Journal of Molecular Graphics and Modelling 29 (2010) 372–381

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Calculation and application of activity discriminants in lead optimization

Xincai Luo[a,*], Jennifer R. Krumrine[a], Ashok B. Shenvi[a], M. Edward Pierson[b], Peter R. Bernstein[a]

[a] Department of Chemistry, AstraZeneca Pharmaceuticals, 1800 Concord Pike, Wilmington, DE 19850, USA
[b] Neuroscience Therapeutic Area, AstraZeneca Pharmaceuticals, 1800 Concord Pike, Wilmington, DE 19850, USA

ARTICLE INFO

Article history:
Received 28 May 2010
Received in revised form 10 July 2010
Accepted 14 July 2010
Available online 30 July 2010

Keywords:
In vitro activity
Linear discriminant analysis
Activity discriminant
Medicinal chemistry descriptors
Lead optimization
QSAR

ABSTRACT

We present a technique for computing activity discriminants of *in vitro* (pharmacological, DMPK, and safety) assays and the application to the prediction of *in vitro* activities of proposed synthetic targets during the lead optimization phase of drug discovery projects. This technique emulates how medicinal chemists perform SAR analysis and activity prediction. The activity discriminants that are functions of 6 commonly used medicinal chemistry descriptors can be interpreted easily by medicinal chemists. Further, visualization with Spotfire allows medicinal chemists to analyze how the query molecule is related to compounds tested previously, and to evaluate easily the relevance of the activity discriminants to the activities of the query molecule. Validation with all compounds synthesized and tested in AstraZeneca Wilmington since 2006 demonstrates that this approach is useful for prioritizing new synthetic targets for synthesis.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

In vivo assays in drug discovery are slow and expensive. To reduce cost, pharmaceutical companies have developed many *in vitro* assays that serve as valuable early assessments of compounds [1–15]. In the discovery process, new compounds are tested in a battery of *in vitro* assays first. Only those that pass the cascade criteria of the *in vitro* assays progress into the *in vivo* tests. *In vitro* assays used in drug discovery projects today include pharmacological, DMPK, and safety assays. They are often run in parallel to optimize various parameters including potency, selectivity, and physical and ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties.

While it is quite easy and often inexpensive to profile compounds with *in vitro* assays, it is still quite expensive to synthesize new compounds. *In silico* models that can accurately predict the *in vitro* activities of proposed synthetic targets help medicinal chemists reduce the number of compounds that need to be synthesized and, therefore, improves the speed and cost of drug discovery.

There are many approaches for predicting *in vitro* activities. The various approaches differ in three primary aspects: molecules used to train the models, descriptors used to characterize the molecular structures, and mathematical learning techniques used to establish relationships between the descriptors and the *in vitro* activities. QSAR models used for activity prediction can be classified as global or local models based on the set of molecules used for training them. Global models are developed with whole data set. Therefore, they model the relationship between the descriptors and the activities of all the molecules. Such models often work well when the data set is not very large. For large data sets, the structural features that are relevant to the activities are not evenly distributed. Some features exist in a large proportion of the molecules while others exist in a smaller set. Global models are biased toward the global features that are in large number of molecules. Local models are trained using subsets of molecules in the data set. The subsets can be obtained by various approaches such as clustering, similarity searches, substructure searches, or by restricting to the compounds in a specific drug discovery project or series. One advantage of local modeling is that a small training set allows the models be updated very frequently, which generally gives rise to improved predictive performance by making the training set more relevant to the chemistry the team is pursuing currently. Another advantage is that local models built upon the subsets of similar compounds have a better chance of capturing relevant structure–activity relationships than global models derived from whole diverse data sets. It has been demonstrated in many cases that local models based on subsets

Abbreviations: LDA, linear discriminant analysis; LD, linear discriminant; MW, molecular weight; *c* Log *P*, calculated water and alcohol partition coefficient; TPSA, topological polar surface area; NHD, number of hydrogen bond donors; NHA, number of hydrogen bond acceptors; NRB, number of rotatable bonds; hCLint, human microsome metabolism intrinsic clearance; hERG, human Ether-a-go-go Related Gene; DMPK, drug metabolism and pharmacokinetics; QSAR, quantitative structure–activity relationship; 2D, two-dimensional; 3D, three-dimensional; PE, percent effect.

* Corresponding author. Tel.: +1 302 886 1030; fax: +1 302 886 4989.
E-mail address: Xincai.Luo@astrazeneca.com (X. Luo).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.07.005

Table 1
In vitro activities modeled in this study. The three pharmacological assays are used in one discovery project; the 6 DMPK assays are used in all the discovery projects; the two safety assays are used in most drug discovery projects in Wilmington. The cutoff values are the screening cascade criteria that compounds need to meet to progress through the project's assay cascade. The number of data points in each assay is number of compounds that have been successfully measured experimentally. The number of compounds in the test set is the number of compounds synthesized in 2006–2008, used for validating the models.

Assay	Cutoff	Number of compounds in the data set	Number of compounds in test set	Number of compounds that are Green in test set	Number of compounds that are Red in test set
Binding K_i	10 nM	2175	727	313	414
GTPγS pIC_{50}	40 nM	933	555	161	394
GTPγS PE	50%	988	556	452	104
Permeability	10^{-5} cm/s	4530	1666	711	955
Efflux Ratio	3	4460	1648	940	708
hCLint	25 μL/min/mg	10,589	3573	1331	2242
Solubility	10 μM	15,509	5372	3649	1723
CYP2D6 pIC_{50}	10 μM	1914	704	449	255
CYP3A4 pIC_{50}	10 μM	2330	771	618	153
hERG pIC_{50}	10 μM	35,713	3685	2508	1177
PLD pEC_{50}	50 μM	888	423	230	193

of similar compounds perform better than global ones within the relevant test set [16–21].

Thousands of descriptors are available for QSAR modeling [22]. Many can be calculated easily with standard packages of commercial software such as Dragon (http://www.talete.mi.it), MolConnZ (http://www.edusoft-lc.com/molconn/), Accelrys Pipeline Pilot (http://accelrys.com/products/scitegic/), MOE (http://www.chemcomp.com/software-chem.htm), Sybyl (http://www.tripos.com), and others. Commonly used descriptors include atom or group counts [23–26], physical chemical properties [27], partial charge [28–30], topological polar surface area [31–33], molecular fingerprints [34–40], molecular connectivity indices [41–44] and E-state indices [44–47], CoMFA [48,49], GRIND [50–52], and Volsurf [53–54].

Although many descriptors can be used for QSAR modeling, very few of them are used by medicinal chemists in their SAR analyses and molecular design. Commonly used descriptors by medicinal chemists are molecular weight (MW), calculated water and alcohol partition coefficient (*c* Log *P*), topological polar surface area (TPSA), number of hydrogen bond donors (NHD), number of hydrogen bond acceptors (NHA), and number of rotatable bonds (NRB). These descriptors describe the size (MW), lipophilicity (*c* Log *P* and TPSA), hydrogen bonding potential (NHD and NHA), and flexibility (NRB) of molecules.

Various machine learning techniques have been applied to classify *in vitro* activities of drug molecules [55,56], including linear discriminant analysis (LDA) [57], decision trees (DT) [58–61], support vector machines (SVM) [62,63] and artificial neural networks (ANN) [64]. Models generated from SVM and ANN are difficult to interpret. They often serve as a black box for classifying activities such that little insight can be gained as to why a molecule belongs to a particular activity class. In contrast, LDA and DT methods generate readily interpretable models that can give insight into the molecular descriptors that distinguish one activity class from another.

Most QSAR models published so far for predicting *in vitro* activities are focused on a single *in vitro* activity. This limits their utility in prioritizing new synthetic targets in lead optimization, which requires prediction of all the *in vitro* activities in the assay cascade. In addition, they employ mathematical learning techniques that are not interpretable or employ complex descriptors, making it difficult for medicinal chemists to utilize the results for guiding new designs.

