

Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel




Jamel Meslamani and Didier Rognan*

Structural Chemogenomics, Laboratory of Therapeutic Innovation, UMR 7200 CNRS, University of Strasbourg, F-67400 Illkirch, France

S Supporting Information

ABSTRACT: Computational chemogenomic (or proteo-chemometric) methods predict target–ligand interactions by training machine learning algorithms on known experimental data in order to distinguish attributes of true from false target–ligand pairs. Many ligand and target descriptors can be used for training and predicting binary associations or even binding affinities. Several chemogenomic studies have not noticed any real benefit in using 3-D structural target descriptors with respect to simpler sequence-based or property-based information. To assess whether this observation results from

inaccurate target description or from the fact that 3-D information is simply not required in chemogenomic modeling, we used a target kernel measuring the distance between target–ligand binding sites of known X-ray structures. When used in combination with a standard ligand kernel in a support vector machine (SVM) classifier, the 3-D target kernel significantly outperforms a sequence-based target kernel in discriminating 2882 target–ligand PDB complexes from 9128 false pairs, whatever the modeling procedure (local or global). The best SVM models could be successfully applied to predict, with very high recall (70%), precision (99%), and specificity (99%), target–ligand associations for an external set of 14 117 ligands and 531 targets. In most of the cases, pooling all data in a global model gave better statistics than just discretizing specific target–ligand subspaces in local models. The current study clearly demonstrates that chemogenomic models taking both ligand and target information outperform simpler ligand-based models. It also permits one to design good modeling practices in predicting target–ligand pairing for a large array of targets: (i) ligand-based models are precise enough if sufficient ligand information (>40–50 diverse ligands) is known; (ii) if not, structure-based chemogenomic models (associating a ligand kernel to a structure-based target kernel) are recommended for proteins of known holostuctures; (iii) sequence-based chemogenomic models (associating a ligand kernel to a sequence-based target kernel) can still be used with a very good accuracy for the remaining targets.

Binding Sites Ligands	lowl	1dm2	2vg7
	0.11	5.51	0.36
	0.01	0.21	0.63
	0.44	0.03	0.72

INTRODUCTION

The ever-increasing availability of target–ligand binding data^{1,2} has significantly modified the scope of computational methods for analyzing and predicting structure–activity relationships. Classical quantitative structure–activity relationships (QSAR) focusing on a particular series of bioactive compounds binding to a single target are unable to extrapolate to different although neighboring chemical and target spaces. Computational chemogenomics³ (also referred to proteochemometrics⁴) circumvents known limitations of pure ligand-based QSAR approaches to predict the binding of several ligands to several targets. Instead of considering only ligand properties, both ligand and target descriptors^{4–6} are used to predict target–ligand associations (binding affinities,^{7,8} complex formation⁹). Since similar receptors are supposed to bind to similar ligands,¹⁰ predicting interactions in a target–ligand interaction matrix can be inferred from known data on similar ligands and/or similar targets. Most chemogenomic approaches have been limited to well-defined target spaces (receptor subtypes or enzyme isoforms),⁴ but extension to larger bioactivity spaces (G protein-coupled

receptors,^{7,9,11} kinases,¹² DrugBank,¹³ PubChem,¹⁴ Target Data Bank,⁸ BindingDB¹⁵) have been reported.

To set up a chemogenomic model, three components are necessary. First, one should carefully choose a data set of experimental data describing target–ligand interactions (three-dimensional structures,¹⁶ binding affinities,^{2,17–20} or simple annotations²¹) and remove target–ligand redundancy. Second, descriptors should be selected for describing ligands and targets.^{4–6} Many descriptors (1-D, 2-D, 3-D) have been reported in the literature,⁴ but no general consensus has been reached to date. It can just be pointed out that 3-D descriptors for either ligands²² or targets^{6,11} did not seem to outperform simpler attributes. Last, a computational method should be chosen for modeling the data. Many methods⁴ (partial least-squares, support vector machines, naïve Bayesian classifier, decision trees, random forest, neural networks, rough set modeling) have been used to model and predict

Received: April 12, 2011

Published: June 06, 2011

chemogenomic data, again with no general consensus on what should be the best approach.

Chemogenomic models have clearly been shown to outperform pure ligand-based models in independent validation studies.^{9,23} This enhancement has been proposed to be mainly due to ligand nearest neighbor effects²³ (high similarity between a target-annotated ligand and a target-orphan compound), although one may argue that the corresponding study was limited to a few related targets. The present study aims at precisely estimating the benefit (or disadvantage) of using an accurate 3-D pocket descriptor in chemogenomic modeling. Up to now, 3-D target descriptors (e.g., fixed-length pharmacophore fingerprint) have mostly been depending on a prior sequence alignment of a consensus set of cavity-lining residues^{9,11} or a full structural alignment of targets.⁶ As far as targets are highly related, the corresponding alignment is straightforward and the resulting similarity score will be relevant. For targets sharing sequence and fold-independent pocket similarities, there is, however, a high risk of underestimating target similarity from a suboptimal alignment.²⁴ Due to the constraints imposed by a 3-D target descriptor (restriction of target space to high-resolution X-ray or NMR target structures) and the paucity of accurate alignment-independent target (binding site) similarity measures,^{24,25} true target–ligand binding site descriptors have not been used in chemogenomic modeling. We herewith report the usage of a real 3-D cavity descriptor (FuzCav)²⁴ recently reported by our group for measuring local binding site similarities among unrelated targets. Briefly, the FuzCav descriptor is a vector of 4834 integers reporting counts of all possible pharmacophoric feature (H-bond acceptor, H-bond donor, positive ionizable, negative ionizable, aromatic, hydrophobic) triplets from binding site-lining residues. Since cavity descriptions and comparisons are fast and independent of a prior 3-D structural alignment, the FuzCav descriptor is perfectly suited to generate a novel target kernel focusing on the most precise information required for describing ligand binding. To be confident in the binding site definition, the current study is limited to high-resolution target–ligand complexes extracted from the sc-PDB data set.²⁶ Using support vector machine (SVM) classifiers, separate kernels for measuring pairwise ligand similarities and target similarities have been investigated. Two conclusions could be drawn from the present report: (i) chemogenomic models clearly outperform simpler ligand-based models, notably when trained on a limited number of ligands (<40), and (ii) a true 3-D cavity descriptor slightly enhances the accuracy of SVM models when compared to a sequence-based target descriptor.

COMPUTATIONAL METHODS

Definitions. A *target* is defined as a macromolecule (protein, nucleic acid) to which a small molecular weight compound (*ligand*) binds to and for which an X-ray structure is available in the Protein Data Bank (PDB).¹⁶ Each target is assigned a unique name (*target name*) according to the UniProt nomenclature.²⁷ For each target–ligand complex, a *binding site* is defined as any residue (amino acid, cofactor, ion) close to the ligand (see precise definition in a previous report²⁶). Target–ligand complexes for which the ligand and the binding site are estimated to be druggable are stored in a subset of the PDB, named sc-PDB.²⁶ sc-PDB targets are clustered according to the pharmacophoric properties of their ligand-binding sites to yield *clusters*. A particular sc-PDB target may thus exhibit different ligand-binding sites located in the same cluster (if sites are

similar) or in different clusters (if sites are dissimilar, e.g. catalytic and allosteric sites). Alternatively, two different targets sharing similar binding sites (e.g., ATP-binding site of protein kinases) may be grouped in the same cluster. In the sc-PDB database, there is no redundancy at the target–ligand complex level, which means that two copies of the same target–ligand complex (obtained at different resolutions, for example) cannot coexist (the complex with the lowest resolution is kept as the single copy).

