0.37$ nM) and highly selective PDE10A inhibitor by targeting a selectivity pocket unique to this isoform.⁵⁶ Neither a search in ChEMBL nor in PubMed for compounds sharing PDE 10A and CysLT1 as targets returned a single positive answer, confirming the novelty of the observed cross-reaction. This study brings yet another confirmation that fine selectivity among related targets is no guarantee to claim for general selectivity and that cross-reactivity may readily jump across target boundaries.⁵⁷

Much more experimental work should be necessary to really estimate the false positive rate of the method. For the five compounds investigated here at the experimental level, we can just state that the mean false positive rate (50%) is rather similar to that reported by the SEA ligand-based similarity method.⁹

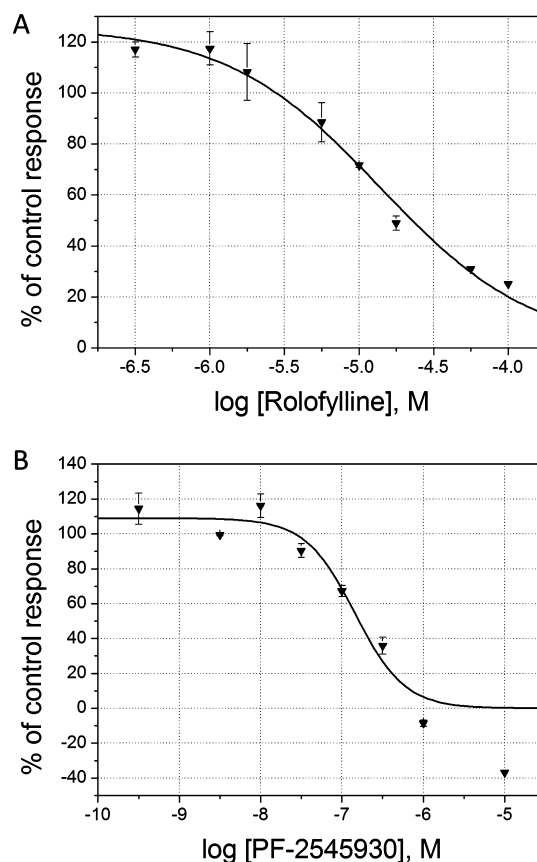


Figure 8. Off-target binding of two clinical candidates. (A) Human platelet PDE5 inhibition by rolofylline. Enzymatic activity was measured upon detection of [3 H]5'GMP by scintillation counting, after incubating 1 μ M [3 H]cGMP and increasing concentrations of the competitor. (B) CystLT1 antagonism by PF-2545930. Receptor activity was measured in HEK-293 cells upon detection of intracellular Ca^{2+} by fluorimetry, after incubating 30 nM LTC4 and increasing concentrations of the competitor. Data represent mean values \pm s.e.m., performed on duplicate experiments.

CONCLUSIONS

The present study describes the first computational profiling protocol that capitalizes on existing knowledge (binding data, structures) regarding target and ligand spaces and uses either ligand-centric or structure-based methods in series to identify, among 4371 proteins, the plausible targets of bioactive compounds. When applied to a set of 189 clinical candidates, the fully automated profiling protocol was able to recover the main target of profiled compounds in 72% of the cases, usually by means of fast ligand-based similarity searches. With respect to the very unfavorable ratio between success rate and CPU screening time, it is however legitimate to question the utility of both structure-based methods in the herein presented hybrid profiling protocol.

Importantly, our profiling protocol permits to deciphering novel cross-reactivity among unrelated targets for two compounds, notably by identifying a nanomolar inhibitor for both a phosphodiesterase (PDE 10A) and a GPCR (CysLT1 receptor). Given its simplicity, accuracy, and speed, computational ligand profiling is likely to play an important role in estimating the most likely target space spanned by bioactive compounds and therefore prioritize early safety assessment.

■ ASSOCIATED CONTENT

■ Supporting Information

Applicability domain of the profiling methods used in PROFILER (Table S1), example of PROFILER output report (Table S2), computational profiling of 189 compounds from the INTEGRITY database (Table S3), distribution of the diversity of bioactivity classes (Figure S1), variation of statistical parameters for predicting pXC50 values by interpolation (Figure S2). This material is available free of charge via the Internet at <http://pubs.acs.org>. The PROFILER Pipeline Pilot protocol (in xml format) is available upon request to D.R.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +33 3 68 85 42 35. Fax: +33 3 68 85 43 10. E-mail: rognan@unistra.fr.