In this paper, we report a technique that can be used to calculate activity discriminants of *in vitro* assays in drug discovery based on local modeling and 6 commonly used medicinal chemistry descriptors. These discriminants can be used by medicinal chemists to determine if a new molecule will pass the assay cascade in lead optimization. We have successfully applied the approach to predict *in vitro* activities in lead optimization in AstraZeneca Wilmington in the past three years to help medicinal chemists prioritize new molecules for synthesis. For each *in vitro* assay, to calculate the activity discriminant of a new molecule, we first perform a similarity search [17,21] against all the compounds that have been tested in the assay and then select a small number of the most similar compounds from the search to perform linear discriminant analysis to find a discriminant that best describes the activity. In this approach, the relevant structure–activity relationships are very well captured and activity discriminants computed are always up-to-date, both of which are key to good predictive performance.

2. Methods

2.1. Data set

We calculated activity discriminants of *in vitro* assays and applied them to many lead optimization projects in AstraZeneca Wilmington from 2006 to the present. In this paper, we report the calculation and application for eleven *in vitro* assays listed in Table 1. Among them, three are pharmacological assays from one lead optimization project, six are physical property and DMPK assays used in all discovery projects, and two are safety assay used in most drug discovery projects.

2.1.1. 5-HT_{1B} receptor SPA radioligand binding assay [65]

The pharmacological assay was run in AstraZeneca Wilmington to support the 5-HT_{1B} antagonist lead optimization project. It quantifies affinity of compounds to the 5-HT_{1B} receptor. Membranes from CHO cells expressing human 5-HT_{1B} receptor were used as the receptor source. The beads, containing a scintillant, are coated with wheat germ agglutinin, which binds to glycosylated sites on the membranes. Binding of the radioactive ligand to its receptor on the membrane stimulates the scintillant within the bead to emit light. Only bound ligand, which is in close proximity to the scintillant, is detected. A K_i value was reported for each compound tested. A compound was considered to be Green (or pass the assay) if the $K_i \leq 10$ nM or Red if $K_i > 10$ nM.

2.1.2. 5-HT_{1B} receptor GTPγS SPA assay [65]

The pharmacological assay was used in AstraZeneca Wilmington to support 5-HT_{1B} antagonist lead optimization project. 5-HT_{1B} receptor belongs to the class of G-protein coupled receptors. The G-proteins are heterotrimeric with α, β, and γ subunits. The assay is based on GTPγS binding to the α subunit activated by human 5-HT_{1B} receptor stimulation. In the assay, the radioactive [^{35}S]GTPγS was added to 5-HT_{1B} membranes in the presence of agonist and the GDP/GTP exchange was monitored by trapping the

[^{35}S]GTPγS/G-protein/GPCR complex onto SPA beads. The assay used CHOK1 membranes expressing human 5-HT_{1B} receptors (15 μg protein/well) and 200 pM GTPγ^{35}S with WGA PVT beads (50 μg/well) to test compounds in an 11pt IC_{50} curve. It classifies compounds as antagonists, agonists, or inverse agonists based on the functional response of the binding. Both IC_{50} (GTPγS IC_{50}) and top %effect (GTPγS PE) were reported for each compound tested. For IC_{50}, a compound was considered to be Green if the $IC_{50} \leq 40$ nM or Red if the $IC_{50} > 40$ nM. For top %effect, a compound was considered to be Green if the antagonist %effect ≥50%.

2.1.3. P-glycoprotein transport assay

P-glycoprotein (Pgp), an ATP-dependent efflux transporter, is highly expressed in blood–brain barrier and intestine. Pgp is regarded as one of the major rate limiting steps for CNS penetration and limits the intestinal absorption of a number of clinically important drugs. The assay was done with Madin-Darby Canine kidney cells (MDCK) expressing human MDR1 Pgp. MDCK cells were grown on Transwell filters and displayed orientation within the established monolayer. Assays were performed three days post seeding. The apical-to-basolateral (A-to-B) apparent Permeability and Efflux Ratio, the ratio of basolateral to apical (B-to-A) versus apical-to-basolateral, were reported for each compound tested. The cutoff for apparent Permeability is 10^{-5} cm/s. A compound is considered Green (or pass) if the Permeability is $\geq 10^{-5}$ cm/s. Otherwise it is Red (or failed). The cutoff for Efflux Ratio is 3 that a compound is considered to be Green (or pass) if the ratio is ≤3 and Red if it is >3.

2.1.4. Human microsome metabolism intrinsic clearance (hCLint) Assay

The hCLint assay was performed using human liver microsomes as the enzyme source. Test compounds of 1 μM were incubated with human liver microsomes and NADPH. At various times, incubations were sampled and analyzed by LC/MS/MS to determine the loss of parent compound. The intrinsic clearance (hCLint) was reported for each compound tested. The cutoff for the activity was 25 μL/min/mg. A compound was considered Green (or pass the assay) if hCLint ≤25 μL/min/mg or Red if hCLint > 25 μL/min/mg.

2.1.5. Solubility assay

Aqueous solubility is an important parameter that needs to be optimized for oral drug molecules; further, poor solubility can limit testing in *in vitro* assays. The solubility data used in this study were all measured at pH = 7.4 using a dried-DMSO method [7]. Our internal validation study showed that the solubility data measured with dried-DMSO correlated well with those measured with solid method. The cutoff for the property was 10 μmol (for most discovery projects). A compound was considered to be Green (or pass) if the solubility is ≥10 μM or Red if it is <10 μM.

2.1.6. CYP3A4 and CYP2D6 inhibition assays

Cytochrome P450 (CYP) enzymes constitute the most important group of drug metabolizing enzymes in the body. The potential of a compound to modulate the activities is an important indication of potential for drug–drug interactions. The assessment of the potential was done by measuring the activities of these specific enzymes using selective probes *in vitro*, in the absence and presence of multiple concentrations of the compounds being tested. IC_{50} values were reported for the compounds tested. Inhibition data of two CYP enzymes, CYP3A4 and CYP2D6, were investigated in this study. The cutoff is 10 μM for both assays. A compound is considered to be Green (or pass the assay) if the $IC_{50} \geq 10$ μM or Red if the $IC_{50} < 10$ μM.

2.1.7. hERG assay

The human Ether-a-go-go related gene safety assay is a medium-throughput electrophysiology-based hERG assay using IonWorks HT [15]. An IC_{50} value was reported for each compound tested. The cutoff for IC_{50} was 10 μmol. A compound was considered to be Green (or pass the assay) if the $IC_{50} \geq 10$ μM or Red if the $IC_{50} < 10$ μM.

2.1.8. Phospholipidosis assay

Phospholipidosis is characterized by the accumulation of polar phospholipids in association with the development of unicentric or multicentric lamellated bodies. The phospholipidosis assay is a fluorescence-based *in vitro* screen using the Cellomics ArrayScan high-content screening platform, which captures and analyzes images from 96-well cell culture microtiter plates using multichannel fluorescence microscopy [14]. An EC_{50} was reported for each compound tested. The cutoff for EC_{50} was 50 μM. A compound was considered to be Green if the $EC_{50} \geq 50$ μmol or Red if the $EC_{50} < 50$ μM.

2.2. Test sets

For each *in vi*tro assay, the test set contains all the compounds synthesized and tested in AstraZeneca Wilmington from 2006 to 2008. The numbers of test set compounds for the *in vitro* assays are listed in Table 1. Since we are interested in using the calculated activity discriminants for prioritizing new molecules for synthesis, we selected only compounds that we synthesized for discovery projects. This excluded the compounds from other sources such as purchasing externally.

2.3. Training sets

For each *in vitro* assay, we performed a specific discriminant analysis for each query molecule in the test set to calculate the activity discriminant. We first performed a similarity search against all the compounds that were tested in the assay to select the N most similar compounds [21]. These were then used to perform discriminant analysis for calculating activity discriminant of the query molecule. For analyzing performance for explaining the existing activity data, all the compounds that have activity data were used in the similarity search. But for analyzing performance for predicting *in vitro* activities of new molecules, only compounds whose assay dates are older than that of the query molecule were used in the search. For a given assay, the number of most similar compounds used in the discriminant analysis was fixed for all the query molecules in the test set. However, the training set size varied from assay to assay and was determined based on the accuracy of the predictions.

2.4. Similarity

The similarity of two molecules was measured with the Tanimoto similarity coefficient of AlFi fingerprints [66] developed in AstraZeneca. The principle behind AlFi fingerprints is analogous to that for Daylight fingerprints (http://www.daylight.com/dayhtml/doc/theory/). It is based on the linear fragments from 1 to 8 atoms in length and hashed to 1024 bits. The fingerprints have been used in AstraZeneca in similarity search and compound clustering for the past five years.