Training Data Set. Target–ligand interaction data were gathered from the sc-PDB data set of druggable target–ligand X-ray structures.²⁶ In the sc-PDB archive, the ligand is considered from a purely pharmacological point of view (detergents, ions, and molecules devoid of any known biological activity are discarded) and explicitly defines the binding site as any surrounding residue (protein amino acid, cofactor, ion). For each entry, atomic coordinates of the target, the binding site, and the atom type-curated ligand are stored. The in-house developed FuzCav algorithm²⁴ was then utilized to systematically measure the pairwise similarity of all 7078 sc-PDB binding sites (release v 2009) to yield a full similarity matrix that was converted to a distance matrix (distance = 1 – similarity score) and clustered by hierarchical clustering with the average linkage method using a FuzCav similarity threshold of 0.16 as a stopping criterion for cluster agglomeration. Only clusters annotated by more than 10 unique ligands were kept to finally yield 87 clusters describing 2882 sc-PDB complexes (581 different targets and 2605 different ligands). For each cluster, a “*mean ligand i*” was identified as the one that is the most similar to all ligands *j* in the cluster as follows:

$$\text{mean}_i = \sum_j d^2(i, j)$$

with $d = (1 - T_c)$ and T_c = the similarity computed by the Tanimoto coefficient on SciTegic ECFP4 circular 2-D fingerprints.²⁸

The 2882 target–ligand PDB complexes were used as positive instances (true complexes) in the SVM classifications. Negative instances (false complexes) were generated by randomly pairing decoy ligands with binding sites of the above-described 581 targets. Decoy ligands were selected from the in-house Bioinfo-DB database of commercially available druglike compounds.²⁹ For each binding site cluster, a ratio of true to false complexes was fixed (20% true complexes, 80% false complexes). Decoy ligands were selected to span similar physicochemical property ranges (molecular weight and log *P*) than active ligands (the ensemble of ligands in a defined cluster) but to be different enough from the “mean ligand”. Any decoy whose Tanimoto coefficient (ECFP4 fingerprint) was less than 0.15 was selected and considered dissimilar to the “mean ligand”. A final random selection of remaining decoys provided the desired ratio of active ligands (true complexes) to inactive ligands (false complexes).

External Validation Set. Additional ligands of the 581 training set targets were retrieved from six bioactivity databases: ChEMBL_02², PDSPKi,³⁰ Accelrys MDDR 2009.2,³¹ DrugBank 2.5,³² BindingDB,¹⁸ and STITCH 2.0.³³

First, target names in bioactivity databases were compared to that of the sc-PDB target name and kept if identical or very similar. Second, all external ligands of the selected targets were retrieved as either SD files or SMILES strings. For each compound, a maximum of 400 conformers were generated using the default settings of Omega.³⁴ All compounds were ionized with Filter³⁵ and their 2-D structures standardized with the ChemAxon standardizer utility.³⁶

External compounds were then compared to the training sc-PDB ligands with ROCS³⁷ and kept if similar to at least one training compound (Colorescore ≥ 0.5 and Comboscore ≥ 1.2). After checking for duplicates and redundancy with sc-PDB ligands, 14 117 additional ligands for 531 out of the initial 581 targets (60 out of 87 clusters) could be finally retrieved. A total of 328 308 positive instances were created as follows: an external ligand L_e was paired to target T_i of ligand L_i if L_i and L_e were found to be similar enough by ROCS (Colorescore ≥ 0.5 and Comboscore ≥ 1.2). In addition, L_e was also paired to target T_j if T_j shares with T_i an identical target name and cluster number. Negative instances were generated by randomly pairing, for each cluster, drug-like

decoys to targets until the number of positive instances is reached. Decoys were retrieved as previously described for setting up the training set and selected only when different from training set decoys.

Ligand Descriptors. Ligands were represented by a hashed ECFP4 extended-connectivity fingerprint.²⁸ The fingerprint was hashed to a 1024 bit string using a specified hash function in PipelinePilot³⁸ as described by Hert et al.³⁹

Target Descriptors. Three descriptors of increasing complexity were used for the targets: (i) the Uniprot name,²⁷ (ii) the SPECTRUM sequence-based descriptor⁴⁰ registering counts for each of the 20³ possible tripeptides when browsing the amino acid sequence from the N-terminus to the C-terminus, (iii) a 3-D structural descriptor of the binding site as implemented in FuzCav.²⁴ Up to six pharmacophoric features (hydrogen-bond acceptor/donor, positive/negative ionizable, aromatic, aliphatic) were mapped to the C- α carbon atom of each sc-PDB binding site residue. Counts of all possible pharmacophoric triplets within specific distance ranges ([0–4.8 Å], [4.8–7.2 Å], [7.2–9.5 Å], [9.5–11.9 Å], [11.9–14.3 Å]) between C- α atoms were stored in a FuzCav fingerprint of 4834 integers.²⁴

Support Vector Machine (SVM) Classification and Kernels. All models were generated using the SVM^{light} package.⁴¹ The similarity between two target–ligand complexes $\langle c_i, c_j \rangle$ was measured as previously described^{6,11} from a target–ligand kernel

Table 1. Descriptors and Kernels

object	descriptor	kernel	index
ligand	ECFP4 fingerprint	Tanimoto	KL
target	UniProt name	Uniprot	KT1
	Tripeptide occurrence	Spectrum	KT2
	FuzCav fingerprint	FuzCav	KT3
target–ligand complex		Tanimoto \otimes	KTL1
		Uniprot Name	
		Tanimoto \otimes	KTL2
		Spectrum	
		Tanimoto \otimes FuzCav	KTL3