Present Addresses

[†]J.M.: Department of Structural and Chemical Biology, Mount Sinai School of Medicine, New York, NY 10029-6574, United States.

[‡]R.B.: Institut de Recherches en Technologies et Sciences pour le Vivant, CEA, F-38054 Grenoble, France.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Conseil regional d'Alsace for funding J.M. and Philippe Dupuis (Cerep, Celle l'Evescault, France) for the in vitro pharmacological studies. The IPHC grid (Strasbourg), CC-IN2P3 (Villeurbanne), and GENCI (Project x2011075024) are acknowledged for providing computational resources. We thank Jérôme Pansanel (IPHC) and Pascal Calvat (CC-IN2P3) for their excellent support.

■ REFERENCES

- (1) Swinney, D. C.; Anthony, J. How were new medicines discovered? *Nat. Rev. Drug. Discovery* **2011**, *10*, S07–S19.
- (2) Kotz, J. Phenotypic screening, take two. *SciBX* **2012**, *5*, 10.1038/scibx.2012.1380
- (3) Scheiber, J.; Jenkins, J. L. Chemogenomic analysis of safety profiling data. *Methods Mol. Biol.* **2009**, *575*, 207–223.
- (4) Hopkins, A. L. Drug discovery: Predicting promiscuity. *Nature* **2009**, *462*, 167–168.
- (5) Raju, T. N. The Nobel chronicles. *Lancet* **2000**, *355*, 1022–1024.
- (6) Rognan, D. Structure-based approaches to target fishing and ligand profiling. *Mol. Inf.* **2010**, *29*, 167–187.
- (7) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.
- (8) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175–181.
- (9) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Cote, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361–367.
- (10) Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the similarity between protein ligand-binding sites? *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 209–220.
- (11) de Franchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS One* **2010**, *5*, e12214.
- (12) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153* (Suppl 1), S7–26.
- (13) Steindl, T. M.; Schuster, D.; Laggner, C.; Langer, T. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J. Chem. Inf. Model* **2006**, *46*, 2146–2157.
- (14) Novikov, F. N.; Zeifman, A. A.; Stroganov, O. V.; Stroylov, V. S.; Kulkov, V.; Chilov, G. G. CSAR scoring challenge reveals the need for new concepts in estimating protein-ligand binding affinity. *J. Chem. Inf. Model* **2011**, *51*, 2090–2096.
- (15) Muller, P.; Lena, G.; Boilard, E.; Bezzine, S.; Lambeau, G.; Guichard, G.; Rognan, D. In silico-guided target identification of a scaffold-focused library: 1,3,5-triazepan-2,6-diones as novel phospholipase A2 inhibitors. *J. Med. Chem.* **2006**, *49*, 6768–6778.
- (16) Yang, L.; Chen, J.; He, L. Harvesting candidate genes responsible for serious adverse drug reactions from a chemical-protein interactome. *PLoS Comput. Biol.* **2009**, *5*, e1000441.
- (17) Yang, L.; Wang, K.; Chen, J.; Jegga, A. G.; Luo, H.; Shi, L.; Wan, C.; Guo, X.; Qin, S.; He, G.; Feng, G.; He, L. Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome–clozapine-induced agranulocytosis as a case study. *PLoS Comput. Biol.* **2011**, *7*, e1002016.
- (18) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model* **2005**, *45*, 160–169.
- (19) Meslamani, J.; Li, J.; Sutter, J.; Stevens, A.; Bertrand, H. O.; Rognan, D. Protein-Ligand-Based Pharmacophores: Generation and Utility Assessment in Computational Ligand Profiling. *J. Chem. Inf. Model* **2012**, *52*, 943–955.
- (20) Pipeline Pilot, version 8.5; Accelrys, Inc.: San Diego, CA, 2011.
- (21) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–1107.
- (22) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012**, *40*, D400–D412.
- (23) Sharman, J. L.; Benson, H. E.; Pawson, A. J.; Lukito, V.; Mpamhanga, C. P.; Bombail, V.; Davenport, A. P.; Peters, J. A.; Spedding, M.; Harmar, A. J.; Nc, I. IUPHAR-DB: updated database content and new features. *Nucleic Acids Res.* **2013**, *41*, D1083–D1088.
- (24) ChEMBLdb. <https://www.ebi.ac.uk/chembl/> (accessed Jul 30, 2013).
- (25) The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **2010**, *38*, D142–148.
- (26) PubChem BioAssay. <http://www.ncbi.nlm.nih.gov/pcassay> (accessed Jul 30, 2013).
- (27) Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discovery Today* **2010**, *15*, 1052–1057.
- (28) Uniprot. ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz (accessed Jul 30, 2013).
- (29) IUPHAR-DB. <http://www.iuphar-db.org/> (accessed Jul 30, 2013).
- (30) ChemAxon Standardizer, version 5.5.0.1; ChemAxon: Budapest, Hungary, 2011.
- (31) Filter, version 2.1.1; OpenEye Scientific Software: Santa Fe, NM, 2011.
- (32) Quacpac, version 1.5.0; OpenEye Scientific Software: Santa Fe, NM, 2011.
- (33) Corina, version 3.1; Molecular Networks GmbH - Computerchemie: Erlangen, Germany, 2011.
- (34) Omega, version 2.4.6; OpenEye Scientific Software: Santa Fe, NM, 2011.
- (35) Batch Entrez. <http://www.ncbi.nlm.nih.gov/sites/batchentrez> (accessed Jul 30, 2013).