2.5. Molecular descriptors

The discriminant analyses were all performed with 6 commonly used medicinal chemistry descriptors: MW, *c* Log *P* (http://www.biobyte.com/bb/prod/bioloom.html), TPSA

[29], NHD [67], NHA [67], and NRB (http://www.daylight.com/dayhtml/doc/theory/theory.mol.html). These 6 descriptors are popular among medicinal chemists for two main reasons. One is that they have clear physical chemical meaning. MW describes the size of the molecules; $c\log P$ and TPSA are related to lipophilicity of the molecules; NHD and NHA describe the hydrogen bonding potential of the molecules; NRB describes the flexibility of the molecules. It is quite easy for medicinal chemists to link each of these descriptors to the actual chemical structures and incorporate information from the activity discriminant back into designing the next molecule. Another reason is that since Lipinski published the Rule of Five a decade ago, it has been quite clear to medicinal chemists how each of the descriptors affect druggability in general [67].

2.6. Discriminant analysis

Discriminant analyses were all done with linear discriminant analysis (LDA) implemented in the R statistical package (http://www.r-project.org). LDA is a statistical technique for finding a discriminant function that can best partition two or more groups of compounds in the training set. In our discriminant analyses of *in vitro* activities, each compound in the training set was classified into one of two classes, Green or Red, according to the screening cascade criteria. In order to perform LDA analyses, there must be some existing (measured) compounds in each class in the training set. For a query molecule in a test set, if all the compounds in the training set were in one class, LDA analysis would not be performed. If all the compounds in the training set were Green, the discriminant (LD) was set to −1.1, a value that the query molecule is predicted to be Green with high confident, and if all the compounds were Red, the discriminant (LD) was set to 1.1, a value that the query molecule is predicted to be Red with high confidence.

When there were compounds in each activity class in the training set, we performed an LDA analysis with R. When LDA was run successfully, its output was the predicted class and discriminant (LD) value of each compound in the training set. However, discriminant values from R cannot be visualized easily with Spotfire. To make the visualization easy, we transformed the discriminants outputted from R into new discriminants. Let *ldr* be the linear discriminant value of a compound outputted from R, TG be the set of compounds in the training set that were classified as Green by the LDA in R, and TR be the set of compounds in the training set that were classified as Red by the LDA in R. We define the following four minimum and maximum values of the *ldr*:

$$g_{\min} = \min\{ldr(c) : c \in \mathrm{TG}\} \tag{1}$$

$$g_{\max} = \max\{ldr(c) : c \in \mathrm{TG}\} \tag{2}$$

$$r_{\min} = \min\{ldr(c) : c \in \mathrm{TR}\} \tag{3}$$

$$r_{\max} = \max\{ldr(c) : c \in \mathrm{TR}\} \tag{4}$$

We can transform the *ldr*, the linear discriminant outputted from R, into a new set of linear discriminants (LD) as follows:

If both TG and TR are not empty:

$$\mathrm{LD} = \begin{cases} ldr - \dfrac{g_{\max} + r_{\min}}{2} & \text{if } g_{\max} < r_{\min} \\ \dfrac{r_{\max} + g_{\min}}{2} - ldr & \text{if } r_{\max} < g_{\min} \end{cases} \tag{5}$$

If TG is empty:

$$\mathrm{LD} = ldr - r_{\min} + 0.1 \tag{6}$$

If TR is empty:

$$\mathrm{LD} = ldr - g_{\max} - 0.1 \tag{7}$$

With this transformation, LD > 0 for all the compounds in TR (the compounds in the training set that were classified to be Red by the LDA model) and LD < 0 for all the compounds in TG (the compounds in the training set that were classified to be Green by the LDA model). For each in vitro assay, the LDs describe activities of compounds. A new molecule in the assay is predicted to be Green if its LD < 0 and Red if its LD > 0. We will refer LDs as activity discriminants of the assay.

LD is linear function of the six medicinal chemistry descriptors since it differs from the *ldr* by a constant and a sign. This allows medicinal chemists to interpret the results easily to gain insight as what structural features lead to the favorable or poor activity.

2.7. Performance measurement

The activity discriminants (LDs) were used for prioritizing new molecules for synthesis. The goal is to be able to select from among the synthetic candidates those that are most likely to have good *in vitro* activities. Let G_{pred} and R_{pred} be sets of compounds predicted to be Green and Red, respectively, and let G_{exp} and R_{exp} be the sets of compounds found to be Green and Red, respectively, from testing them experimentally in the assay. The percent of compounds ($P_{\text{G-to-G}}$) in G_{pred} that are found to be Green experimentally in the assay can be calculated by:

$$P_{\text{G-to-G}} = \frac{N_{G_{\mathrm{pred}} \cap G_{\mathrm{exp}}}}{N_{G_{\mathrm{pred}}}} \tag{8}$$

In above equation, $N_{G_{\mathrm{pred}}}$ is the total number of compounds that were predicted to be Green and $N_{G_{\mathrm{pred}} \cap G_{\mathrm{exp}}}$ is the number of those compounds, predicted to be Green, that are actually verified to be Green experimentally with the assay. The percent of compounds ($P_{\text{G-to-R}}$) in G_{pred} that are found to be Red experimentally in the assay is 100% – $P_{\text{G-to-G}}$ since there are only two classes (Green and Red) of compounds experimentally as defined by the assay cascade.

Similarly, the percent of compounds ($P_{\text{R-to-G}}$) in R_{pred} that are found to be Green experimentally in the assay can be calculated by:

$$P_{\text{R-to-G}} = \frac{N_{R_{\mathrm{pred}} \cap G_{\mathrm{exp}}}}{N_{R_{\mathrm{pred}}}} \tag{9}$$

and the percent of compounds ($P_{\text{R-to-R}}$) in R_{pred} that are found to be Red experimentally in the assay is 100% – $P_{\text{R-to-G}}$.

The ratio (f_{sep}) of $P_{\text{G-to-G}}$ over $P_{\text{R-to-G}}$, calculated with Eq. (10), measures how good the class of molecules, predicted to be Green, than the class of molecules, predicted to be Red, is for meeting the assay cascade criterion.

$$f_{\mathrm{sep}} = \frac{P_{\text{G-to-G}}}{P_{\text{R-to-G}}} \tag{10}$$

The overall accuracy of prediction is calculated by:

$$\mathrm{Accuracy} = \frac{N_{G_{\mathrm{pred}} \cap G_{\mathrm{exp}}} + N_{R_{\mathrm{pred}} \cap R_{\mathrm{exp}}}}{N_{G_{\mathrm{pred}} \cup R_{\mathrm{pred}}}} \tag{11}$$

In this report, we will use accuracy, $P_{\text{G-to-G}}$, and f_{sep} to measure performance of the activity discriminants for prioritizing new molecules for meeting the assay cascade criteria. In the situation that there are only two predicted classes (Green and Red), $P_{\text{G-to-G}}$ is equal to precision. Accuracy and precision are often used to measure performance of classification models.

2.8. Application workflow

All of our *in vitro* screening data were in the AstraZeneca corporate screening database. Our prediction of *in vitro* activities was

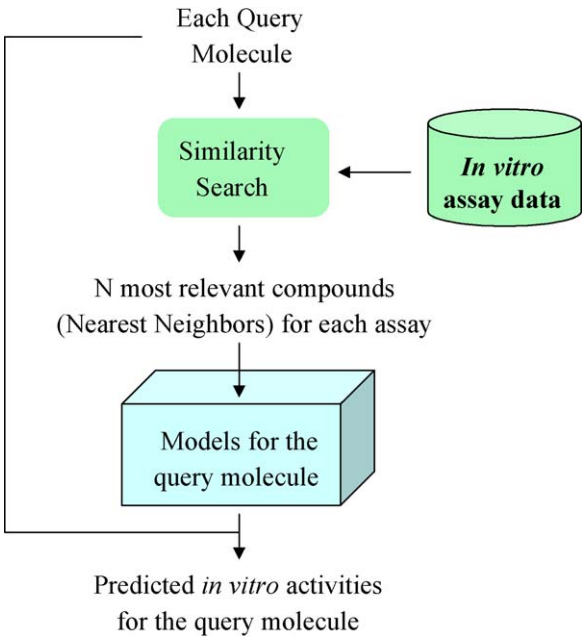

Fig. 1. Calculation of activity discriminants for predicting *in vitro* activities of a query molecule. For each in vitro assay, a similarity search was performed against the compounds that the activities were known and N most similar compounds were selected for performing linear discriminant analysis to compute the activity discriminants.

project-based. Each project had a specific list of assays and cascade criteria that the compounds need to meet in the *in vitro* pharmacological, DMPK and safety activities. Compounds were divided into two activity classes (Green or Red) for each *in vitro* assay based on the cascade criteria.