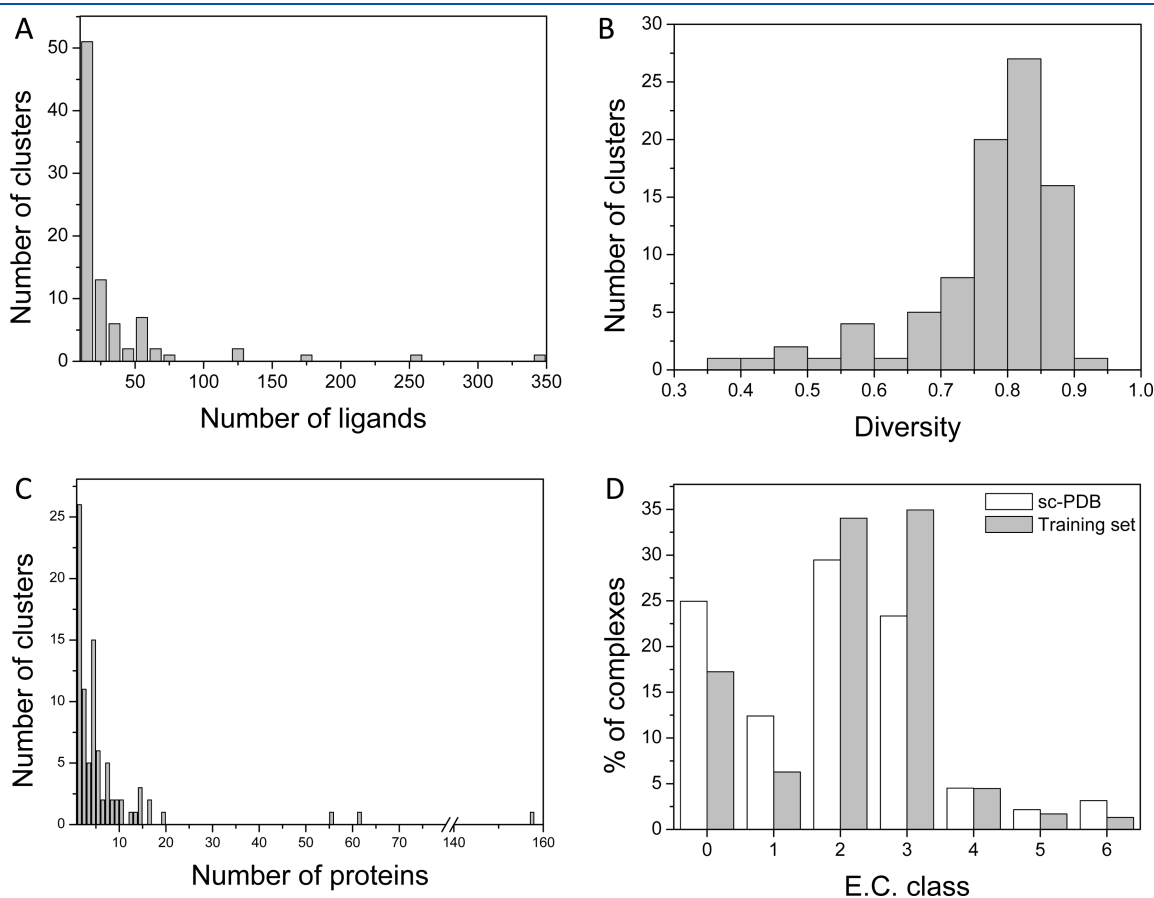


Figure 1. Diversity analysis of the target–ligand training set. (A) Distribution of the number of training set ligands for 87 binding site clusters. (B) Distribution of the chemical diversity of training set ligands for 87 binding site clusters. (C) Distribution of the number of training set targets for 87 binding site clusters. (D) Functional annotation of training set targets by Enzyme Commission (EC) number: 0, no EC number; 1, oxidoreductase; 2, transferase; 3, hydrolase; 4, lyase; 5, isomerase; 6, ligase.

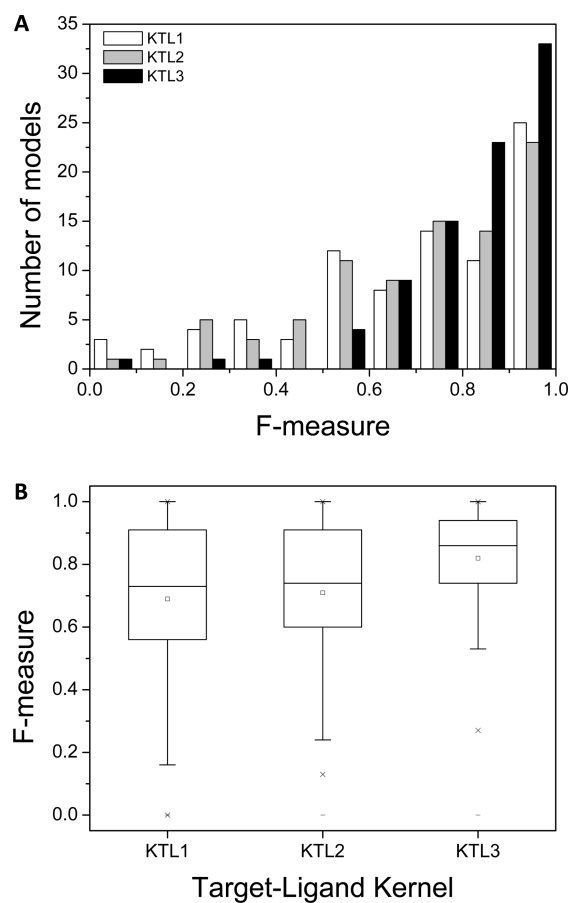


Figure 2. Accuracy, estimated by the *F*-measure, of 87 local SVM models for predicting target–ligand pairing (2882 true pairs and 11528 false pairs) using a pure ligand-based (KTL1) and two chemogenomic approaches based on a target sequence (KTL2) and target 3-D structure (KTL3) kernel. (A) Distribution of the *F*-measure. (B) Box-and-whisker plot of *F*-measure distributions for the three target–ligand kernels. The box delimit 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 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

$K(c_i, c_j)$ as the product of two separate kernels for the target pair $K_{\text{target}}(T_i, T_j)$ and the ligand pair $K_{\text{ligand}}(l_i, l_j)$

$$\langle c_i, c_j \rangle = K(c_i, c_j) = K_{\text{ligand}}(l_i, l_j) \times K_{\text{target}}(t_i, t_j)$$

The Tanimoto kernel was used to calculate the similarity between pairs of ligands represented by their hashed ECFP4 fingerprints.

$$KL = K_{\text{ligand}}(l_i, l_j) = \frac{\langle l_i, l_j \rangle}{\langle l_i, l_i \rangle + \langle l_j, l_j \rangle - \langle l_i, l_j \rangle}$$

Three kernels based on the above-described three target descriptors were utilized to measure pairwise target similarities as follows:

1. *The Uniprot Name Kernel.* The Uniprot name kernel was used to differentiate targets according to their names

$$KT1 = 1 \text{ if } t_i = t_j$$

$$KT1 = 0 \text{ if } t_i \neq t_j$$

This kernel implies that only ligand information is used for each target separately in the learning process. From here

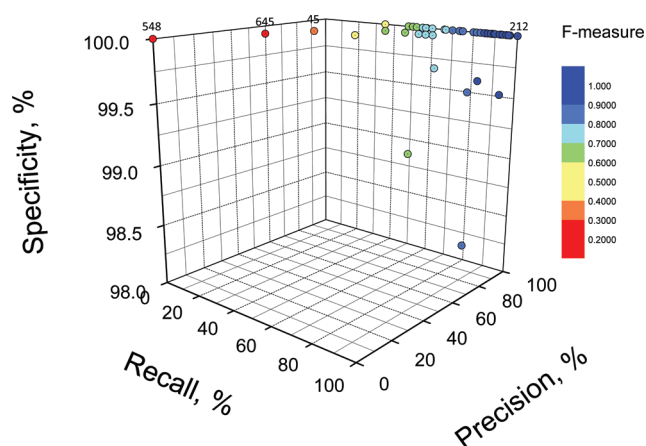


Figure 3. Recall, specificity, and precision of 87 SVM local models using the KTL3 target–ligand kernel. Data are color-coded according to the *F*-measure value. Binding site clusters discussed in the text are labeled according to their number.

on, using this kernel will be referred as a control ligand-based approach.

2. *The Spectrum Kernel.* It is a sequence similarity kernel used for target classification.⁴⁰ Each target is represented by a 20^3 dimensional vector counting occurrences of all possible tripeptides in the sequence. The Spectrum kernel between two targets is then computed as

$$KT2 = \frac{\langle t_i, t_j \rangle}{\sqrt{\langle t_i, t_i \rangle} \sqrt{\langle t_j, t_j \rangle}}$$

3. *The FuzCav Similarity Kernel.* Targets are represented by their FuzCav descriptors, and the similarity between two cavities is measured by a standard Tanimoto coefficient as follows:

$$KT3 = K_{\text{target}}(t_i, t_j) = \frac{\langle t_i, t_j \rangle}{\langle t_i, t_i \rangle + \langle t_j, t_j \rangle - \langle t_i, t_j \rangle}$$

Table 1 shows a summary of all descriptors and kernels used in this study.

A 5-fold cross-validation procedure was used to split each of the 87 clusters five times into a training (four-fifths of the data set) and a test set (one-fifth of the data set) and analyze the predictivity of SVM models on the remaining test sets, using the best trade-off *C* value optimized for each model

Model Evaluation. Statistical parameters for evaluating the different SVM models were the recall, precision, specificity, and *F*-measure:

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

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

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

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

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

RESULTS AND DISCUSSION

Composition and Analysis of the Training Data Set. The current paper specifically aims at unambiguously evaluating the benefit of using true 3-D binding site information in chemogenomic classification models of target–ligand binary associations.

