JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naïve Bayes and Parzen-Rosenblatt Window

Alexios Koutsoukas,[†] Robert Lowe,[†] Yasaman KalantarMotamedi,[†] Hamse Y. Mussa,[†] Werner Klaffke,[‡] John B. O. Mitchell,[§] Robert C. Glen,*[,†] and Andreas Bender*[,†]

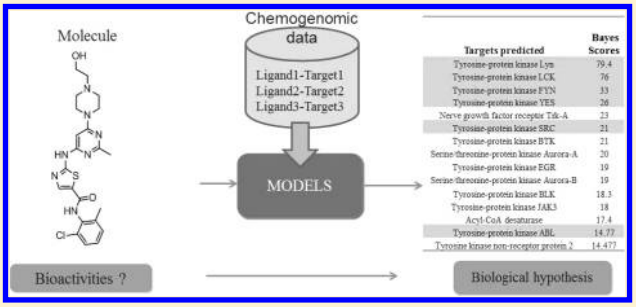
[†]Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

[‡]Unilever R&D Vlaardingen, Olivier van Noortlaan 120, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands

[§]Biomedical Sciences Research Complex and EaStCHEM School of Chemistry, Purdie Building, University of St Andrews, North Haugh, St. Andrews, Scotland KY16 9ST, U.K.

Supporting Information

ABSTRACT: In this study, two probabilistic machine-learning algorithms were compared for in silico target prediction of bioactive molecules, namely the well-established Laplacian-modified Naïve Bayes classifier (NB) and the more recently introduced (to Cheminformatics) Parzen-Rosenblatt Window. Both classifiers were trained in conjunction with circular fingerprints on a large data set of bioactive compounds extracted from ChEMBL, covering 894 human protein targets with more than 155,000 ligand-protein pairs. This data set is also provided as a benchmark data set for future target prediction methods due to its size as well as the number of bioactivity classes it contains. In addition to evaluating the methods, different performance measures were explored. This is not as straightforward as in binary classification settings, due to the number of classes, the possibility of multiple class memberships, and the need to translate model scores into "yes/no" predictions for assessing model performance. Both algorithms achieved a recall of correct targets that exceeds 80% in the top 1% of predictions. Performance depends significantly on the underlying diversity and size of a given class of bioactive compounds, with small classes and low structural similarity affecting both algorithms to different degrees. When tested on an external test set extracted from WOMBAT covering more than 500 targets by excluding all compounds with Tanimoto similarity above 0.8 to compounds from the ChEMBL data set, the current methodologies achieved a recall of 63.3% and 66.6% among the top 1% for Naïve Bayes and Parzen-Rosenblatt Window, respectively. While those numbers seem to indicate lower performance, they are also more realistic for settings where protein targets need to be established for novel chemical substances.



INTRODUCTION

Phenotypic screening is increasingly being used in drug discovery for the identification of compounds that elicit a therapeutically beneficial response in diseased cells, organs, or model organisms. Novel "first-in-class" therapeutics are already being discovered *via* this route, compared to single target-based screening.[1] In many cases phenotypic screening can provide accurate information and understanding of the effects that a drug has on complex biological systems,[2] taking properties such as absorption-distribution-metabolism-exertion (ADME) properties but also intracellular signaling networks into account. However, in phenotypic screens target identification is often a nontrivial step which follows the discovery of small molecules that can elicit a biological phenotype.[2] The problem has been termed target "deconvolution" and is an important aspect of current drug discovery processes, as knowledge of molecular targets can significantly aid in the discovery of new chemical entities and rationalize the actions of compounds.[3]

A number of experimental technologies are currently available and can be employed to elicit the underlying molecular target involved in complex biological mechanisms, such as affinity chromatography-based methods that measure the strength of binding of proteins to ligands attached to resins, or methods based on expression of mRNA from cDNA libraries, which allows for the identification of genes that encode proteins targets.[3] Nevertheless, these methods are often time- and resource-intensive, making it unfeasible to test all possible ligand-target interactions, especially when little or no previous knowledge of potential molecular targets is available.

The problem of unintended secondary protein targets for a drug is of considerable practical relevance, since a significant number even of marketed drugs present off-target interactions that are responsible for a wide range of adverse drug

Received: September 13, 2012
Published: July 8, 2013

reactions (ADRs). This relates both to compounds in development, many of which fail to reach the market due to unforeseen side-effects,[4,5] and also to reassessment of marketed drugs with clinically observed side-effects. Well-documented examples here include the case of Cerivastatin,[6] which was withdrawn due to drug-induced rhabdomyolysis; and e.g. the examples of cisapride, terfenadine, astemizole, sertindole, and grepafloxacin which were withdrawn due to undesirable interaction with the hERG channel.[7] The goal is now to comprehensively anticipate drug-target interactions early on, in order to direct experimentation toward the most likely targets first.

The relationship between chemical structures and biological activities against protein targets, which is at the core of current medicinal chemistry efforts, has been an active field of research in recent decades. The concept of chemical similarity, which states that structurally similar compounds are more likely than not to exhibit similar properties, has been used as the basis of many "rational" drug design approaches.[8,9] Taking into consideration the exponential growth of bioactivity databases and the wealth of information stored in them (databases such as PubChem,[10] ChEMBL,[11] and KiDB[12]), it is now becoming feasible to map the known "chemical space" and "biological space" into models that will enable us to generate biological "spectra" which can serve as hypotheses for the prediction of phenotypic activity of new molecules, based on their chemical structures and the known bioactivities of structurally similar compounds. Some of the first attempts to map global pharmacological space based on chemical structural similarity were reported by Paolini et al.[13] and Fabian et al.[14] Those studies have contributed significantly to our understanding of the extent of polypharmacology present in pharmacological space, both regarding the extent of polypharmacology encountered in bioactive compounds and also by relating particular targets (and their ligands) to each other.

Making use of the bioactivity data available today, the focus of the work presented here is on predicting protein targets, based on experimental bioactivity data available from databases, and describing ligands active at these targets in computational models in an appropriate way. While a number of computational target prediction approaches exist,[15] such as those based on docking[16] or pharmacophores,[17] in this work we will restrict analysis to ligand-based target prediction methods which utilize fingerprints combined with machine learning approaches.

In silico methods can exploit prior knowledge of ligand-target interactions, collected e.g. from literature and/or patents, which is organized in chemogenomic databases (in particular where the use of standardized target names or identifiers is of crucial importance). These data are then computationally analyzed in order to make knowledge-based predictions for new, untested molecules or to suggest new drug-target interactions for already marketed compounds. By combining known structure activity relationships (SARs) and machine learning methods, it is possible to explore a large chemogenomic space (spanning up to the order of a million compounds and thousands of targets) and to identify patterns present in chemical space to generate a mechanism of action (MOA) hypothesis for a previously unseen molecular structure.[13,18] An example of such a prediction is presented in Figure 1 for Dasatinib, a multi-BCR/ABL and Src family tyrosine kinase inhibitor approved for use in patients with chronic myelogenous leukemia (CML). Highlighted proteins indicate correct target predictions generated by the models, while additional predicted targets are potentially new interactions, which can then be tested in experiments. The interested reader is referred to a recent comprehensive review on in silico target prediction, bioactivity databases, and multitarget drug design by Koutsoukas et al.[19]