To prioritize a list of synthetic targets in a given drug discovery project, all the *in vitro* activities were predicted. The activity prediction for each *in vitro* assay is given in Fig. 1. The workflow is as follows:

1. All the data for the *in vitro* assay was extracted from the corporate database and compounds were grouped into two classes. The compounds that cannot be assigned into an activity class were discarded and the compounds that can be classified as Green or Red were maintained for a similarity search for generating training sets of compounds.
2. The fingerprint Tanimoto similarity coefficients between each query compound in the list of synthetic targets and the compounds of known activity classes from the previous step were calculated. For each query molecule, the N most similar compounds were selected.
3. For each query molecule in the list of synthetic targets, a linear discriminant analysis (LDA) was performed with R statistical package using the N most similar compounds and the 6 descriptors.
4. The discriminants of each query molecules and the N most similar compounds were calculated using the technique described in the discriminant analysis section.

After steps 1–4 above for each *in vitro* assay, the results of all the assays were combined and imported into Spotfire (http://www.spotfire.com) for visualization.

In the above workflow, the N most similar compounds were selected for performing the discriminant analysis. We used a fixed number of the most similar compounds in the training set. This number was an input number for each run. Since N was fixed, the fingerprint similarity coefficients between query molecule and the N most similar compounds in the training set varied with query molecules. We used Spotfire to visualize the fingerprint similarity and activity discriminant of each molecule. The visualization allowed medicinal chemists to quickly identify the similarities between the compounds in the training set and the query molecule.

Rather than fixing number of compounds, one can also fix the similarity cutoff to select compounds into the training set. In this case, the number of compounds in the training set would vary with query molecules. For some query molecules, the number of compounds in the training set could be too small for performing the discriminant analysis or could be too large for facile visualization of the analysis results. Although this approach was an attractive option from a purely computational chemistry point of view, it was not particularly useful for prioritizing new molecules for synthesis. Therefore, this approach was not used to support drug discovery projects.

2.9. Visualization of activity discriminants

All the activity discriminants were imported into Spotfire for visualization. Spotfire is a flexible tool for visualizing multi-dimensional data. It can be used with any information source and is widely used in drug discovery. Within the Spotfire window, users can select any individual variable as the *x*-axis and any individual variable as the *y*-axis in the multi-dimensional space to generate a 2D plot, a projection of the multi-dimensional data onto the *x*–*y* plane. In addition, users can quickly select any subset of data that are displayed on the plot by choice buttons or radio buttons or sliders.

In our visualization of *in vitro* activity prediction with activity discriminants, we chose the fingerprint similarity between the compounds in the training set and the query molecule as the *x*-axis and activity discriminant as the *y*-axis. The calculated activity discriminants and other related information of query molecules and the compounds in the training set were imported into Spotfire. In the Spotfire window, each individual assay was associated with a radio button or choice button. Users can generate plots of data of any individual assay or data of all the assays by selecting the radio button or plot the data of any subgroup of assays but selecting the group of the choice buttons. Similarly, users can generate plot of any individual query molecule or any subgroup of query molecules.

Fig. 2a–c shows our Spotfire visualization of the Permeability prediction for three new synthetic targets. For compounds in the training set, the Permeability is coded with our color convention where Green compounds meet our criterion for this assay and Red compounds do not. Since the synthetic target had not been synthesized, the activity is not known and the color is gray. The *x*-axis is the fingerprint similarity between the new synthetic target and the compounds in the training set. By definition, compounds in the training set have fingerprint similarity scores between zero and one, and the synthetic target has similarity score of one. The *y*-axis is LD, the linear discriminant, calculated from the Permeability model.

In addition to providing a Green ($LD < 0$) or Red ($LD > 0$) prediction for the query molecule, the Spotfire visualization aids in evaluating the activity discriminants in two ways: (1) assessing how well the activity discriminants separate Green and Red compounds in the training set, and (2) assessing the relevance of the discriminant analysis to the query molecule in terms of how similar the query molecule is to the compounds in the training set. As shown in Fig. 2a, compounds classified as Green and compounds classified as Red are well separated with their activity discriminants, whereas in Fig. 2b and c, the Green and Red compounds in the training set are somewhat less well separated.

In Fig. 2a, all compounds in the training set that pass the Permeability criterion lie on the lower part of the plot ($LD < 0$) and all compounds in the training set that do not pass the Permeability

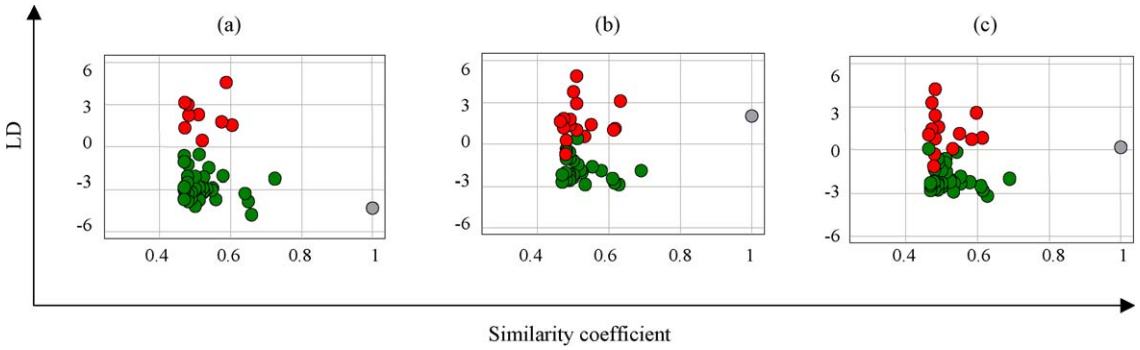

Fig. 2. Visualization of activity discriminants with Spotfire. The *x*-axis is fingerprint similarity between the query molecule and the compounds in the training set. The *y*-axis is the discriminant that best distinguishes the activity. Compounds in the training set that meet the cascade criterion are colored in Green and those that do not meet the cascade criterion are colored in Red. The query molecule, for which the activity is not known, is colored in Grey. (a) A query molecule that will likely meet the cascade criterion; (b) a query molecule that will likely not meet the cascade criterion; (c) a query molecule for which it is difficult predict from the activity discriminant whether the molecule is likely to meet the cascade criterion or not.

assay lie on the upper part (LD > 0). The new synthetic target lies on the lower part of the plot and the activity discriminant is −4.24. Based on this plot, one can predict that the new synthetic target will pass the Permeability assay.

Fig. 2b shows the Permeability prediction of another synthetic target. The new target lies on the upper part of the plot and the activity discriminant is 2.04. While the Red and Green compounds in the training set are not as well separated as in Fig. 2a, the query compound lies well into the range of activity discriminants where all the compounds in the training set fail the Permeability assay. Based on the plot, this new target is predicted to fail.

Fig. 2c shows the Permeability prediction for another potential synthetic target. The activity discriminant of the new target is close to zero, and lies in a range where the activity discriminants do not achieve separation between experimentally Green and Red compounds. This is an example where it was not possible to provide an informative prediction using this technique.

3. Results and discussion

Medicinal chemists often use a computed descriptor to describe a molecular property. One example is *c* Log *P* that describes lipophilicity of a new molecule. Here, we extend into the *in vitro* activities in drug discovery. For each *in vitro* assay, by dividing compounds that have been tested into two activity classes, Green and Red, with the transition criterion defined in the assay cascade and performing linear discriminant analysis with a small number of most similar compounds, we can calculate the activity discriminant of a new molecule. This activity discriminant describes the activity of the new molecule in the assay. If the activity discriminant is less than 0, the new molecule is predicted to pass the assay. Otherwise, it is predicted to fail in the assay. Like other molecular properties, these activity discriminants of *in vitro* assays can be used to prioritize new synthetic targets in order to make molecules that will pass all the *in vitro* assays in the assay cascade of a drug discovery project.