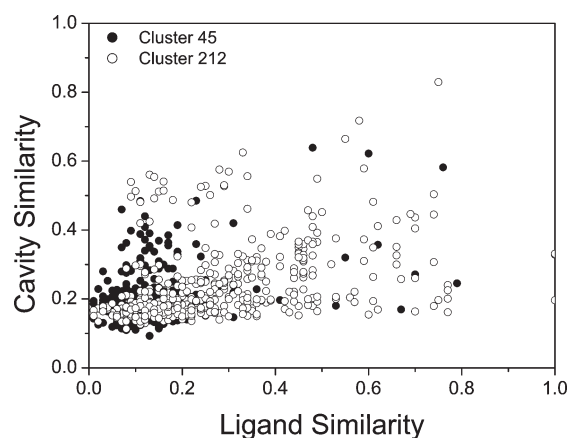


Figure 4. Pairwise ligand and cavity similarities for target–ligand complexes yielding poor (cluster 45, F -measure = 0.48) and perfect (cluster 212; F -measure = 1.0) SVM classification models. Cluster 45 comprises seven targets and 28 ligands, whereas cluster 212 is composed of five targets and 24 ligands (see Supporting Information Table 1). Similarities are computed according to the Tanimoto coefficient on hashed ECFP4 ligand fingerprints and FuzCav cavity fingerprints (see Computational Methods).

Since the target kernel used in this study focuses on ligand-binding site similarity and local models were derived for target clusters of similar binding sites, the present analysis is constrained to a reduced set of 87 clusters covering 581 unique targets for which (i) a high-resolution X-ray structure of a ligand-bound state is available and (ii) at least 10 ligands have been cocrystallized with similar binding sites. There is no real consensus about the minimal number of ligands required to describe an activity class. In one of the most exhaustive studies published to date,⁴² a threshold of five different ligands was used for druglike compounds. Since PDB ligands are notoriously less diverse than druglike compounds, we therefore chose a slightly higher value of 10 compounds.

Most of the 87 target clusters are paired with less than 20 compounds (Figure 1A) and only four clusters (see a complete description of cluster contents in Supporting Information Table 1) are populated by more than 100 different compounds. The corresponding binding sites and target families have either been heavily investigated (e.g., cluster 30, HIV-1 protease catalytic site; cluster 57, serine/threonine target kinase ATP-binding site; cluster 50, trypsin-related catalytic site) or exhibit ligand promiscuity (cluster 5, nucleotide-binding sites). The diversity of each training ligand set, estimated as the mean pairwise intermolecular dissimilarity,⁴³ is remarkably high (Figure 1B). Seventy-two out of the 86 training sets show a diversity above 0.7. Since target space has been discretized by binding site diversity, the number of unique targets in each cluster varies considerably from a single target (26 out of 87 clusters) to a maximum of 157 (cluster 5). A very large majority of clusters regroup less than five different targets (Figure 1C). The annotation of the target training set according to the Enzyme Commission (EC) number⁴⁴ reveals no major change with respect to the sc-PDB data set, suggesting that our selection of target–ligand complexes has not introduced any major bias toward a particular target space (Figure 1D).

Cross-Validation of SVM Models on the Training Data Set Clearly Indicates a Superior Performance of Chemogenomic with Respect to Ligand-Based Models. Prediction of target–ligand pairing (2882 true pairs; 11 528 decoys) was realized using

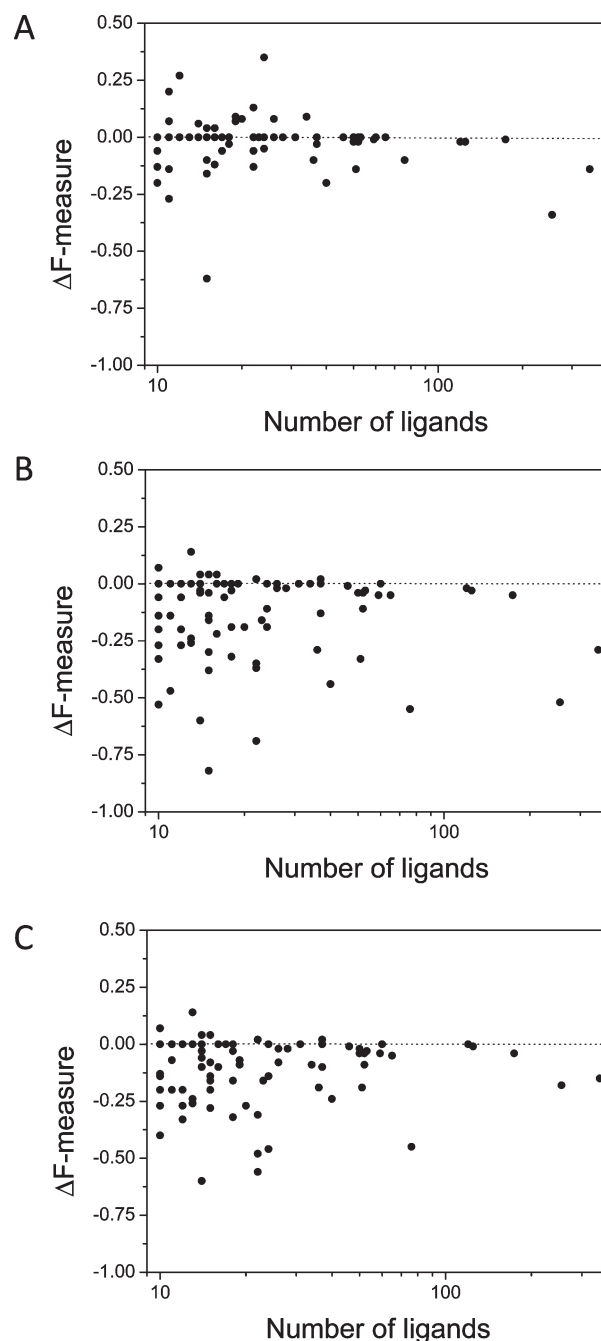


Figure 5. Difference in the F -measure value (ΔF -measure) of three chemogenomic models (KTL1, KTL2, KTL3) in classifying 2882 true target–ligand PDB complexes and 11528 target–ligand decoys: (A) $F(KTL1) - F(KTL2)$, (B) $F(KTL1) - F(KTL3)$, and (C) $F(KTL2) - F(KTL3)$. Values are plotted according to the number of true ligands for each binding site cluster.

87 local SVM models and a 5-fold cross-validation protocol. Each local model addresses targets whose binding sites are grouped in the same cluster, their corresponding PDB ligands, and the related decoys. The accuracy of all models was estimated from the F -measure parameter (see Computational Methods), which presents the advantage to take both recall and precision into account. Varying the target kernel on this data set unambiguously demonstrates that the best models were obtained using a 3-D