In recent years, a number of target prediction methodologies have been suggested by training machine learning algorithms on bioactivity data, represented by suitable descriptors, in order to develop predictive models. Nidhi et al.[20] applied a multiclass Naïve Bayes trained model on chemogenomic space extracted from the WOMBAT database and applied this to predict the top three most likely protein targets for 10 MDDR (MDL Drug Database Report) activity classes, managing in 77% of the cases to predict the correct protein targets of the compounds. Later, Nigsch et al.[21] compared the performance of the Winnow algorithm against Naïve Bayes using 20 pharmaceutical activity classes extracted from WOMBAT database. It was found that, although both algorithms achieved similar performance on

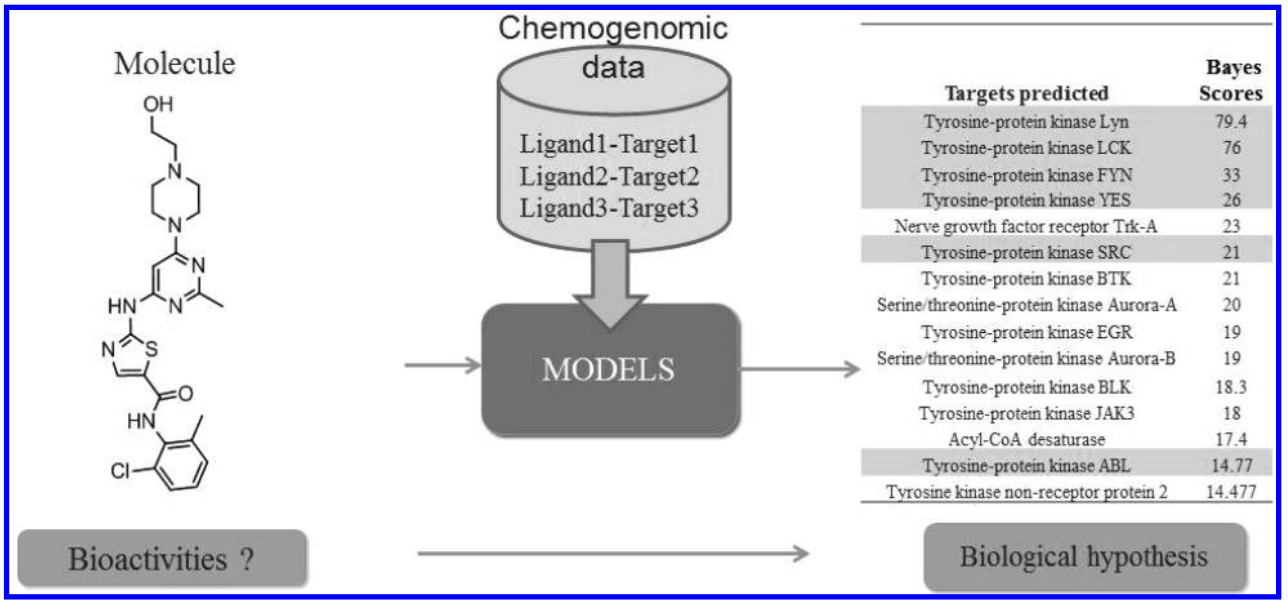


Figure 1. A protein target hypothesis generated by the multiclass Naïve Bayes for the kinase inhibitor dasatinib. Experimentally established targets of dasatinib are highlighted, while the remaining targets are potential new interactions which need to be verified (or disproved) experimentally.

average, differences were observed in performance among classes and among individual structures. More recently, Wale et al.[22] compared the performance of Naïve Bayes, ranking SVM and Perceptron methods using SARs extracted from the PubChem database. The data set used in that study covered 231 targets and 27,205 compounds. In that study recall and precision among top-k positions were used to assess the performance of the models, with SVMs achieving the best performance among the algorithms. Here, the main finding was that nonlinear classifiers outperformed linear methods such as Naïve Bayes; however, considering the computational complexity and number of classes, such methods as SVMs and Perceptron, might be too complex when dealing with thousands of targets and hundreds of thousands (or even millions) of compounds. Keiser et al.[23,24] developed a statistical model which relates biological targets based on ligand similarities and ranks the significance of the resulting similarity scores. The method was named Similarity Ensemble Approach (SEA), which employs a BLAST[25]-derived algorithm to develop minimal spanning trees considering chemical similarity.

On the application side, a number of studies have successfully demonstrated that by data mining available structure-activities relationships (SARs) it is possible, within the applicability domains both in chemical and biological space, to predict the bioactivities of small molecules in a practically relevant context.[23,24,26,27] Examples of successful applications of in silico target prediction methodologies include target deconvolution of antitubercular compounds,[28] the identification of off-target inhibition of the enzyme protein farnesyltransferase (PFTase) by the marketed drugs Loratadine and Miconazol,[29] as well as the rationalization of false positive readouts in reporter gene assays and the phenotypic readouts of high-content screening data.[27] More recently, Lounkine and Keiser et al.[30] performed a large scale prediction and testing of drug activity on side-effect targets. More than 600 marketed drugs were computationally screened against a set of 73 protein targets, and approximately half of the positive predictions were subsequently experimentally confirmed.

Computationally, target prediction can be viewed as a category-ranking problem, and given the large number of classes (of the order of 1000), the uneven distribution of data points (molecules) in the classes, the different chemical diversity per ligand activity class, and often subjective bioactivity cutoffs (a K_d or IC_{50} value of e.g. 10 μM may possess pharmacologically different significance, depending on the assay employed), it is a challenging one. The goal is still to develop a model, based on the known chemogenomic space, that a given test molecule is able to rank potential protein targets by their likelihood of binding to the molecule being analyzed.

In this study, large scale human ligand-protein predictive models were developed based on bioactivity data extracted from ChEMBL,[31] with several aims in mind. First, the performance of the well established Laplacian-modified multiclass Naïve Bayes classifier was compared with the Parzen-Rosenblatt Window method that was more recently introduced in the cheminformatics area[21,32] in the context of target prediction. Finally, identifying and quantifying factors that may affect the performance of target prediction models were investigated, in particular class size and intraclass chemical diversity.

In addition to the computational aspects above, an important point regarding the benchmarking of target prediction methods is that currently no common benchmark data set is available in the public domain selected for this purpose. Hence, by choosing the ChEMBL database as the source of structure–activity relationships for this study (and by providing all compound identifiers in the Supporting Information) the first publicly available data set is provided that enables proper comparison of the performance of new in silico target prediction methodologies to existing ones.