Usefulness of a QSAR model is often validated with a test set of compounds not used in training the model. In most QSAR models, assay dates are not considered when dividing data into training and test sets such that the assay dates of compounds in the training set could be later than that of compounds in the test set. In this case, the results of compounds in the test set only show how well the model explains the existing data (predicting the past) but does not necessarily provide insight toward predicting the future. On the other hand, one can also consider experimental dates when dividing compounds into training and test sets. In this case, the experimental assay dates of all compounds used for training the model are older than the assay dates of the compounds for which the model is used to predict the activities. Therefore, the results measure how well the model can truly predict future measurements. We present results of our activity discriminants in both cases. First, we present the activity discriminants calculated by selecting compounds into training set without considering the experimental assay dates in the data set. This allows us to compare our approach with other approaches where data were separated into training and test sets without considering the dates. We refer to this as the performance for explaining the existing data. We then present the activity discriminants calculated with analysis of only compounds where the experimental assay dates are older than those of the compounds for which we are predicting the activities. The results of the latter case align well with practical usage of the activity discriminants in drug discovery projects for predicting activities of molecules that have not yet been synthesized, or have not been tested experimentally in the assays. We refer this as performance for predicting activities of new molecules.

3.1. Performance of activity discriminants for explaining the existing activity data

To predict the *in vitro* activity of a query molecule in the test set, we searched for N most similar compounds in the entire data set and used these similar compounds to perform discriminant analysis. The number of most similar compounds that were selected into the training sets was the same for all the query molecules in the assay. To find the best number of similar compounds to use, we calculated the prediction accuracy with various numbers of neighbors selected into the training set and chose the number for which the accuracy is highest.

Fig. 3 shows accuracy versus training set size for Permeability. The predicted activities have the highest accuracy when the training set size is 50. When the training set size is less than 50, the accuracy is lower, indicating that there are not enough compounds to formulate the structure–property relationship. When the training set size is larger than 50, the accuracy is also lower, indicating structures of compounds in the training set may be too diverse to formulate the specific structure property relationship.

Table 2 lists the best training set size for all the *in vitro* assay activities. For most *in vitro* activities, the accuracy is highest when the 50 most similar compounds are used for performing the discriminant analysis.

The performance of the activity discriminants for predicting activities of the 11 in vitro assays is listed in Table 3. There are three accuracy measures for each assay. The column named “all” is the accuracy calculated with all the compounds in the test set.

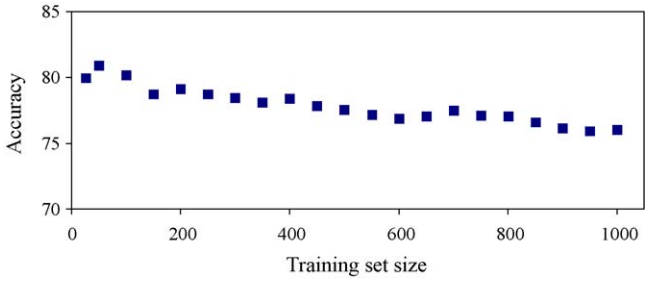

Fig. 3. Accuracy versus training set size for the permeability prediction. For each query molecule, the training set size is the number of most similar compounds selected from the similarity search.

Table 2
Training set sizes that give the most accurate activity discriminants used for explaining the existing *in vitro* activity data. For each assay, the training set size is number of most similar compounds used for training the model to predict the activity of each query molecule.

Model	Training set size that give the most accurate model
Binding K_i	50
GTPγS IC_{50}	100
GTPγS PE	25
Permeability	50
Efflux Ratio	50
hCLint	15
Solubility	50
CYP2D6 IC_{50}	50
CYP3A4 IC_{50}	50
hERG IC_{50}	50
PLD EC_{50}	100

The computed activity discriminants of five assays (GTPγS PE, Solubility, CYP2D6 IC_{50}, CYP3A4 IC_{50}, and hERG IC_{50}) have accuracy very close to 85%. Their accuracy values are 86.7%, 85.3%, 84.2%, 85.1%, and 84.4%, respectively. The predictions of other assays have accuracy close to 80%.

For CYP3A4 IC_{50}, the assay data were used by our colleagues to develop a global model using a support vector machine (SVM) learning technique and very large number of descriptors. The accuracy of the global model was reported to be 83% [68]. The accuracy of our calculated discriminants (85%) is very close to that of the global SVM model (83%). For other activities, no global models of the same data sets were published. However, the prediction accuracy of our activity discriminants for our corporate data sets are close to some recently published global models of corporate data sets of other pharmaceutical companies [61,69]. For hERG activity, the accuracy of a global model of a corporate data set developed with support vector machine was reported to be 85% [69], close to that of our local model (84.4%). For solubility, the accuracy of global models developed with recursive partitioning method and a corporate data set was reported to be between 78% and 81.1% [61], lower than prediction of our calculated activity discriminants (85.3%). While activity discriminant and global model both reach the same level of accuracy for explaining the existing data, our activity discriminant has several other advantages. One advantage is that our activity discriminants could be interpreted easily by medicinal chemists as they are a linear function of only 6 medicinal chemistry descriptors. Another advantage is that our activity discriminants are always up-to-date, whereas the global model may not be updated for many months.

Table 3
Performance of the activity discriminants for explaining the existing *in vitro* activity data.

Model	Accuracy			$P_{G\text{-to-}G}$	$P_{Y\text{-to-}G}$	$P_{R\text{-to-}G}$	f_{sep}
	$\|LD\| \geq 1$	$\|LD\| < 1$	All				
Binding K_i	86.11	60.76	73.31	82.31	48.77	11.74	7.01
GTPγS IC_{50}	84.05	63.78	77.30	65.52	50.27	14.37	4.56
GTPγS PE	89.87	72.55	86.69	92.21	68.63	26.79	3.44
Permeability	88.20	66.37	80.91	86.10	46.40	10.39	8.29
Efflux Ratio	89.52	67.89	81.67	90.27	54.18	11.82	7.64
hCLint	83.30	59.81	77.16	76.70	45.48	13.41	5.72
Solubility	93.30	66.91	85.25	95.08	52.75	12.13	7.84
CYP2D6 IC_{50}	93.86	66.53	84.23	95.48	47.98	10.48	9.11
CYP3A4 IC_{50}	91.17	63.74	85.08	91.23	60.82	9.76	9.35
hERG IC_{50}	92.17	68.08	84.42	92.92	53.94	10.50	8.85
PLD EC_{50}	88.06	70.27	78.72	86.23	47.75	7.94	10.87

In the in vitro activity prediction based on the activity discriminants, the activity of a query molecule is predicted to be Green if the LD is smaller than 0. In contrast, the activity is predicted to be Red if the LD is larger than 0. When the LD is 0, no prediction can be made. Since the LD is continuous, one would expect that the prediction would not be accurate for the query molecules with an LD close to 0. Also, we found empirically that the predictions are consistently more accurate when $|LD| \geq 1$. To account the accuracy difference for molecules with LD close to 0 and LD far from 0, for each assay, we divided the compounds in the test set into two groups: one with $|LD| \geq 1$ and another one with $|LD| < 1$. Accuracy of the two groups are listed in the columns labeled $|LD| \geq 1$ and $|LD| < 1$, respectively. For $|LD| \geq 1$, the accuracy for the 11 assays range from 83.3% (for hCLint) to 93.9% (for CYP2D6 pIC_{50}). There are 4 (solubility, CYP2D6 IC_{50}, CYP3A4 IC_{50}, hERG IC_{50}) assays with accuracy higher than 90%. Four other assays (GTPγS PE, Permeability, Efflux Ratio, PLD EC_{50}) have accuracy below but very close to 90%. Their accuracy values are 89.9%, 88.2%, 89.5%, and 88.1%, respectively. Only two assays (GTPγS IC_{50} and hCLint) have accuracy below 85%. For $|LD| < 1$, the accuracy for the 11 assays ranges from 59.8% (for hCLint) to 72.6% (for GTPγS PE), significantly lower than that for $|LD| \geq 1$.