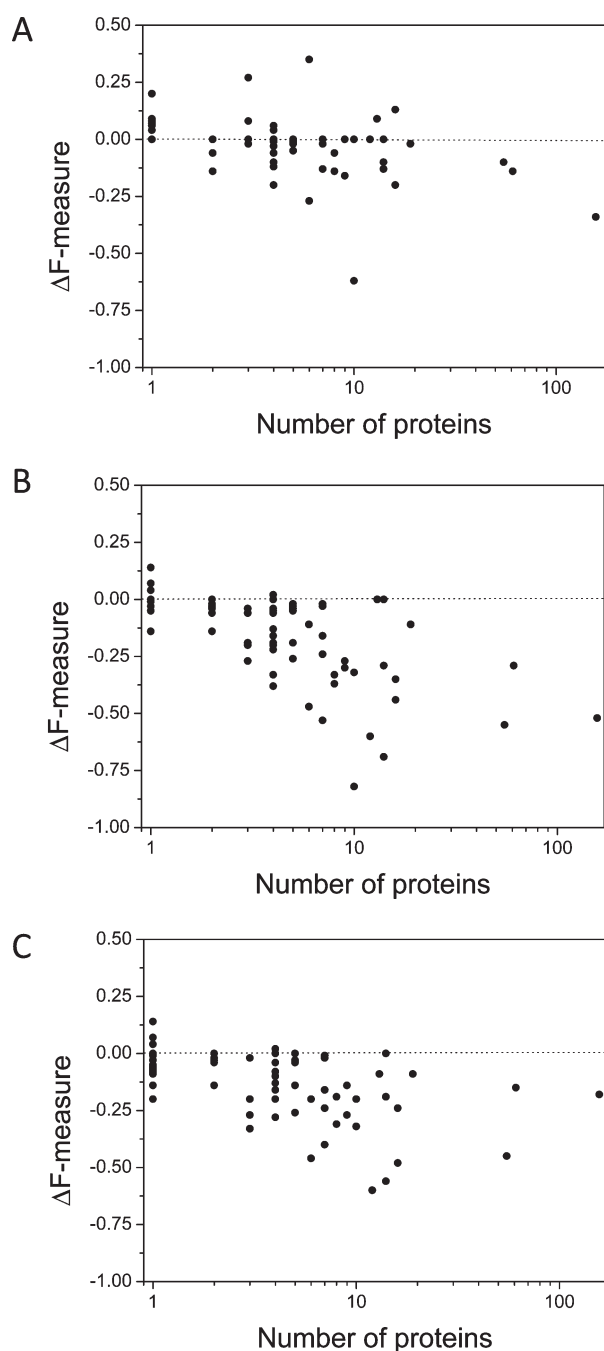


Figure 6. Difference in the F -measure value (ΔF -measure) of three chemogenomic models (KTL1, KTL2, KTL3) in classifying 2882 true target–ligand PDB complexes and 9128 target–ligand decoys: (A) $F(\text{KTL1}) - F(\text{KTL2})$, (B) $F(\text{KTL1}) - F(\text{KTL3})$, (C) $F(\text{KTL2}) - F(\text{KTL3})$. Values are plotted according to the number of different targets for each binding site cluster.

binding set kernel (KTL3), then with a target sequence-based kernel (KTL2), and last with a purely ligand-based approach (Figure 2). Considering an F -measure value above 0.5 as acceptable, only 3 out of 87 models (models 45, 548, and 645; see complete results in Supporting Information Table 2) are unsatisfactory (Figure 3). The low F -measure value observed for these three models are mostly attributable to low recall values, the specificity being still always above 90% (Figure 3). Inspecting

Table 2. Performance of a Global and Local Models (F -measure) on the Cross-Validated Training Set

model	local		global
	mean	median	
KTL1	0.689	0.730	0.750
KTL2	0.709	0.740	0.900
KLT3	0.819	0.860	0.910

the training ligands and cavities of these difficult clusters (e.g., cluster 45) reveals much poorer pairwise ligand and pairwise cavity similarities with respect to clusters (e.g., cluster 212), yielding nearly perfect SVM classification models (Figure 4). Failure of SVM models to recall true target–ligand complexes belonging to these three clusters may thus be attributable first to an incomplete coverage of the corresponding target–ligand space by existing PDB complexes and then to the promiscuity of the corresponding binding sites toward different chemotypes.

We next carefully inspected cross-validation results for the three target–ligand kernels and all 87 clusters, notably the impact of using target information either in the form of amino acid composition (KLT2 kernel) or binding site 3-D structure (KTL3 kernel). Plotting the difference of the F -measure for each local model clearly shows a benefit of combining ligand and target descriptors for almost all models. Interestingly, the benefit tends to be more important for models trained with a limited number of positive instances, whatever the kernel used (Figure 5). Likewise, chemogenomic models also profit from the number of targets taken into account (Figure 6). Single target models are not suitable for the kind of approach developed herein, a benefit being seen when considering more than four or five targets (Figure 6). Conversely to previous studies^{6,11} that did not noticed any advantage of using target 3-D information with respect to much simpler descriptors, the present report demonstrates a significant superiority of the target kernel (KT3) measuring binding site 3-D similarities with respect to the sequence-based Spectrum KT2 kernel (Figures 5 and 6). The observed discrepancy to previous results may be explained by three important factors. First, target space has been here discretized by a novel approach considering binding site 3-D similarity, irrespective of the target name and family. Second, it is the first time to the best of our knowledge that a true binding site 3-D descriptor (FuzCav),²⁴ which has shown its ability to discriminate true similar from true dissimilar cavities, has ever been applied to chemogenomic modeling. Third, the target–ligand data set has been restricted to target–ligand complexes of known X-ray structures in which a pharmacological ligand binds to a druggable cavity.²⁶

Assuming that similar receptors bind similar ligands,¹⁰ it is therefore logical that focusing on the target binding sites at the most precise level provides a true advantage in predicting target–ligand associations. Of course, we acknowledge that the applicability domain of the current approach is limited to a tiny target–ligand space and may exclude important targets (e.g., membrane receptors) for which precise 3-D structural information is missing. The main advantage of using local models is that results may be examined for different target–ligand subspaces to infer possible guidelines for best practice chemogenomic modeling. Of course, all positive and negative instances may be pooled into a single SVM model. A 10-fold cross-validation procedure was applied to the entire data set and recall, selectivity, specificity, and F -measure

of the global model were determined as previously described for the local model. Overall statistics were in favor of a global modeling procedure, whatever the target kernel used (Table 2). Notably, the

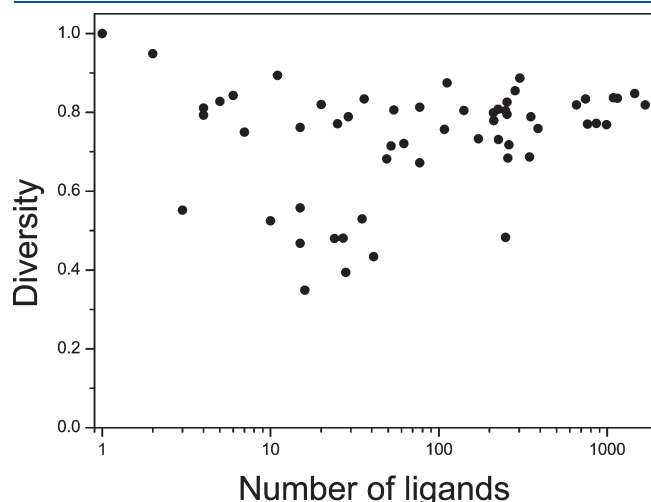


Figure 7. Diversity of the 60 external test sets, estimated from the mean pairwise intermolecular dissimilarity⁴³ of the corresponding ligand ECFP4 fingerprints, as a function of their size (number of ligands).

sequence-based target model (KTL2) appears as good as the structure-based kernel (KTL3) in a global SVM modeling.

We, however, acknowledge that cross-validated models described here may exhibit overestimated statistics due to the decoys selection protocol. Selecting decoys is indeed a tricky process for which many routes may be followed. The selection based on ECFP4 similarity to a mean ligand was just done to be sure that actives and decoys do not share the same scaffolds. What is important is that the models are extensively challenged by external test sets and that selectivity, specificity, and precision remain acceptable. We therefore designed the largest possible external ligand set to challenge the cross-validated models.