MATERIALS AND METHODS

Data Set. In this study ChEMBL (version 10) was used, containing more than a million annotated compounds, comprising more than 4 million bioactivities covering more than 8,000 targets, all abstracted from primary scientific literature.[11,31] Compounds with reported activities (K_i/K_d/IC50/EC50) equal or better than 10 μM against human protein targets with a confidence score of 8 or 9 were extracted and used for subsequent model generation. These rules were used to ensure that compounds showing reasonably measurable effects with more reliable target annotations were selected for the study. While the cutoff for bioactivities of equal to or better than 10 μM may represent very active compounds in some assays, this will also include marginally active compounds. However, given the constraints of data availability, this value represents a suitable trade-off between biologically relevant activity and data set size (and hence coverage of chemical space). A confidence score is provided in ChEMBL as a measure of confidence in the assay-to-target relationships represented in the database. Confidence scores of 8 and 9 represent cases where homologous single proteins or direct single protein targets are assigned, respectively.[11] The SARs tables extracted were further processed, and ligand structures were standardized using Pipeline Pilot[35] "strip salts" and "standardize molecules" components, with options "standardize stereo" and "standardize charges" checked on. Ambiguous reported SARs identified (e.g., cases were relation of compound - binding affinities reported with ">" or being "null" or annotated with conflicting binding affinities active/inactive against protein targets) were discarded and not taken into further consideration as they represented a small fraction of data points (in our case less than 1%). Next, utilizing ligands' canonical SMILES representations generated based on standardized structures, duplicate ligands annotated multiple times against a protein target were merged to retain only unique ligand-protein pairs per protein target. Moreover, only protein classes with at least 20 compounds were used in order to ensure that enough information is available to describe the chemical space associated with each protein target. The total number of structures extracted includes 105,946 compounds, covering 894 human protein targets (provided in the Supporting Information) with a total of 155,208 ligand-target pairs. No attempt was made to further categorize the data (e.g., to distinguish between agonists and antagonists or competitive or allosteric inhibitors etc., as this type of information is rarely present for reported small molecules); instead all compounds with binding capability were considered together. Table 1 shows the number of annotated targets per compound and Figure 2 the distribution of targets among protein classes. It can be seen that a significant number of compounds in the current data set are annotated against multiple protein targets; more than 20% of all compounds. Protein kinases and enzymes represent 34% and 30%, respectively, of biological targets included in the data set, indicating the significance of these targets in modern pharmacology and also the prevalence of kinases as drug targets in the past decade.

Table 1. Number of Annotated Protein Targets Per Compound[a]

parameter	value									
no. of compds	*82931*	*15166*	*4379*	*1654*	*709*	*267*	*160*	*109*	*96*	*390*
% of total	*78.340%*	*14.326%*	*4.137%*	*1.562%*	*0.670%*	*0.252%*	*0.151%*	*0.103%*	*0.091%*	*0.368%*
annotated targets	1	2	3	4	5	6	7	8	9	≥10

[a]Although the majority of the compounds in the data set extracted from ChEMBL are anotated against one protein target (78%), a significant fraction of compounds (22%) are anotated against two or more (even >10) protein targets.

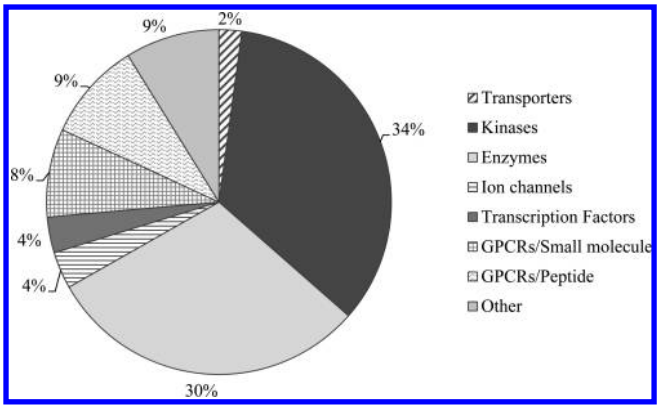


Figure 2. Protein target class distribution in the training data set employed in the current study. It can be seen that the majority of targets for which bioactivity data are available are kinases and enzymes, with GPCRs following next. Despite the similar number of data points in the model, given that enzyme inhibitors are structurally more diverse than kinase inhibitors, the performance for predicting enzymes as targets is still significantly better than for predicting which kinases a particular inhibitor is able to interact with.

Although ChEMBL constitutes a comprehensive resource for the deduction of SARs, it is (as are necessarily all resources of this type) far from being a complete representation of the chemical space reported in the literature. Creating databases of this type is a complex and difficult task. As a recently reported analysis of commercial and public bioactivity databases reported, data for specific molecules may be abstracted from different sources (giving rise to incompatible or different values) and are subject to transcription errors.[33] It was reported that inconsistencies between databases are common even when the *same* sources are used for data extraction, often due to human error or interpretation of results. Similar observations were reported in an earlier work,[34] where it was pointed out that the observed differences found across databases often result from different data selection and extraction strategies. Therefore integration/comparison of data from additional databases could significantly enrich the chemical and biological space used to train models and subsequently improve and expand the applicability of such predictive models, in both chemical and biological space.

MOLECULAR DESCRIPTORS

Extended Connectivity Fingerprint (ECFP) descriptors were used as implemented in Pipeline Pilot Student Edition 6.1.34 ECFP descriptors belong to the class of circular topological substructural fingerprints and are based on a modified version of the Morgan algorithm.[35,36] For this study the ECFP_4 descriptors were used as they have previously been demonstrated to effectively capture chemical structural information relevant to bioactivity as shown in studies.[21,37,38] All molecules were standardized prior to fingerprint generation in PipelinePilot by using the "strip salts" and "standardize molecules" components, with the options "standardizestereo" and "standardizecharges" checked in the latter.[39] The total number of unique ECFP_4 features generated for this data set was 103,428.

Target Prediction Algorithms. *A.) Parzen-Rosenblatt Window (PRW).* The first target algorithm employed here is the Parzen-Rosenblatt Window (PRW) based method.[32,40] For each possible protein target ω_α, the similarity of a test molecule, $\boldsymbol{x}$, to all the training molecules $\boldsymbol{x}_j$ in the class ω_α is calculated by the kernel function $K(\boldsymbol{x},\boldsymbol{x}_j;h)$. For each class ω_α, it is assumed that $p(\boldsymbol{x}|\omega_\alpha)$, the class-conditional probability density (mass) function for molecule $\boldsymbol{x}$ given class ω_α, is directly proportional to this average similarity as described in eq 1

$$p(\boldsymbol{x}|\omega_\alpha) = \frac{1}{N_{\omega_\alpha}} \sum_{\boldsymbol{x}_j \in \omega_\alpha} K(\boldsymbol{x}, \boldsymbol{x}_j; h) \tag{1}$$

where N_{ω_α} denotes the number of the training data points in class ω_α, and h is the smoothing parameter.

Using Bayes' theorem, the class a posteriori probability of a new molecule $\boldsymbol{x}$ being associated with a specific protein target $p(\omega_\alpha|\boldsymbol{x})$ can be calculated according to

$$p(\omega_\alpha|\boldsymbol{x}) = \frac{p(\omega_\alpha)p(\boldsymbol{x}|\omega_\alpha)}{p(\boldsymbol{x})} \tag{2}$$

The prior class probability $p(\omega_\alpha)$ is assumed to be equivalent to the proportion of molecules in that class for the training set, $p(\omega_\alpha) = (N_{\omega_\alpha})/(N)$, where N refers to total number of instances in the training set; and $p(\boldsymbol{x})$ can be expressed as $\sum_{\beta=1}^{L}((N_{\omega_\beta})/(N))p(\boldsymbol{x}|\omega_\beta)$ with L being the number of classes.

The targets are ordered with decreasing probability, $p(\omega_\alpha|\boldsymbol{x})$, for molecule $\boldsymbol{x}$. In the case of $p(\omega_\alpha|\boldsymbol{x}) = p(\omega_\beta|\boldsymbol{x})$, the targets are ranked arbitrarily.