Because of the low accuracy with $|LD| < 1$, we assigned a new predicted class ("Yellow") to these compounds to indicate that when applying our activity discriminants to predict *in vitro* activities in drug discovery projects, we have lower confidence in these predictions than for those where $|LD| \geq 1$. With this assignment, for each assay, there are three classes from our activity discriminants: Green with $LD \leq -1$, Red with $LD \geq 1$, and Yellow with $|LD| < 1$. The percentages of compounds in these three predicted classes where the activities were found experimentally to be Green in the assays are listed in Table 3 with column named $P_{G\text{-to-}G}$, $P_{R\text{-to-}G}$, and $P_{Y\text{-to-}G}$, respectively. For those compounds predicted to be Green, the percentages that were found to be Green experimentally is higher than 90% for 6 assays (GTPγS PE, Efflux Ratio, Solubility, CYP2D6 IC_{50}, CYP3A4 IC_{50}, and hERG IC_{50}), more than half of the 11 assays. The lowest value for the 11 assays is 65.5% for GTPγS IC_{50}. However, for this assay, the percentage of experimentally Green compounds in the test set is only 29.0%. Therefore, while the correct prediction of 65.5% may seem less impressive than the predictions for other assays, this level of accuracy is still considered to be useful in a project context; compounds that are predicted to be Green are more than twice as likely to be experimentally Green in the GTPγS assay compared to having no prediction at all.

For compounds predicted to be Red, the percentage of compounds that were found to be Green experimentally is close to 10% for 10 of the 11 assays. The one exception is for GTPγS PE with 26.8% of compounds predicted to be Red but found to be Green experimentally. However, for this assay, the percentage of Green compounds in the test set is 81.3%. Therefore, for this assay, 26.8% is considered to be reasonable since the prediction leads to more than a threefold enrichment compared to having no prediction.

f_{sep}, the ratio of $P_{G\text{-to-}G}$ over $P_{R\text{-to-}G}$, for the 11 assays ranged from 3.4 (for GTPγS PE) to 10.9 (for PLD EC_{50}). The calculated activity discriminants were used by our medicinal chemists to prioritized new molecules for synthesis. The strategy was to synthesize the

Table 4
Training set sizes that give the most accurate activity discriminants used for predicting *in vitro* activities of new molecules that have not been tested in the assays and/or have not been synthesized.

Model	Training set size that give the most accurate model
Binding K_i	100
GTPγS IC_{50}	100
GTPγS PE	100
Permeability	100
Efflux Ratio	50
hCLint	50
Solubility	50
CYP2D6 IC_{50}	25
CYP3A4 IC_{50}	200
hERG IC_{50}	150
PLD EC_{50}	100

Table 5
Performance of the activity discriminants for predicting *in vitro* activities of new molecules that have not been tested in the assays and/or have not been synthesized.

Assay	Accuracy			$P_{G\text{-to-}G}$	$P_{Y\text{-to-}G}$	$P_{R\text{-to-}G}$	f_{sep}
	$\|LD\| \geq 1$	$\|LD\| < 1$	All				
Binding K_i	79.46	56.58	68.23	70.19	48.18	13.40	5.24
GTPγS IC_{50}	80.11	52.53	70.27	51.72	44.95	17.38	2.98
GTPγS PE	87.31	68.83	82.19	88.31	68.83	35.29	2.50
Permeability	87.32	67.23	77.97	83.99	50.32	10.94	7.68
Efflux Ratio	87.36	62.82	77.43	87.14	57.42	12.31	7.08
hCLint	81.55	58.62	72.46	75.17	45.41	16.13	4.66
Solubility	90.88	63.45	82.15	92.96	55.26	15.58	5.97
CYP2D6 IC_{50}	86.37	64.71	79.51	88.60	52.49	19.26	4.60
CYP3A4 IC_{50}	86.80	66.67	82.10	86.85	63.33	15.38	5.65
hERG IC_{50}	89.66	64.93	78.70	91.86	54.90	20.00	4.59
PLD EC_{50}	81.14	65.64	74.00	81.01	45.64	18.57	4.36

molecules that were predicted to be Green in most assays and not synthesize the molecules that were predicted to be Red in any assay. The high ratio indicates that the predictions would be useful for reducing the number of compounds that needed to be synthesized in order to find compounds that were Green for all the *in vitro* activities. Since the activity discriminants were calculated with analysis of compounds selected into the training set without considering the dates, the results in this section do not reflect the actual performance for prioritizing the new molecules.

3.2. Performance of activity discriminants for predicting in vitro activities of new molecules

To evaluate the performance for predicting activities of new molecules, we took into account the assay dates while selecting compounds for the training sets. To predict the *in vitro* activities of a query molecule in the test set, we first extracted compounds from the data set where the assay dates were older than the assay date of the query molecule. These compounds with older test dates were then used for performing the similarity searches to select the N most similar compounds and subsequently for calculating activity discriminants to predict the activities of the query molecule. This calculation of activity discriminants emulates the actual usage of activity discriminants in predicting activities of new molecules, where either the new molecules had not been synthesized or, if in the compound library, had never been tested in the assays.

Table 4 lists the optimal training set sizes for calculating activity discriminants of new molecules. When moving from explaining existing activities to predicting activities of new molecules, the optimal training set size remains the same for 4 assays (GTPγS IC_{50}, Efflux Ratio, Solubility, PLD EC_{50}), increases for 6 assays (Binding K_i, GTPγS PE, Permeability, hCLint, CYP3A4 IC_{50}, and hERG IC_{50}) and decreases for 1 assay (CYP2D6 IC_{50}).

Table 5 shows the performance of activity discriminants used for predicting activities of new molecules. The activity discriminants were computed with optimal training sizes listed in Table 4. The performance for predicting activities of new molecules is lower than that of explaining existing data for all the 11 assays. It is consistent with our expectation that activity discriminants appear to be more accurate when explaining the past versus predicting the future; further, it is important to keep in mind that validation for predicting the future is a better representation of the performance a project team can expect when using the activity discriminants.

For $|LD| \geq 1$, the accuracy of predicting activities of new molecules ranges from 79.5% to 90.9%, lower than that of explaining the existing activity data listed in Table 3. However, the drop in accuracy when moving to predicting activities of new molecules is quite small for most assays. We believe this is a consequence of the dynamic nature of the discriminant analysis in that compounds used for the analysis are always up-to-date. The largest drop was found to be 6.9% for PLD EC_{50}. The drop is below 3% for all the 4 most commonly used DMPK properties (Permeability, Efflux Ratio, Solubility, and hCLint).

Column $P_{G\text{-to-}G}$ in Table 5 lists the percent of compounds that were predicted to be Green and later found to be Green in the assays. These numbers are most interesting to the medicinal chemists. When medicinal chemists use the models to select molecules for synthesis, this is the expected percentage that will be Green after compounds have been synthesized and tested in the assays. For most assays, the percentage is higher than 80%. The lowest value is 51.7% for GTPγS IC_{50}. However, as explained in the previous section, this value is considered to be quite good for this assay, since molecules selected with the model are still positively enriched by nearly a factor of 2.

The ratio (f_{sep}) of $P_{G\text{-to-}G}$ over $P_{R\text{-to-}G}$ listed in Table 5 shows how useful the activity discriminants can be used for prioritizing compounds for synthesis. The values indicate how much better the molecules predicted to be Green are than molecules predicted to be Red. Most assays have values higher than 4. Based on the numbers, when medicinal chemists synthesize compounds that are predicted to be Green and do not synthesize compounds that are predicted to be Red, the number of compounds that need to be synthesized to get one or more compounds to pass the *in vitro* assay cascade in its entirety will be reduced very significantly.

Our activity discriminants have been used by medicinal chemists to prioritize new molecules for synthesis for more than 7 lead optimization projects in the past three years. Although we had not developed an advanced informatics infrastructure for medicinal chemists to track the accuracy of the prediction, the assay dates recorded in our screening database have allowed us to go back to any given date to generate the activity discriminants and make predictions. By selecting only those compounds that had been tested experimentally in assays before the query molecule, the activity discriminants generated now are exactly the same as the ones that could have been developed before the query molecule was actually synthesized. The performance therefore reflects the actual performance for predicting activities of new molecules.