Large-Scale Prediction of Target–Ligand Associations from an External Data Set. The best models derived after cross-validation may not be the most predictive when applied to an external data set.^{9,45} Likewise, the superiority of the KTL3 kernel noticed in the cross-validation study may disappear when applied to an external test set. We therefore decided to validate the previously reported models, using exactly the same three target–ligand kernels, on the largest possible external data set. For each of the 581 sc-PDB targets, six external ligand sources (ChEMBL,² PDSPKi,³⁰ MDDR,²¹ DrugBank,³² BindingDB,¹⁸ and STITCH 2.0.³³) were browsed to retrieve 14 117 compounds for 531 targets of the training set. Since, target space is organized by binding site and not target name, we had to verify

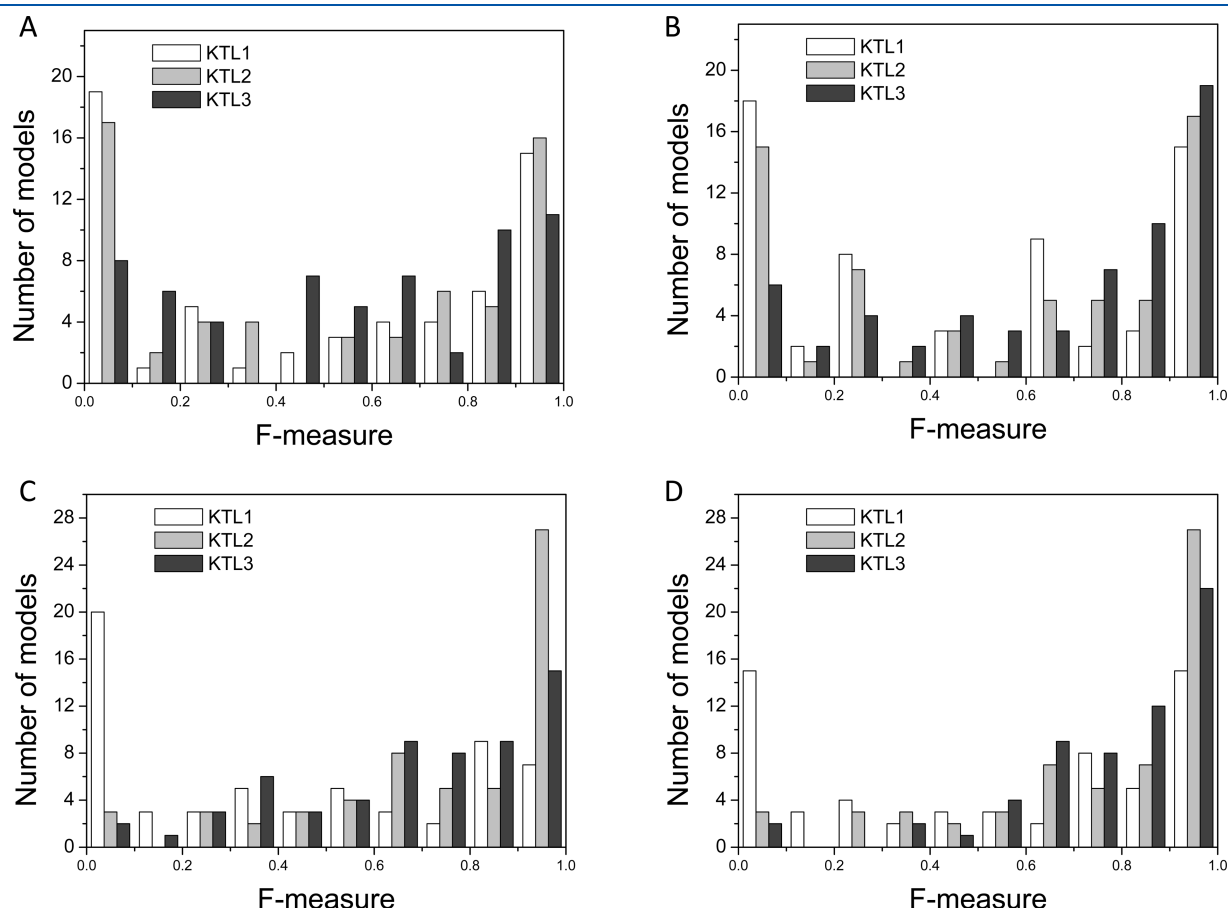


Figure 8. Distribution of the *F*-measure values for three SVM target–ligand kernels (KTL1, KTL2, KTL3) predicting target–ligand binary association (PDB entry prediction mode) for an external test set (14 114 ligands, 531 targets, 328 308 true complexes seeded with the same number of false target–ligand decoys): (A) local models, prediction by PDB entry; (B) local models, prediction by target binding site; (C) global model, prediction by PDB entry; (D) global model, prediction by target binding site.

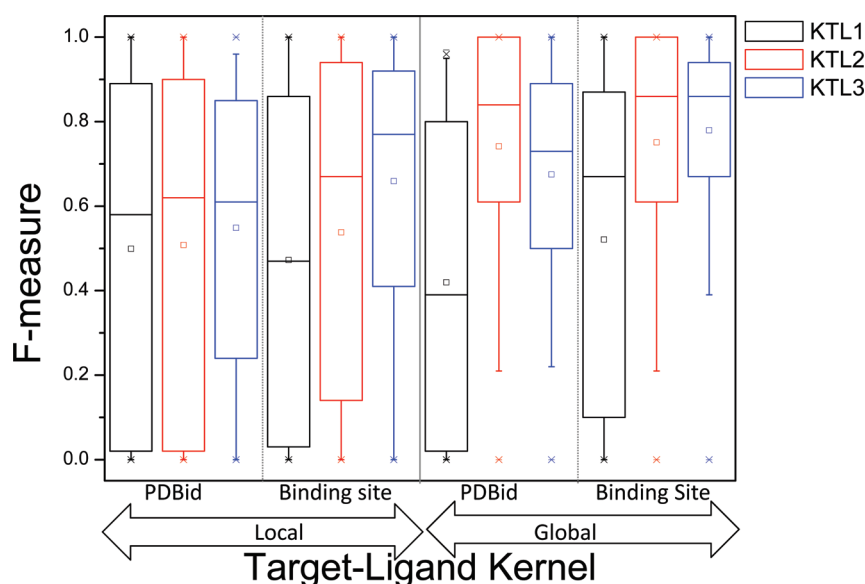


Figure 9. Box-and-whisker plot of *F*-measure distributions for three SVM target–ligand kernels (KTL1, KTL2, KTL3) predicting target–ligand binary association (PDB entry prediction mode) for an external test set (14 114 ligands, 531 targets, 328 308 true complexes seeded with the same number of false target–ligand decoys). The box delimit 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 1% and 99th percentiles, respectively. Minimum and maximum values are indicated by a dash.

that both training and external ligands are likely to share the same binding site by computing their pharmacophore-annotated shape similarity and only to retain pairs for which the similarity, as estimated by ROCS, is above a certain threshold (Colorescore ≥ 0.5 and Comboscore ≥ 1.2). For each of the 60 remaining target clusters, an external test set of ligands could be defined exhibiting in most of the cases more than 10 compounds and an acceptable chemical diversity (Figure 7). In the current external validation, the ratio of positive to negative instances was fixed to 1 for all clusters. Modification of this ratio by varying the number of target–ligand decoys did not affect the obtained results. Target–ligand pairing was predicted in two possible modes, by PDB entry and by target binding site.

Predicting target–ligand binding on a PDB entry basis corresponds to answer the following question: “To which PDB entry (e.g., 1dm2) is this compound predicted to bind to?” For many entries, however, multiple copies of the same target binding site are available. For example, cluster 57 regroups 347 PDB entries out of which 99 correspond the ATP binding site of the cell division protein kinase 2 (CDK2). Predicting the binding of a true CDK2 inhibitor from the test set may fail for some of the 99 entries but succeed for some others. Fusing the results for each target binding site thus enables one to escape from particular binding site singularities (e.g., site-directed mutagenesis, induced fit, amino acid omission in the PDB entry). In the second prediction mode, a target–ligand pairing was then predicted for each target binding site; in other words, to answer the question, “To which binding site (e.g., CDK2 ATP-binding site) is this compound predicted to bind to?” In order to directly compare predictions from local models with that from the global one, the 60 external test sets were iteratively used for predicting target–ligand pairing with the global model.