A Gaussian kernel for $K(\boldsymbol{x}, \boldsymbol{x}_j; h)$ is utilized, which is given as

$$K(\boldsymbol{x}, \boldsymbol{x}_j; h) = \frac{1}{(\sqrt{2\pi}h)^d} \exp\left(-\frac{\left\|x - x_j\right\|^2}{2h^2}\right) \tag{3}$$

Here, h is the smoothing parameter, and d is the dimension of the feature vectors $\boldsymbol{x}$ and $\boldsymbol{x}_j = \{f_1, f_2, ..., f_d\}$, where f_i denote the features. Interested readers are referred to the works of Hand et al.[41] and Bishop et al.[42] for further information on Kernel Density Estimation methods. The smoothing parameter was optimized on the data set used, based on recall of true positives (TP) in top-1 position, and it was found to be $h = 0.9$ (Supporting Information Figure S1).

B.) Multiclass Laplacian-Modified Naïve Bayes (NB). The Laplacian-Modified Naïve Bayes Classifier was implemented according to Nigsch et al.,[21] which is summarized briefly here. Starting from Bayes' rule, eq 2, and making the assumption that features f_i are independent, $p(\omega_\alpha|\boldsymbol{x})$ can be estimated according to

$$p(\omega_\alpha|\boldsymbol{x}) = \frac{1}{p(\boldsymbol{x})} p(\omega_\alpha) \prod_{i=1}^{d} p(f_i|\omega_\alpha) \quad (4)$$

The prior on each class was set as before, $p(\omega_\alpha) = (N_{\omega_\alpha})/(N)$ where N_{ω_α} is the number of instances in class ω_ω and N is the total number of instances. Let $N^+_{i\omega_\alpha}$ be the total number of occurrences of feature f_i in ω_α and N^+_i the total number of occurrences of feature f_i in the whole of the training set, then the noncorrected class conditional probability of an instance is given by

$$p(f_i|\omega_\alpha) = \frac{N^+_{i\omega_\alpha}}{N^+_i} \quad (5)$$

From eq 5 it can be seen that if $N^+_{i\omega_\alpha} = 0$, then $p(f_i|\omega_\alpha) = 0$ and hence $p(\omega_\alpha|\boldsymbol{x}) = 0$, independent of all other $p(f_i|\omega_\alpha)$. The solution desired is that in the limit when feature f_i is rarely present, then $p(f_i|\omega_\alpha) = p(\omega_\alpha)$. Therefore, if a feature is sampled D more times it would be expected that $p(\omega_\alpha) * D$ samples to belong to class ω_α. Hence, by adding D virtual samples eq 5 becomes

$$p'(f_i|\omega_\alpha) = \frac{N^+_{i\omega_n} + D^*p(\omega_\alpha)}{N^+_i + D} \quad (6)$$

then in the limit as $N^+_{i\omega_\alpha} \to 0$ and $N^+_i \to 0$

$$\lim_{N^+_{i\omega_\alpha}, N^+_i \to 0} p'^{(f_i|\omega_\alpha)} = \lim_{N^+_{i\omega_\alpha}, N^+_i \to 0} \frac{N^+_{i\omega_\alpha} + D^*p(\omega_\alpha)}{N^+_i + D} = p(\omega_\alpha) \quad (7)$$

In the Laplace correction, $D = p(\omega_\alpha)^{-1}$, and to produce a relative probability we divide by $p(\omega_\alpha)$ to give

$$p_{rel}(f_i|\omega_\alpha) = \frac{p'(f_i|\omega_\alpha)}{p(\omega_\alpha)} = \frac{N^+_{i\omega_\alpha} + 1}{N^+_i{}^*p(\omega_\alpha) + 1} \quad (8)$$

The logarithm is used to avoid numerical problems when values become small, hence the score for class ω_α of a new molecule $\boldsymbol{x}$, can be calculated as

$$S_{\omega_\alpha}(\boldsymbol{x}) = \sum_i f_i \log\left(\frac{N^+_{i\omega_\alpha} + 1}{N^+_i{}^*p(\omega_\alpha) + 1}\right) \quad (9)$$

where f_i refers to the binary-valued ith feature for the test molecule $\boldsymbol{x}$.

NB was implemented in C# and the PRW in python (version 2.7). Implementations of both algorithms will be made publicly available with this study in the Supporting Information.

Assigning Target Classes Based on Ranking. In the first part of the work, the performance of each method was measured by taking a particular number of the highest ranks into account. To this end, the annotated targets of each ligand-target instance in the database were considered as having a "true" label; subsequently the performance was evaluated by measuring the positions where the "true" labels were retrieved in the ranked lists of targets for a compound. 5-fold cross-validation (5FCV) was applied, and the recall was calculated in the top-k (k = 3, 6, 9, 27, 54, 72, 90, and 180) positions of the ranked target list averaged over the 5-folds. This approach was pursued since in real-world applications only a small number of top-k most likely targets for a given compound possibly are tested experimentally.

Performance per Protein Class. In order to investigate how prediction performance varies between protein targets, the performance per protein class was investigated. Variation in class size and structural diversity of compounds describing each protein class were assumed to affect the performance of each algorithm in correctly identifying instances of each class in the 5-fold cross-validation. Furthermore, overlap of classes was expected to decrease model performance, especially in the protein superfamily of kinases, where strong evolutionary conservation of the binding sites of ATP is present and developing selective inhibitors remains a challenge.[43,44] For this analysis, mean and median ranking per class averaged over 5-fold cross-validation were calculated to measure prediction performance on a per-class basis and to subsequently investigate the influence of data set size and structural diversity.

External Validation Data Set. In order to assess predictive performance of the models, an external data set of bioactive molecules from the WOMBAT (WOrld of Molecular BioAcTivity version 2011.1)[46] database was assembled. This approximates the situation when such predictive methods are applied to generate biological hypotheses for molecules external to the bioactive molecules used for training the models (i.e., novel scaffolds). The external data set was extracted from WOMBAT (version 2011.1) according to the following criteria: I) Only binders were considered (K_i/IC50/K_d/EC50) with 10 μM or better. II) Only compounds with Tanimoto similarity of less than 0.8 (measured in ECFP_4 fingerprint space) to the training set of compounds from ChEMBL were included, with the most similar compounds in the range of Tc of 0.8 and most dissimilar in the range of 0.2; these had a mean of Tc 0.575 and median 0.583. In total 46,306 active compounds were used for testing, covering approximately 530 therapeutically important human protein targets (provided in the Supporting Information). Performance was established by measuring recall in the 9 and 27 top positions (which correspond to the top 1% and 3% of the total number of targets, respectively).

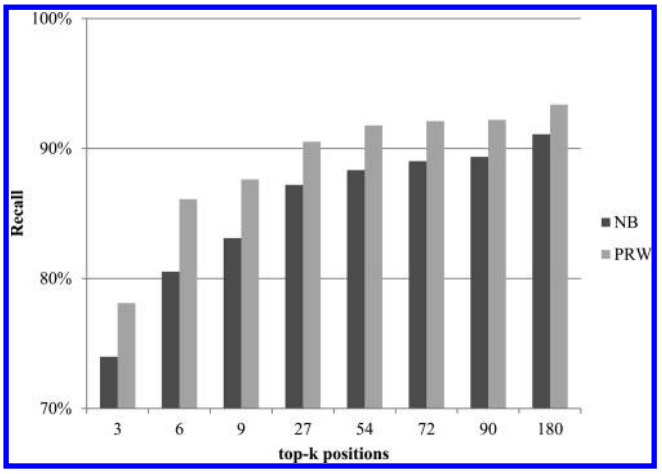


Figure 3. Recall achieved by the Naïve Bayes (NB) and Parzen-Rosenblatt Window (PRW) method in the top 3, 6, 9, 27, 54, 72, 90, and 180 positions, respectively. PRW achieved higher recall than NB among all measured top-k positions. Recall exceeded 90% by PRW in top-27 (top 3%) on this data set, while similar performance was measured for NB in the top 90 positions (almost top 10%), showing that PRW is able to retrieve correct targets higher on the ranked list. Experimentally, that could be interpreted as meaning that using PRW one would have to test fewer protein targets to identify the protein targets against which a test molecule is active.

dx.doi.org/10.1021/ci300435j | J. Chem. Inf. Model. 2013, 53, 1957−1966

Table 2. Recall in the Top-k Positions Achieved by Two Methods (Parzen-Rosenblatt Window (PRW) and Naïve Bayes (NB)[a]

top-k positions	PRW	NB
3	78%	74%
6	86%	80.5%
9	87.6%	83.1%
27	90.5%	87.2%
54	91.8%	88.3%
72	92.1%	89.0%
90	92.2%	89.4%
180	93.4%	91.1%

[a]In order to achieve 90% recall, PRW requires only 27 positions to be included, while NB needs to include about 90 positions.