3.3. Application domain

The drug discovery process is often divided into lead generation and lead optimization phases. During lead generation, lead compounds with new scaffolds or of different chemical classes, which may have the potential to be modified into candidates for clinical trials, are identified. These lead compounds often have activities that are good for some *in vitro* assays but bad for some others. They need to be modified in lead optimization in order to meet the criteria of all the assays in the assay cascade. During lead optimization, medicinal chemists often analyze structure and activities of compounds that have been tested, make some small changes to

the structures in order to make the new molecules that will have better activities for the assays.

To calculate activity discriminants of a new molecule, some similar compounds that have been synthesized and tested previously are required. For this reason, the approach presented in this paper can only be applied to the projects in lead optimization phase in which medicinal chemists are making small change to the compounds in order to improve the activities. For predicting activities of molecules of new chemical classes in lead generation, the approach cannot be applied since no similar compounds have been tested.

3.4. Future direction

The approach presented in this paper is not good for the discovery of new chemical classes, which often requires docking and/or aligning to 3D pharmacophore models. We are working to extend the QSAR models to include descriptors generated from aligning query molecules to the multiple pharmacophore models and descriptors of ligand–receptor interaction generated from docking in order to predict activities of new chemical classes.

The current approach utilized a fixed number of similar compounds for each in vitro assay to perform linear discriminant analysis and assessed the usability of the activity discriminant by visually inspecting the Spotfire plots. We are working on extracting useful information analytically from the plots to characterize the relationship between the query molecule and the compounds in the training set and on varying number of compounds used in training based on the similarity profile in order to extend the approach to query molecules of low similarity. The variation allows training set size ranging from 50 for those query molecules that have many very similar compounds to the total number of compounds that have been tested (the training set for the global model) for the query molecules that have no similar compounds.

4. Further comments and conclusion

We demonstrated here that we could compute activity discriminants and use them to predict *in vitro* pharmacological, DMPK, and safety activities of new molecules quite accurately. Our approach emulates how medicinal chemists perform SAR analysis and activity prediction. Since the activity discriminants are linear functions of 6 medicinal chemistry descriptors, they can be interpreted easily by medicinal chemists and therefore are useful for guiding design of new molecules. The results show that our activity discriminants are as accurate as those predictive models trained with very large number of descriptors for explaining existing activity data. Since our discriminant analyses are always up-to-date, the drop in accuracy is very small when moving (from just explaining the existing data) to predicting activities of new molecules, making our activity discriminants particularly useful for medicinal chemists to prioritize new molecules for synthesis.

Acknowledgements

We are grateful for David Cosgrove for providing the similarity search tool that is an important part of our modeling and predicting workflow. We thank Steven Wesolowski and James Damewood for helpful discussion and editing.

References

[1] K.C. Cheng, C. Li, A.S. Uss, Prediction of oral drug absorption in humans – from cultured cell lines and experimental animals, Expert Opin. Drug Metab. Toxicol. 4 (2008) 581–590.
[2] L. Di, E.H. Kerns, X.J. Ma, Y. Huang, G.T. Carter, Applications of high throughput microsomal stability assay in drug discovery, Comb. Chem. High Throughput Screen. 11 (2008) 469–476.
[3] H. Wan, A.G. Holmen, High throughput screening of physicochemical properties and *in vitro* ADME profiling in drug discovery, Comb. Chem. High Throughput Screen. 12 (2009) 315–329.
[4] J. Wang, L. Urban, D. Bojanic, Maximizing use of *in vitro* ADMET tools to predict *in vivo* bioavailability and safety, Expert Opin. Drug Metab. Toxicol. 3 (2007) 641–665.
[5] M. Zientek, H. Miller, D. Smith, M.B. Dunklee, L. Heinle, A. Thurston, C. Lee, R. Hyland, O. Fahmi, D. Burdette, Development of an in vitro drug–drug interaction assay to simultaneously monitor five cytochrome P450 isoforms and performance assessment using drug library compounds, J. Pharmacol. Toxicol. Methods 58 (2008) 206–214.
[6] Y.W. Alelyunas, R. Liu, L. Pelosi-Kilby, C. Shen, Application of a dried-DMSO rapid throughput 24-h equilibrium solubility in advancing discovery candidates, Eur. J. Pharm. Sci. 37 (2009) 172–182.
[7] G. Zlokarnik, P.D.J. Grootenhuis, J.B. Watson, High throughput P450 inhibition screens in early drug discovery, Drug Discov. Today 10 (2005) 1443–1450.
[8] M. Culot, S. Lundquist, D. Vanuxeem, S. Nion, C. Landry, Y. Delplace, M. Dehouck, V. Berezowski, L. Fenart, R. Cecchelli, An *in vitro* blood–brain barrier model for high throughput (HTS) toxicological screening, Toxicol. In Vitro 22 (2008) 799–811.
[9] T.J. Raub, B.S. Lutzke, P.K. Andrus, G.A. Sawada, B.A. Staton, Early preclinical evaluation of brain exposure in support of hit identification and lead optimization, Biotechnol.: Pharm. Aspects 4 (2006) 355–410.
[10] M. Cik, M.R. Jurzak, High-throughput and high-content screening, Compr. Med. Chem. II 3 (2006) 679–696.
[11] A. Weissman, J. Keefer, A. Miagkov, M. Sathyamoorthy, S. Perschke, F.L. Wang, Cell-based screening assays, Compr. Med. Chem. II 3 (2006) 617–646.
[12] B. Faller, J. Wang, A. Zimmerlin, L. Bell, J. Hamon, S. Whitebread, K. Azzaoui, D. Bojanic, L. Urban, High-throughput *in vitro* profiling assays: lessons learnt from experiences at Novartis, Expert Opin. Drug Metab. Toxicol. 2 (2006) 823–833.
[13] S. Whitebread, J. Hamon, D. Bojanic, L. Urban, Keynote review: *in vitro* safety pharmacology profiling: an essential tool for successful drug development, Drug Discov. Today 10 (2005) 1421–1433.
[14] J.K. Morelli, M. Buehrle, F. Pognan, L.R. Barone, W. Fieles, P.J. Ciaccio, Validation of an *in vitro* screen for phospholipidosis using a high-content biology platform, Cell Biol. Toxicol. 22 (2006) 15–27.
[15] M.H. Bridgland-Taylor, A.C. Hargreaves, A. Easter, A. Orme, D.C. Henthorn, M. Ding, A.M. Davis, B.G. Small, C.G. Heapy, N. Abi-Gerges, F. Persson, I. Jacobson, M. Sullivan, N. Albertson, T.G. Hammond, E. Sullivan, J.P. Valentin, C.E. Pollard, Optimisation and validation of a medium-throughput electrophysiology-based hERG assay using ionworks HT, J. Pharmacol. Toxicol. Methods 54 (2006) 189–199.
[16] S. Sommer, S. Kramer, Three data mining techniques to improve lazy structure–activity relationships for noncongeneric compounds, J. Chem. Inf. Model. 47 (2007) 2035–2043.
[17] R. Guha, D. Dutta, P.C. Jurs, T. Chen, Local lazy regression: making use of the neighborhood to improve QSAR predictions, J. Chem. Inf. Model. 46 (2006) 1836–1847.
[18] H. Yuan, Y. Wang, Y. Cheng, Mode of action-based local QSAR modeling for the prediction of acute toxicity in the fathead minnow, J. Mol. Graph. Model. 26 (2007) 327–335.
[19] H. Yuan, Y. Wang, Y. Cheng, Local and global quantitative structure–activity relationship modeling and prediction for the baseline toxicity, J. Chem. Inf. Model. 47 (2007) 159–169.
[20] S. Zhang, A. Golbraikh, S. Oloff, H. Kohn, A. Tropsha, A novel automated lazy learning QSAR(ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models, J. Chem. Inf. Model. 46 (2006) 1984–1995.
[21] H. Zhang, H.Y. Ando, L. Chen, P.H. Lee, On-the-fly selection of a training set for aqueous solubility prediction, Mol. Pharm. 4 (2007) 489–497.
[22] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, 2000.
[23] V.N. Viswanadhan, A.K. Ghose, G.R. Revankar, R.K. Robins, Atomic physicochemical parameters for three-dimensional structure directed quantitative structure–activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics, J. Chem. Inf. Comput. Sci. 29 (1989) 163–172.
[24] H. Zhu, A. Sedykh, S.K. Chakravarti, G. Klopman, A new group contribution approach to the calculation of log *P*, Curr. Comput. Aided Drug Des. 1 (2005) 3–9.
[25] S. Bhavani, A. Nagargadde, A. Thawani, V. Sridhar, N. Chandra, Substructure-based support vector machine classifiers for prediction of adverse effects in diverse classes of drugs, J. Chem. Inf. Model. 46 (2006) 2478–2486.
[26] T.J. Hou, K. Xia, W. Zhang, X.J. Xu, ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach, J. Chem. Inf. Comput. Sci. 44 (2004) 266–275.
[27] J. Kotecha, S. Shah, I. Rathod, G. Subbaiah, Prediction of oral absorption in humans by experimental immobilized artificial membrane chromatography indices and physicochemical descriptors, Int. J. Pharm. 360 (2008) 96–106.
[28] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges, Tetrahedron 36 (1980) 3219–3222.
[29] P.A. Labute, Widely applicable set of descriptors, J. Mol. Graph. Model. 18 (2000) 464–477.