- (36) Meslamani, J.; Rognan, D.; Kellenberger, E. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics* **2011**, *27*, 1324–1326.
- (37) sc-PDB. <http://bioinfo-pharma.u-strasbg.fr/scPDB> (accessed Jul 30, 2013).
- (38) Rose, P. W.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dimitropoulos, D.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Prlic, A.; Quesada, M.; Quinn, G. B.; Ramos, A. G.; Westbrook, J. D.; Young, J.; Zardecki, C.; Berman, H. M.; Bourne, P. E. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.* **2012**, *41*, D475–D482.
- (39) SVMlight, version 6.02. <http://svmlight.joachims.org/> (accessed Jul 30, 2013).
- (40) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model* **2010**, *50*, 742–754.
- (41) Meslamani, J.; Rognan, D. Enhancing the accuracy of chemogenomic models with a three-dimensional binding site kernel. *J. Chem. Inf. Model* **2011**, *51*, 1593–1603.
- (42) Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18–22.
- (43) Shepard, D. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, Las Vegas, NV, August 27–29; ACM: New York, 1968; pp 517–524.
- (44) ROCS, version 3.1.2; OpenEye Scientific Software: Santa Fe, NM, 2011.
- (45) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287–2299.
- (46) Spitzer, R.; Jain, A. N. Surflex-Dock: Docking benchmarks and real-world application. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 687–699.
- (47) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model* **2007**, *47*, 195–207.
- (48) *Discovery Studio*, version 3.5; Accelrys, Inc.: San Diego, CA, 2012.
- (49) *Integrity*. <https://integrity.thomson-pharma.com/integrity/xmlxsl/> (accessed Jul 30), 2013.
- (50) KEGG-BRITE. <http://www.genome.jp/kegg/brite.html> (accessed Jul 30, 2013).
- (51) Rask-Andersen, M.; Almen, M. S.; Schioth, H. B. Trends in the exploitation of novel drug targets. *Nat. Rev. Drug Discovery* **2011**, *10*, 579–590.
- (52) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476–488.
- (53) Strömbergsson, H.; Lapins, M.; Kleywegt, G. J.; Wikberg, J. E. S. Towards Proteome-Wide Interaction Models Using the Proteochemometrics Approach. *Mol. Inf.* **2010**, *29*, 499–508.
- (54) Vidal, D.; Mestres, J. In Silico Receptorome Screening of Antipsychotic Drugs. *Mol. Inf.* **2010**, *29*, 543–551.
- (55) Han, C.; Zhang, J.; Chen, L.; Chen, K.; Shen, X.; Jiang, H. Discovery of *Helicobacter pylori* shikimate kinase inhibitors: bioassay and molecular modeling. *Bioorg. Med. Chem.* **2007**, *15*, 656–662.
- (56) Verhoest, P. R.; Chapin, D. S.; Corman, M.; Fonseca, K.; Harms, J. F.; Hou, X.; Marr, E. S.; Menniti, F. S.; Nelson, F.; O'Connor, R.; Pandit, J.; Proulx-Lafrance, C.; Schmidt, A. W.; Schmidt, C. J.; Suiciak, J. A.; Liras, S. Discovery of a novel class of phosphodiesterase 10A inhibitors and identification of clinical candidate 2-[4-(1-methyl-4-pyridin-4-yl-1H-pyrazol-3-yl)-phenoxy-methyl]-quinoline (PF-2545920) for the treatment of schizophrenia. *J. Med. Chem.* **2009**, *52*, 5188–5196.
- (57) Lin, H.; Sassano, M. F.; Roth, B. L.; Shoichet, B. K. A pharmacological organization of G protein-coupled receptors. *Nat. Methods* **2013**, *10*, 140–146.