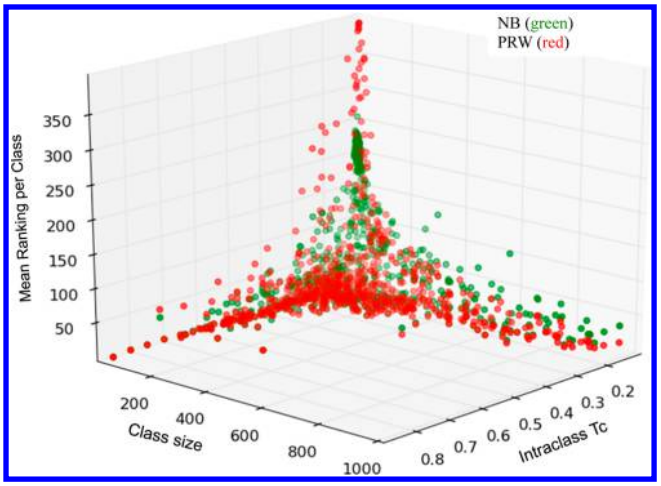


Figure 4. Effect of class size and intraclass Tanimoto similarity (in ECFP_4 space) on the performance of Parzen-Rosenblatt Window (PRW) and Naïve Bayes (NB) classifiers, respectively. Classes with intraclass Tc ≥ 0.4 perform overall well and are retrieved among the top 100 or better positions (shown on the axis on the left). Similar performance was observed among classes with a large number of ≥200 data points. On the contrary, classes with a small number of data points and low intraclass Tc (middle bottom section) are more difficult to predict and present large variation in performance (distribution on vertical *z*-axis). In the case of PRW, 55% of the classes were measured with mean ranking among top 3%, while in the case of NB the corresponding figure was 47%. A point's color intensity serves as an indication of its position in 3D space; lighter shades indicate classes located in the background.

RESULTS AND DISCUSSION

Assigning Target Classes Based on Ranking. Results obtained by considering a given number of top-ranked targets as positive predictions are presented averaged over 5-folds in Figure 3 and Table 2, with every fold containing approximately 31,000 ligand-protein pairs (one-fifth of the total number of data points employed). It can be observed that both algorithms performed rather well overall, achieving in the top nine positions (corresponding to 1% of the total number of targets) a recall higher than 80%, with PRW achieving a recall of 87% and NB achieving a recall of 83%. While those numbers seem rather similar, there is also another way to interpret them, which better illustrates the difference in performance: namely, the PRW method achieved (in the top 27 positions) a recall of 90%, while to achieve a very similar recall (89.35%) in the case of NB the top 90 positions are required; so more than three times the number of positions as is required for PRW. Our results in the case of NB are also close to those reported by Nidhi et al.,[20] where in the top three positions, on average 77% of the correct targets were correctly predicted for compounds for 10 MDDR activity classes.

Performance per Class: Influence of the Number of Data Points and Chemical Diversity. Although both methods demonstrated good overall performance, which exceeded a recall of 80% in the top 27 positions, performance per class exhibited large variations. In the following, the top 27 positions were considered as a cutoff for comparison between classes, which represents approximately the top 3% of total number of targets (out of 894). Using this performance measure, PRW was able to achieve a mean ranking better than this threshold for 493 out of 894 protein classes (55.15%), while NB was only able to retrieve slightly fewer than 419 classes (49.9%) (Supporting Information).

In order to assess the influence of the number of data points and the data set diversity on per-class performance, the model performance was assessed based on those two variables. For this purpose, both class size and average intraclass Tanimoto similarity coefficient were calculated utilizing ECFP_4 fingerprints (Figure 4; more data are provided in Supporting Information Model Statistics). The results show that both class size and intraclass structural similarity significantly affect the performance of these target prediction algorithms. Classes with more than 200 data points perform significantly better overall than smaller ones; in the case of PRW the average mean rank for classes with more than 200 data points was 13.48 and for NB 26.35, respectively (219 classes in the data set), while in the

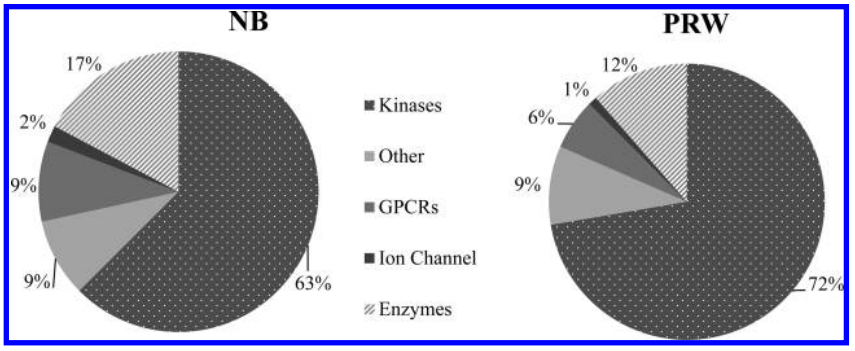
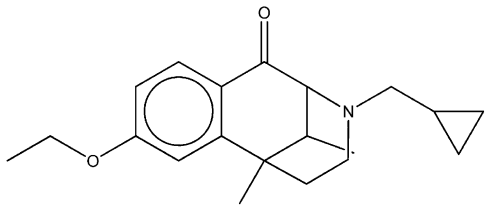


Figure 5. Distribution of protein targets found among classes with "poor performance" (defined as classes with a mean ranking below the top 27 positions (top 3%)). In the case of the Parzen-Rosenblatt Window (PRW), the total number of poorly predicted classes was measured to be 401, while for Naïve Bayes this number is 475. The majority of the "difficult" classes are kinases, which constitute 72% of the cases for PRW and 63% for NB, followed by enzymes with 17% and 12%, respectively.