[30] J. Zhang, T. Kleinoeder, J. Gasteiger, Prediction of pK_a values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors, J. Chem. Inf. Model. 46 (2006) 2256–2266.
[31] P. Ertl, B. Rohde, P. Selzer, Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties, J. Med. Chem. 43 (2000) 3714–3717.
[32] S. Prasanna, R.J. Doerksen, Topological polar surface area: a useful descriptor in 2D-QSAR, Curr. Med. Chem. 16 (2009) 21–41.
[33] T.J. Hou, X.J. Xu, ADME evaluation in drug discovery. 3. Modeling blood–brain barrier partitioning using simple molecular descriptors, J. Chem. Inf. Comput. Sci. 43 (2003) 2137–2152.
[34] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of MDL keys for use in drug discovery, J. Chem. Inf. Comput. Sci. 42 (2002) 1273–1280.
[35] R.E. Carhart, D.H. Smith, R. Venkataraghavan, Atom pairs as molecular features in structure–activity studies: definition and applications, J. Chem. Inf. Comput. Sci. 25 (1985) 64–73.
[36] M. Hassan, R.D. Brown, S. Varma-O'brien, D. Rogers, Cheminformatics analysis and learning in a data pipelining environment, Mol. Divers. 10 (2006) 283–299.
[37] U. Fechner, J. Paetz, G. Schneider, Comparison of three holographic fingerprint descriptors and their binary counterparts, QSAR Comb. Sci. 24 (2005) 961–967.
[38] Y. Marrero-Ponce, R. Medina-Marrero, F. Torrens, Y. Martinez, V. Romero-Zaldivar, E.A. Castro, Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity, Bioorg. Med. Chem. 13 (2005) 2881–2899.
[39] S. Askjaer, M. Langgrd, Combining pharmacophore fingerprints and PLS-discriminant analysis for virtual screening and SAR elucidation, J. Chem. Inf. Model. 48 (2008) 476–488.
[40] F. Bonachera, D. Horvath, Fuzzy tricentric pharmacophore fingerprints. Application of topological fuzzy pharmacophore triplets in quantitative structure–activity relationships, J. Chem. Inf. Model. 48 (2008) 409–425.
[41] L.B. Kier, H. Hall, Molecular Connectivity in Structure–Activity Analysis, Wiley, NY, 1986.
[42] X.H. Li, A.F. Jalbout, M. Solimannejad, Definition and application of a novel valence molecular connectivity index, THEOCHEM 663 (2003) 81–85.
[43] Q.N. Hu, Y.Z. Liang, H. Yin, X.L. Peng, K.T. Fang, Structural interpretation of the topological index. 2. The molecular connectivity index, the kappa index, and the atom-type E-state index, J. Chem. Inf. Comput. Sci. 44 (2004) 1193–1201.
[44] L.B. Kier, L.H. Hall, The prediction of ADMET properties using structure information representations, Chem. Biodivers. 2 (2005) 1428–1437.
[45] L.B. Kier, L.H. Hall, Molecular Structure Description. The Electrotopological State, Academic Press, NY, 1999.
[46] O. Ivanciuc, Electrotopological state indices, Methods Princ. Med. Chem. 37 (2008) 85–109.
[47] L.H. Hall, L.B. Kier, L.M. Hall, Electrotopological state indices to assess molecular and absorption, distribution, metabolism, excretion, and toxicity properties, Compr. Med. Chem. II 5 (2006) 555–576.
[48] R.D. Cramer, D.E. Patterson, J.D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, J. Am. Chem. Soc. 110 (1988) 5959–5967.
[49] A. Hillebrech, G. Klebe, Use of 3D QSAR models for database screening: a feasibility study, J. Chem. Inf. Model. 48 (2008) 384–396.
[50] M. Pastor, G. Cruciani, I. McLay, S. Pickett, S. Clementi, GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors, J. Med. Chem. 43 (2000) 3233–3243.
[51] M. Pastor, Alignment-independent descriptors from molecular interaction fields, Methods Princ. Med. Chem. 27 (2006) 117–143.
[52] P. Gedeck, B. Rohde, C. Bartels, QSAR – how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets, J. Chem. Inf. Model. 46 (2006) 1924–1936.
[53] G. Cruciani, P. Crivori, P.A. Carrupt, B. Testa, Molecular fields in quantitative structure–permeation relationships: the VolSurf approach, THEOCHEM 503 (2000) 17–30.
[54] R. Mannhold, G. Berellini, E. Carosati, P. Benedetti, Use of MIF-based VolSurf descriptors in physicochemical and pharmacokinetic studies, Methods Princ. Med. Chem. 27 (2006) 173–196.
[55] B.B. Goldman, W.P. Walters, Machine learning in computational chemistry, Annu. Rep. Comput. Chem. 2 (2006) 127–140.
[56] C.L. Bruce, J.L. Melville, S.D. Pickett, J.D. Hirst, Contemporary QSAR classifiers compared, J. Chem. Inf. Model. 47 (2007) 219–227.
[57] F.J. Prado-Prado, V.O. Martinez de la, E. Uriarte, M. Ubeira, M. Florencio, K.C. Chou, H. Gonzalez-Diaz, Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug–drug complex networks, Bioorg. Med. Chem. 17 (2009) 569–575.
[58] R. Lior, M. Oded, Data Mining with Decision Trees: Theory and Applications, World Scientific Publishing, Singapore, 2008.
[59] A. Rusinko, M.W. Farmen, C.G. Lambert, P.L. Brown, S.S. Young, Analysis of a large structure biological activity data set using recursive partitioning, J. Chem. Inf. Comput. Sci. 39 (1999) 1017–1026.
[60] W. Tong, H. Hong, H. Fang, Q. Xie, R. Perkins, Decision forest: combining the predictions of multiple independent decision tree models, J. Chem. Inf. Comput. Sci. 43 (2003) 525–531.
[61] C. Lamanna, M. Bellini, A. Padova, G. Westerberg, L. Maccari, Straightforward recursive partitioning model for discarding insoluble compounds in the drug discovery process, J. Med. Chem. 51 (2008) 2891–2897.
[62] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, NY, 1995.
[63] J. Doucet, F. Barbault, H. Xia, A. Panaye, B. Fan, Nonlinear SVM approaches to QSPR/QSAR studies and drug design, Curr. Comput. Aided Drug Des. 3 (2007) 263–289.
[64] D.A. Winkler, F.R. Burden, Bayesian neural nets for modeling in drug discovery, Drug Discov. Today: BIOSILICO 2 (2004) 104–111.
[65] D.A. Nugiel, J.R. Krumrine, D.C. Hill, J.R. Damewood, P.R. Bernstein, C.D. Sobotka-Briner, J. Liu, A. Zacco, M.E. Pierson, De novo design of a picomolar nonbasic 5-HT1B receptor antagonist, J. Med. Chem. 53 (2010) 1876–1880.
[66] D. Cosgrove, AlFi – an alternative to daylight fingerprints, AstraZeneca Internal Software Document.
[67] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 23 (1997) 3–25.
[68] D. Zhou, R. Liu, S.A. Otmani, S.W. Grimm, R.J. Zauhar, I. Zamora, Rapid classification of CYP3A4 inhibition potential using support vector machine approach, Lett. Drug Des. Discov. 4 (2007) 192–200.
[69] L. Jia, H. Sun, Support vector machines classification of hERG liabilities based on atom types, Bioorg. Med. Chem. 16 (2008) 6252–6260.