Four main observations could be drawn by analyzing the prediction statistics (Figures 8 and 9 and Table 3):

- Chemogenomic models (KTL2, KTL3) almost always outperform the ligand-based model (KTL1 kernel), as evidenced by the higher proportion of models with *F*-measure higher than 0.5, the narrower distribution, and the higher mean and median values,
- Among chemogenomic models, using a structure-based kernel (KTL3) leads to slightly better predictions than a sequence-based kernel (KTL2).
- Global models perform better than local models.
- As expected from the data fusion, predicting target–ligand pairing on a target binding site basis is preferred to a PDB entry basis, whatever the modeling procedure (local or global).

Out of the 12 prediction modes (see complete results in Supporting Information Tables 3–6), the best predictions are obtained using a global model with a structure-based kernel (KTL3) and predicting pairing on a target binding site basis. Since precision and specificity are excellent (99%) for all 60 external test sets (see complete results in Supporting Information Table 6), the predictive property of the global model is therefore almost dependent on the recall capability (71%). Altogether, an *F*-measure value higher than 0.5 was obtained for 42 out of the 60 external test sets. Examining the recall values against the diversity and the size (number of external ligands) of the external test sets did not reveal clear trends. Considering the unprecedented diversity and size of the external test set (>14 000 ligands and 531 targets), the performance of this chemogenomic model is remarkable and suggests the use of cavity 3-D kernels whenever possible to predict target–ligand pairing. Using a sequence-based target kernel in a global model also provides very good statistics, but with a wider distribution of *F*-measure values (Figure 9) and slightly more poor models (*F*-measure < 0.5) with respect to the structure-based kernel (Figure 8). The accuracy of the target-sequence kernel is however quite promising, notably for predicting ligand pairing to

Table 3. Performance of a Global and Local Models (F-measure) on the External Test Set

model	local		global	
	mean	median	mean	median
KTL1	0.499	0.575	0.419	0.385
KTL2	0.508	0.595	0.741	0.840
KLT3	0.548	0.605	0.675	0.725
KTL1	0.472	0.460	0.521	0.680
KTL2	0.538	0.655	0.751	0.860
KLT3	0.659	0.770	0.779	0.865

targets of unknown 3-D structure, and therefore considerably extends the applicability domain of chemogenomic QSAR modeling.

Altogether, we therefore recommend different possible strategies with respect to the question that has to be answered. If one wants to exactly know to which protein a ligand may bind, it is better to use a global model (KTL3 kernel), which gives a probability of target–ligand association for every target of our data set. However, if the issue is to know to which kind of binding site (e.g., ATP-binding site of Ser/Thr protein kinases, catalytic site of of serine endopeptidases) a ligand (or a focused library) may bind to, it is better to use local models focusing on well-defined target spaces. Moreover, the present study permits one to design good chemogenomic modeling practices with respect to the existing knowledge on targets and their ligands. Taking into account target information (sequence or structure) makes sense only when the number of known ligands for this peculiar target is sufficiently low (roughly below 40–50). When this is the case, using structural information about the ligands binding site clearly provides an advantage with respect to simpler sequence information in predicting novel target–ligand associations. Since these conclusions are similarly drawn by models derived from the training set and more importantly from the external test set, we therefore propose a pragmatic *in silico* target profiling strategy taking advantage of three possible situations. In the first one (3-D structure of the target is available and less than 50 ligands are known), we propose to combine a ligand descriptor and a 3-D binding site descriptor in separate kernels. In scenario 2 (3-D structure of the target is not available and less than 50 ligands are known), we propose to combine a ligand descriptor and a target sequence descriptor in separate kernels. Last, in the case where more than 50 ligands for a particular target are known, we propose to use a simple ligand similarity kernel and a SVM classification model. These proposals are based on the statistics (mainly the *F*-measure) derived from models applied to the external data set. Like any model, the conclusions are of course partly dependent on the input data and therefore the decoy ligand selection and known actives. We, however, believe that the general trends indicated in the present study are data-set-independent, notably because of the very large external test set used to challenge the current SVM models.

CONCLUSIONS

Chemogenomic (or proteochemometric) QSAR modeling methods are taking an increasing importance in predicting, at a very high throughput, the binding of numerous ligands to numerous targets. A key advantage of chemogenomic modeling with respect to ligand centric methods is the applicability to orphan targets or at least to targets for which ligand information is sparse.

By looking at known data on the neighboring target–ligand space, novel target–ligand relationships may be inferred. Various approaches for predicting target–ligand binary association or binding affinities have been proposed and shown to be remarkably efficient and predictive. Whatever the method, a target–ligand space to which it is applicable must be defined. Defining this space will strongly influence both the methods and descriptors used to describe targets and their ligands. An exhaustive definition (e.g., any biologically relevant target with more than five ligands) implies a relatively fuzzy and rough definition of target space, usually at the amino acid sequence level, since precise information on binding site location and target–ligand interactions are missing. Conversely, a restricted definition as the one used in the current study (any biologically relevant target cocrystallized with drug-like compounds) considerably limits the applicability range of the method, but it enables a fine modeling of target cavity attributes responsible for ligand binding. In the current report, we unambiguously demonstrate that target binding site descriptors, when available, enhance the performance of chemogenomic methods in predicting target–ligand binary associations, with respect to simpler sequence-based target attributes and pure ligand-based modeling. The proposed method, despite its accuracy, still suffers from two drawbacks: (i) it is only applicable to 531 targets (mostly enzymes) of known high-resolution X-ray structure and (ii) it just predicts the likelihood of target–ligand binding but not a binding affinity nor functional effects (agonist vs antagonist, competitive vs noncompetitive inhibition). Extending the approach to a much larger target–ligand space (e.g., membrane receptor and their ligands) requires either changing target descriptors (e.g., sequence-based attributes), although we know this change may be slightly detrimental to the model accuracy, or following a more pragmatic target-based approach utilizing various models for different target–ligand spaces (e.g., 3-D binding site kernels for PDB targets and sequence-based kernels for other targets). One may even imagine varying the property to predict according to known data (binding affinities or target–ligand association) for particular target–ligand subspaces. In many instances (e.g., biogenic amine G protein-coupled receptor ligands), so many data are available that chemogenomic methods are not required or even not suitable, as evidenced by the present report, for accurate predictions. Along with the ever increasing amount of public target–ligand binding data, we believe that target class-specific methods, descriptors, and property predictions will enhance the applicability of ligand profiling methods to a large array of biologically relevant targets and propose usable computational preclinical safety profiles in early drug discovery phases.

ASSOCIATED CONTENT

S Supporting Information. Target–ligand binding site annotations (cluster id, target name, number of ligands, number of targets), model statistics (recall, precision, specificity, *F*-measure) for the 87 cross-validated training sets and the 60 external test sets, and mapping external database to sc-PDB target names. This material is available free of charge via the Internet at <http://pubs.acs.org>

AUTHOR INFORMATION

Corresponding Author

*E-mail: rognaun@unistra.fr.

■ ACKNOWLEDGMENT

We thank the Conseil Régional d'Alsace for a grant to J.M. The Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) and the GENCI (Project x2010075024) are acknowledged for allocation of computing time. We sincerely thank Dr. I. Baskin, Dr. N. Weill, and Dr. J. P. Vert for much advice and critical reading of the manuscript.