Ethylketazocine (EKT)

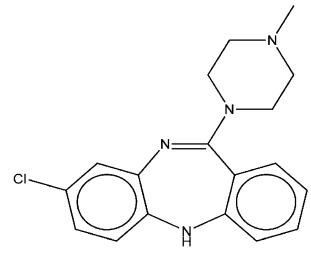
A)



PRW predicted targets	PRW scores	NB predicted targets	NB scores
OPRM1	5.00E-01	OPRK1	57.9396817
OPRK1	5.00E-01	OPRM1	55.7881506
OPRD1	1.14E-18	OPRD1	41.8024668
NQO1	8.18E-27	SIGMAR1	11.3543285
PTPRC	1.14E-37	KCND3	10.6352099
CDC25B	1.14E-37		

Clozapine

B)



PRW predicted targets	PRW scores	NB predicted targets	NB scores
HTR6	1.00E+00	DRD1	31.55594067
HTT	7.19E-09	DRD2	20.60258295
HRH4	1.39E-11	FGFR1	19.73198779
AHR	5.18E-17	PDGFRB	17.50642553
PARP1	5.40E-17	HRH4	15.03831259
CDK1	2.27E-18	HRH1	13.22001758
IDO1	2.04E-19	HTR2A	12.95348615
CDC25B	2.00E-19	ALK	12.48778456
SIGMAR1	1.00E-19	AURKC	12.09882538
ERCC5	9.99E-20	CSF1R	11.6631539
CYP1A2	9.99E-20	HTR7	11.41717679
MAP2K1	9.99E-20	RET	11.08990942

Figure 6. Examples of predictions generated by PRW and NB for compounds EKT and clozapine. A) ethylketazocine (CHEMBL278086), a kappa opioid receptor agonist, for which both algorithms generated good predictions and retrieved the main targets of the opioid family and B) clozapine (CHEMBL42), an atypical antipsychotic agent, for which NB generated better predictions and retrieved the main targets, DRD 1 and 2, among the top positions. The two methods retrieve surprisingly different lists of targets, depending on the particular structure; hence they both have their value when analyzing the mechanism of action (MOA) of compounds where phenotypic information is known.

case of the latter much more variation in performance was observed. On the other hand, small classes (<100 data points) with low intraclass structural similarity (Tc < 0.2 using ECFP_4 fingerprints) had a mean rank of 342.4 and 160 for PRW and NB, respectively (280 classes in total), illustrating that diverse bioactivity classes are apparently difficult to capture, in combination with circular fingerprints. On the other hand, classes with larger intraclass similarity (Tc >0.2) were measured with a mean rank of 22.51 and 18.86 for PRW and NB, respectively, underlining the influence of compound diversity on performance.

When examining which protein classes are difficult to predict, it becomes apparent that kinases represent the majority of the less successful cases for both algorithms (Figure 5). Kinase inhibitors are known for their promiscuous polypharmacological interactions (at least within kinase targets), which are supported by sequence and structural conservation specifically at the ATP binding site core, the site that the majority of inhibitors target and where they compete with ATP.[43,44] Also, studies on cross reactivity of clinical kinase inhibitors have demonstrated the presence of promiscuity of those inhibitors across multiple protein families,[45] rendering this result understandable. While it is quite possible to anticipate that a compound inhibits a kinase, it is much more difficult to predict which kinase (out of a large set) is the most likely one to be inhibited. Given the great similarity between kinase inhibitors in chemical space, it appears that it will be necessary to extend this work by investigating more thoroughly the relationships in the pharmacological space of kinases, which may lead to the exploration of methods to improve performance, such as collapsing bioactivity classes, where appropriate. This could substantially improve the performance of the models, while still being able to differentiate those subsets of kinases, which can be realistically distinguished from another.

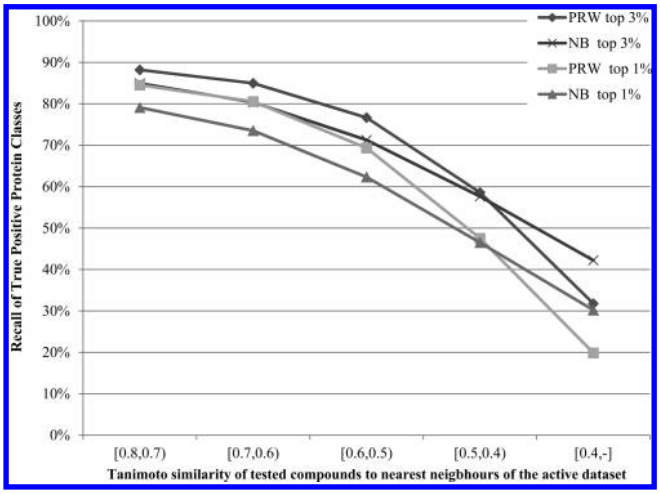


Figure 7. Recall among the top 1% and 3% positions achieved by NB and PRW versus the similarity to the nearest neighbor of the training set measured by Tanimoto coefficient in ECFP4 space. It can be seen that for compounds with a Tc larger than 0.4, PRW performed better than NB and achieved a higher recall, while for compounds with a Tc smaller than 0.4 (i.e., for cases where no similar compounds are present in the training set), NB achieved higher recall (42% vs 30.20%, for the top 3% and top 1%, respectively) than PRW (31.76% and 19.87%). These results suggest that NB gives better extrapolation to new chemical space than PRW, likely by being able to link molecular features in novel ways.

In order to illustrate the difference in behavior between PRW and NB in a particular case, target predictions generated for ethylketazocine (EKT), an opioid receptor agonist (CHEMBL id= CHEMBL278086), and clozapine, an atypical antipsychotic agent (compound id = CHEMBL42), are presented in

dx.doi.org/10.1021/ci300435j | J. Chem. Inf. Model. 2013, 53, 1957−1966

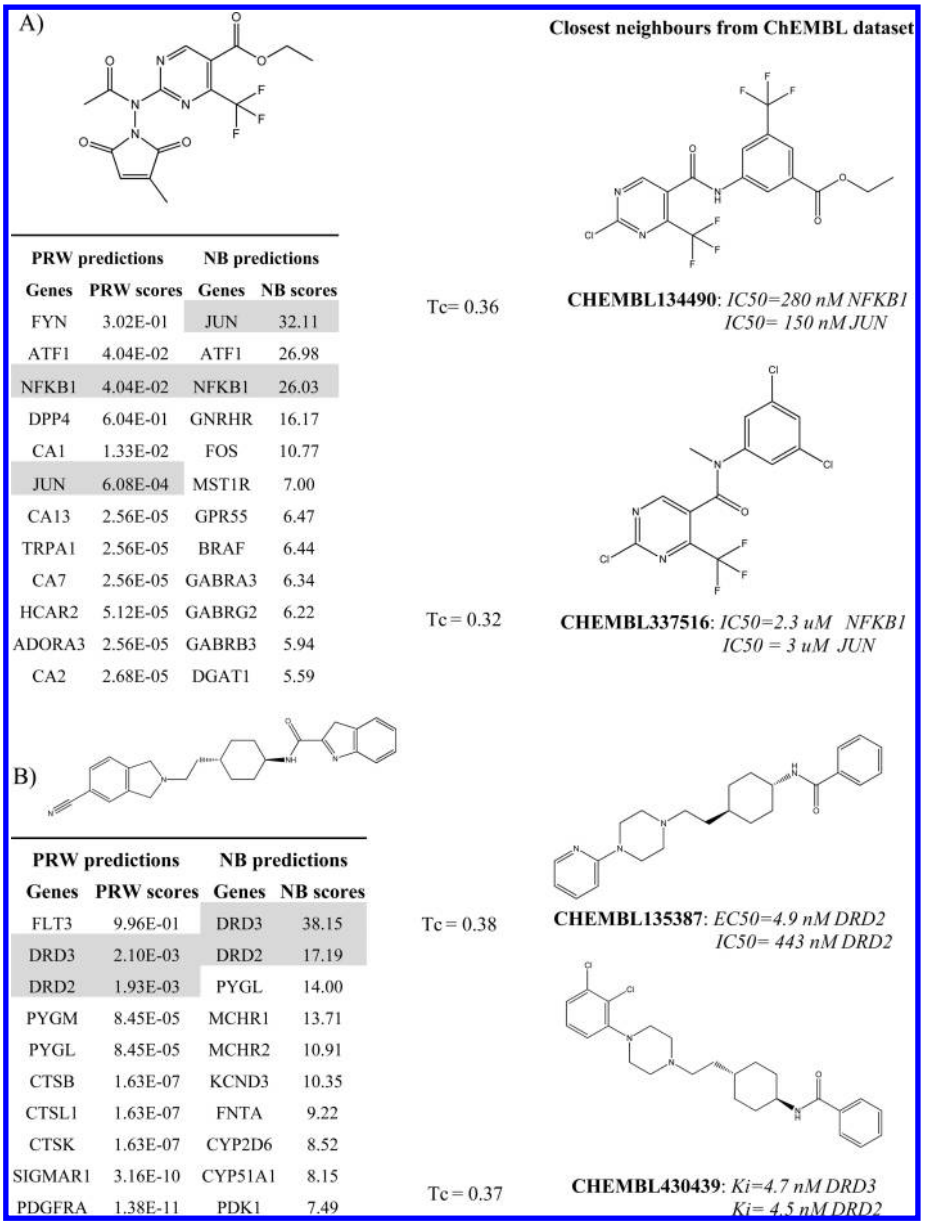


Figure 8. Examples among tested external compounds from WOMBAT where the models achieved good extrapolation to new chemical space (i.e., where no similar compounds were given in the training set). A) An antagonist of human transcription factor Ap-1 (*JUN_HUMAN*) and nuclear factor NF-kappa-B (*NFKB1_HUMAN*). Closest neighbors in the ChEMBL training data set were found to be in the similarity range of Tc = 0.36 (CHEMBL134490) and Tc = 0.32 (CHEMBL337516). B) An antagonist of human dopamine receptor 2 and 3 (*DRD2_HUMAN and DRD3_HUMAN*). Closest training set neighbors were found in the range of Tc = 0.38 (CHEMBL135387) and Tc = 0.37 (CHEMBL430439).

Figures 6 A and B, respectively. It can be seen that for the case of EKT both algorithms retrieved the main targets of the opioid family (Mu, Kappa, and Delta) in the top three positions. However, in the case of clozapine PRW failed to retrieve the correct targets among the top ranked positions. This is due to the different ways in which both algorithms draw conclusions from the known bioactive compounds to make predictions for a novel structure. Such observations suggest that it would be beneficial to use multiple algorithms and not depend solely on one algorithm, which might fail in a particular case and that is in many cases difficult to anticipate.

External Validation Using WOMBAT data. In order to test the applicability of the target prediction algorithms on novel chemical substances, the performance of the predictions was tested on a data set retrieved from a different source, WOMBAT,[46] with very similar compounds excluded (more precisely, compounds in this external test set were less than 0.8 similar in ECFP_4/Tanimoto space from the ChEMBL data set that was used for testing and training the models). In the case of this external test set, the results obtained for PRW achieved a recall in the top 9 positions of 66.61% and in the top 27 positions of 73.9%. In the case of NB the results obtained were 63.3% and 72.1%, respectively. While lower than in the case of the training/test set splits performed above, these results are very likely to be closer to the real world applicability of such predictive models, where novel molecules are most interesting if they come from as yet unexplored areas of chemical space.

Furthermore, in order to quantify the relationship between retrieval rate among the top-k positions and the distance of compounds to the nearest neighbor in the training set

(i.e., to establish the applicability domain of the model), the recall achieved in the top 9 and 27 positions was plotted against the distance in chemical space, which is shown in Figure 7. It can be seen that for compounds with a Tc larger than 0.4, PRW achieved better recall among the top 9 and 27 positions than NB. On the other hand, NB achieved better performance for compounds with a Tc less than 0.4, achieving more than 10% better performance among the top 9 and 27 positions compared to PRW. These results suggest that, when testing new chemical compounds with no close neighbors, the chance of retrieving the correct targets is better with NB than with PRW. This is in line with the expectation that the NB method is able to "pool" features from multiple compounds in a given bioactivity class to make predictions for as yet unseen compounds, while this is not possible for the PRW method (which considers the distance to a set of individual features instead). This explanation is in line with previous results obtained from fragment-based screening, where models based on the Bayes classifier were also able to combine information from multiple active fragments in order to identify novel bioactive molecules.[47]

Surprisingly, even for many compounds with low structural similarity to known examples in the training set, both algorithms were able to retrieve the correct targets. Two such examples are presented in Figure 8, where although no closely related structure was present (all training set compounds had a similarity of less than 0.4 in Tanimoto space), the algorithms were still able to correctly annotate the compounds with targets. In the case of the antagonist of the human transcription factors Ap-1 and NFκB, this was possible despite the different scaffolds of the structures; the same is true for the dopamine receptor ligand, where structures seem at first sight to be rather similar, but where extrapolation at the terminal ring systems and as well as at the piperazine moiety are still necessary to achieve successful target extrapolation. These examples demonstrate that, while in silico target prediction models are based on and limited by the data availability, given a suitable choice of descriptors, extrapolation can still be performed successfully to related structures.

CONCLUSIONS

In this work, large scale in silico target prediction models were developed using bioactivity data extracted from the ChEMBL database, and the performance of the well-established Naïve Bayes classifier was compared with the newly proposed Parzen-Rosenblatt Window method. The PRW outperformed the NB on this data set. The data set used in this study has been made available to serve as a benchmarking data set for future studies.

Furthermore, the influence of class size and diversity on the performance of such predictive models was explored. It was found that both class size and chemical diversity play critical roles in the ability of the models to predict the correct protein targets for a chemical compound. Classes with a large number of data points, those with at least 200, tend to be predicted better than smaller ones due to more information being included in those classes. Similarly, classes with high intraclass similarity (low chemical diversity) were found to be predicted better than those with lower similarity. Although no strong conclusions can be made on whether the number of data points available per class or structural diversity among compounds describing each class is decisive, it should be taken into consideration when developing such models.

Testing the performance of the current models on data extracted from the WOMBAT database served as a realistic test of how such modeling approaches would perform in realistic scenarios, where the ability of a method to extrapolate to new chemical space is important. Examples of successful cases were found even for compounds of very different chemistry from those in the training set. This demonstrated that in silico prediction methods could provide valuable information to medical chemists and assist in prioritizing lead compounds.

ASSOCIATED CONTENT

Supporting Information

(1) Benchmarking data set. List of ligand-protein pairs extracted from ChEMBL with compound identifiers (using ChEMBL molregno identifier), compound structures are provided in canonical SMILES format and protein target identifiers (using chembl_id identifier) (ChEMBL_human_SARs). (2) Protein-Target identifiers. Protein target ChEMBL identifiers, Uniprot identifiers, protein description and genes names. MySQL query. (3) MySQL query used to extract human Structure Activity Relationships from ChEMBL (SQL_query_human_SARs). (4) Optimization of smoothing parameter (h) of PRW on the data set (Supporting Information, Figure S1). (5) Table with mean ranking, median ranking, intraclass Tanimoto similarity, and class size per target obtained from 5-fold cross-validation (Model statistics). (6) External test set. External test set with WOMBAT compound identifier and annotated protein targets with ChEMBL identifiers. No structures of small molecules or activity measurements from WOMBAT are included. Tanimoto similarity (in ECFP_4) for each molecule in test set to its closest neighbor from the ChEMBL training data set. (7) Implementation of PRW in Python (version 2.7) and NB in C#. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: ab454@cam.ac.uk (A.B.), rcg28@cam.ac.uk (R.C.G.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

A.K., A.B., and R.C.G. acknowledge support by Unilever. J.B.O.M. thanks the Scottish Universities Life Sciences Alliance (SULSA) for funding.