■ REFERENCES

- (1) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–33.
- (2) Overington, J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput. Aided Mol. Des.* **2009**, *23*, 195–8.
- (3) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.
- (4) van Westen, G. J. P.; Wegner, J. K.; Ijzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm* **2011**, *2*, 16–30.
- (5) Vert, J. P.; Jacob, L. Machine learning for in silico virtual screening and chemical genomics: New strategies. *Comb. Chem. High Throughput Screen* **2008**, *11*, 677–85.
- (6) Wassermann, A. M.; Geppert, H.; Bajorath, J. Ligand prediction for orphan targets using support vector machines and various target–ligand kernels is dominated by nearest neighbor effects. *J. Chem. Inf. Model.* **2009**, *49*, 2155–67.
- (7) Bock, J. R.; Gough, D. A. Virtual screen for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1402–14.
- (8) Strombergsson, H.; Daniluk, P.; Kryshchavych, A.; Fidelis, K.; Wikberg, J. E.; Kleywegt, G. J.; Hvidsten, T. R. Interaction model based on local protein substructures generalizes to the entire structural enzyme–ligand space. *J. Chem. Inf. Model.* **2008**, *48*, 2278–88.
- (9) Weill, N.; Rognan, D. Development and validation of a novel protein–ligand fingerprint to mine chemogenomic space: Application to G protein-coupled receptors and their ligands. *J. Chem. Inf. Model.* **2009**, *49*, 1049–62.
- (10) Klabunde, T. Chemogenomic approaches to drug discovery: Similar receptors bind similar ligands. *Br. J. Pharmacol.* **2007**, *152*, 5–7.
- (11) Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J. P. Virtual screening of GPCRs: An in silico chemogenomics approach. *BMC Bioinf.* **2008**, *9*, 363.
- (12) Lapins, M.; Wikberg, J. E. Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. *BMC Bioinf.* **2010**, *11*, 339.
- (13) Nagamine, N.; Shirakawa, T.; Minato, Y.; Torii, K.; Kobayashi, H.; Imoto, M.; Sakakibara, Y. Integrating statistical predictions and experimental verifications for enhancing protein–chemical interaction predictions in virtual screening. *PLoS Comput. Biol.* **2009**, *5*, e1000397.
- (14) Ning, X.; Rangwala, H.; Karypis, G. Multi-assay-based structure–activity relationship models: Improving structure–activity relationship models by incorporating activity information from related targets. *J. Chem. Inf. Model.* **2009**, *49*, 2444–56.
- (15) Strombergsson, H.; Lapins, M.; Kleywegt, G. J.; Wikberg, J. E. S. Towards proteome-wide interaction model using the proteochemometrics approach. *Mol. Inf.* **2010**, *29*, 499–508.
- (16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–42.
- (17) Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G. AffinDB: A freely accessible database of affinities for protein–ligand complexes from the PDB. *Nucleic Acids Res.* **2006**, *34*, D522–6.
- (18) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–201.
- (19) Roth, B. L.; Sheffler, D. J.; Kroeze, W. K. Magic shotguns versus magic bullets: Selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discovery* **2004**, *3*, 353–9.
- (20) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–80.
- (21) Jagarlapudi, S. A.; Kishan, K. V. Database systems for knowledge-based discovery. *Methods Mol. Biol.* **2009**, *577*, 159–72.
- (22) Mahe, P.; Ralaivola, L.; Stoven, V.; Vert, J. P. The pharmacophore kernel for virtual screening with support vector machines. *J. Chem. Inf. Model.* **2006**, *46*, 2003–14.
- (23) Geppert, H.; Humrich, J.; Stumpfe, D.; Gartner, T.; Bajorath, J. Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 767–79.
- (24) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein–ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–35.
- (25) Yeturu, K.; Chandra, N. PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinf.* **2008**, *9*, 543.
- (26) Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: An annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46*, 717–27.
- (27) Apweiler, R.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Alam-Farouque, Y.; Antunes, R.; Barrell, D.; Bely, B.; Bingley, M.; Binns, D.; Bower, L.; Browne, P.; Chan, W. M.; Dimmer, E.; Eberhardt, R.; Fedotov, A.; Foulger, R.; Garavelli, J.; Huntley, R.; Jacobsen, J.; Kleen, M.; Laiho, K.; Leinonen, R.; Legge, D.; Lin, Q.; Liu, W. D.; Luo, J.; Orchard, S.; Patient, S.; Poggioni, D.; Pruess, M.; Corbett, M.; di Martino, G.; Donnelly, M.; van Rensburg, P.; Bairoch, A.; Bougueleret, L.; Xenarios, I.; Altairac, S.; Auchincloss, A.; Argoud-Puy, G.; Alselsen, K.; Baratin, D.; Blatter, M. C.; Boeckmann, B.; Bolleman, J.; Bollondi, L.; Boutet, E.; Quintaje, S. B.; Breuza, L.; Bridge, A.; deCastro, E.; Ciapina, L.; Coral, D.; Coudert, E.; Cusin, I.; Delbard, G.; Doche, M.; Dornevil, D.; Roggli, P. D.; Duvaud, S.; Estreicher, A.; Famiglietti, L.; Feuerhahn, M.; Gehant, S.; Farriol-Mathis, N.; Ferro, S.; Gasteiger, E.; Gateau, A.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hulo, N.; James, J.; Jimenez, S.; Jungo, F.; Kappler, T.; Keller, G.; Lachaize, C.; Lane-Guermonprez, L.; Langendijk-Genevaux, P.; Lara, V.; Lemerrier, P.; Lieberherr, D.; Lima, T. D.; Mangold, V.; Martin, X.; Masson, P.; Moinat, M.; Morgat, A.; Mottaz, A.; Paesano, S.; Pedruzzi, I.; Pilboud, S.; Pillet, V.; Poux, S.; Pozzato, M.; Redaschi, N.; Rivoire, C.; Roehert, B.; Schneider, M.; Sigrist, C.; Sonesson, K.; Staehli, S.; Stanley, E.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Veuthey, A. L.; Yip, L. N.; Zuletta, L.; Wu, C.; Arighi, C.; Arminski, L.; Barker, W.; Chen, C. M.; Chen, Y. X.; Hu, Z. Z.; Huang, H. Z.; Mazumder, R.; McGarvey, P.; Natale, D. A.; Nchoutmboube, J.; Petrova, N.; Subramanian, N.; Suzek, B. E.; Ugochukwu, U.; Vasudevan, S.; Vinayaka, C. R.; Yeh, L. S.; Zhang, J.; Consortium, U. The universal protein resource (UniProt) in 2010. *Nucleic Acids Res.* **2010**, *38*, D142–D148.
- (28) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–54.
- (29) <http://bioinfo-pharma.u-strasbg.fr/bioinfo> (accessed June 16, 2011).
- (30) <http://pdsp.med.unc.edu> (accessed May 20, 2011).
- (31) <http://accelrys.com/products/databases/bioactivity/mddr.html> (accessed May 20, 2011).
- (32) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–6.
- (33) Kuhn, M.; Szklarczyk, D.; Franceschini, A.; Campillos, M.; von Mering, C.; Jensen, L. J.; Beyer, A.; Bork, P. STITCH 2: An interaction network database for small molecules and proteins. *Nucleic Acids Res.* **2010**, *38*, D552–6.

- (34) *Omega*, version 2.4.3; OpenEye Scientific Software: Santa Fe, NM.
- (35) *Filter*, version 2.1.1; OpenEye Scientific Software: Santa Fe, NM.
- (36) *Standardizer*, version 5.5.0.1; ChemAxon Kft.: Budapest, Hungary.
- (37) *ROCS*, version 3.0.0; OpenEye Scientific Software: Santa Fe, NM.
- (38) *Pipeline Pilot*, version 7.5; Accelrys Software Inc.: San Diego, CA.
- (39) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
- (40) Leslie, C.; Eskin, E.; Noble, W. S. The spectrum kernel: A string kernel for SVM protein classification. *Pac. Symp. Biocomput.* **2002**, 564–75.
- (41) *SVMlight*, version 6.02, <http://svmlight.joachims.org/> (accessed May 20, 2011).
- (42) 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.
- (43) 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.
- (44) <http://www.chem.qmul.ac.uk/iubmb/> (accessed May 20, 2011).
- (45) Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* **2010**, *29*, 476–488.