REFERENCES

(1) Swinney, D. C.; Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discovery* **2011**, *10*, 507−519.

(2) Feng, Y.; Mitchison, T. J.; Bender, A.; Young, D. W.; Tallarico, J. A. Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat. Rev. Drug Discovery* **2009**, *8*, 567−578.

(3) Terstappen, G. C.; Schlupen, C.; Raggiaschi, R.; Gaviraghi, G. Target deconvolution strategies in drug discovery. *Nat. Rev. Drug Discovery* **2007**, *6*, 891−903.

(4) Arrowsmith, J. Trial watch: Phase II failures: 2008−2010. *Nat. Rev. Drug Discovery* **2011**, *10*, 328−329.

(5) Kubinyi, H. Drug research: myths, hype and reality. *Nat. Rev. Drug Discovery* **2003**, *2*, 665−668.

(6) Scheiber, J.; Chen, B.; Milik, M.; Sukuru, S. C. K.; Bender, A.; Mikhailov, D.; Whitebread, S.; Hamon, J.; Azzaoui, K.; Urban, L.; Glick, M.; Davies, J. W.; Jenkins, J. L. Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J. Chem. Inf. Model.* **2009**, *49*, 308−317.

(7) Ekins, S. Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discovery Today* **2004**, *9*, 276−285.
(8) Johnson, M. A.; Maggiora, G. M. In *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(9) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.
(10) 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*, D400−4012.
(11) 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.* **2011**, *40*, D1100−D1107.
(12) Jensen, N. H.; Roth, B. L. Massively parallel screening of the receptorome. *Comb. Chem. High Throughput Screening* **2008**, *11*, 420−426.
(13) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805−815.
(14) Fabian, M. A.; Biggs, W. H., 3rd; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lelias, J. M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329−336.
(15) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38−52.
(16) Chen, Y. Z.; Zhi, D. G. Ligand-protein inverse docking and its potential use in the computer search of protein targets of small molecules. *Proteins* **2001**, *42*, 217−226.
(17) Schuster, D. 3D pharmacophores as tools for activity profiling. *Drug Discovery Today: Technol.* **2010**, *7*, e205−e211.
(18) Jenkins, J. L.; Bender, A.; Davies, J. W. In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today: Technol.* **2006**, *3*, 413−421.
(19) Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P. J.; Whitmore, A. V.; Zimmer, S.; Young, M. P.; Jenkins, J. L.; Glick, M.; Glen, R. C.; Bender, A. From in silico target prediction to multi-target drug design: current databases, methods and applications. *J. Proteomics* **2011**, *74*, 2554−74.
(20) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124−33.
(21) Nigsch, F.; Bender, A.; Jenkins, J. L.; Mitchell, J. B.O. Ligand-target prediction using Winnow and naive Bayesian algorithms and the implications of overall performance statistics. *J. Chem. Inf. Model.* **2008**, *48*, 2313−25.
(22) Wale, N.; Karypis, G. Target fishing for chemical compounds using target-ligand activity data and ranking based methods. *J. Chem. Inf. Model.* **2009**, *49*, 2190−2201.
(23) 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.
(24) Keiser, M. J.; Irwin, J. J.; Shoichet, B. K. The chemical basis of pharmacology. *Biochemistry* **2010**, *49*, 10267−10276.
(25) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(26) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261−266.
(27) Bender, A.; Mikhailov, D.; Glick, M. S.; Davies, J.; Cleaver, W. J.; Marshall, S.; Tallarico, S.; Harrington, J.; Cornella-Taracido, E.; Jenkins, J. L. Use of ligand based models for protein domains to predict novel molecular targets and applications. *J. Proteome Res.* **2009**, *8*, 2575−2585.
(28) Prathipati, P.; Ma, N. L.; Manjunatha, U. H.; Bender, A. Fishing the target of antitubercular compounds: in silico target deconvolution model development and validation. *J. Proteome. Res.* **2009**, *8*, 2788−2798.
(29) DeGraw, A. J.; Keiser, M. J.; Ochocki, J. D.; Shoichet, B. K.; Distefano, M. D. Prediction and evaluation of protein farnesyltransferase inhibition by commercial drugs. *J. Med. Chem.* **2010**, *53*, 2464−2471.
(30) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361−367.
(31) Bender, A. Databases: Compound bioactivities go public. *Nat. Chem. Biol.* **2010**, *6*, 309−309.
(32) Lowe, R.; Mussa, H. Y.; Nigsch, F.; Glen, R. C.; Mitchell, J. B.O. Predicting the mechanism of phospholipidosis. *J. Cheminf.* [Online] **2012**, *4*, Article 2. http://www.jcheminf.com/content/4/1/2 (accessed Oct 19, 2012).
(33) Tiikkainen, P.; Franke, L. Analysis of commercial and public bioactivity databases. *J. Chem. Inf. Model.* **2012**, *52*, 319−326.
(34) Southan, C.; Boppana, K.; Jagarlapudi, S. A. R. P.; Muresan, S. Analysis of in vitro bioactivity data extracted from drug discovery literature and patents: Ranking 1654 human protein targets by assayed compounds and molecular scaffolds. *J. Cheminf.* [Online] **2011**, *3*, Article 14. http://www.jcheminf.com/content/3/1/14 (accessed Oct 19, 2012).
(35) *Pipeline Pilot*, version 6.1.5.0 Student ed.; Accelrys: San Diego, 2007.
(36) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9*, 199−204.
(37) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(38) Bender, A. How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opin. Drug Discovery* **2010**, *5*, 1141−1151.
(39) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.
(40) Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Statist.* **1962**, *33*, 1065−1076.
(41) Hand, D. J. *Discrimination and Classification,* 1st ed.; John Wiley & Sons: Chichester — Brisbane — New York — Toronto, 1981; pp 24−26.
(42) Bishop, M. C. *Pattern Recognition and Machine Learning*, 1st ed.; Springer: Berlin, 2006; p 123.
(43) Verkhivker, G. M. Imprint of evolutionary conservation and protein structure variation on the binding function of protein tyrosine kinases. *Bioinformatics* **2006**, *22*, 1846−1854.
(44) Scheeff, E. D.; Bourne, P. E. Structural evolution of the protein kinase-like superfamily. *PLoS Comput. Biol.* **2005**, *1*, e49.
(45) Fabian, M. A.; Biggs, W. H., III; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, M. J.; Galvin, M.; Gelach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Lai, A. G.; Lelias, J. M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lackhart, D. J. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329−336.
(46) Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mracec, M.; Oprea, T. I. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*; Schreiber, S. L., Kapoor, T. M., Wess, G., Eds.; Wiley-VCH: New York, 2007; pp 760−786.
(47) Crisman, T. J.; Bender, A.; Milik, M.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Fejzo, J.; Hommel, U.; Davies, J. W.; Glick, M. "Virtual fragment linking": An approach to identify potent binders from low affinity fragment hits. *J. Med. Chem.* **2008**, *51*, 2481−2491.