JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Capturing Structure−Activity Relationships from Chemogenomic Spaces

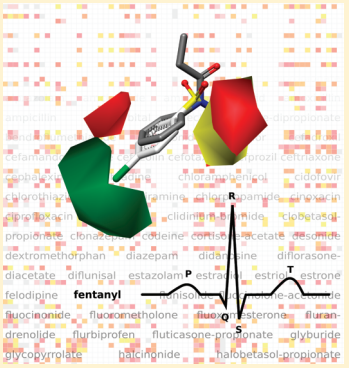
Bernd Wendt,[*,†,§] Ulrike Uhrig,[‡] and Fabian Bös[‡]

[†]European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, Heidelberg, D-69117 Germany

[‡]Tripos International, Martin-Kollar-Strasse 17, Munich, D-81829 Germany

Supporting Information

ABSTRACT: Modeling off-target effects is one major goal of chemical biology, particularly in its applications to drug discovery. Here, we describe a new approach that allows the extraction of structure−activity relationships from large chemogenomic spaces starting from a single chemical structure. Several public source databases, offering a vast amount of data on structure and activity for a large number of different targets, have been investigated for their usefulness in automated structure−activity relationships (SAR) extraction. SAR tables were constructed by assembling similar structures around each query structure that have an activity record for a particular target. Quantitative series enrichment analysis (QSEA) was applied to these SAR tables to identify trends and to transform these trends into topomer CoMFA models. Overall more than 1700 SAR tables with topomer CoMFA models have been obtained from the ChEMBL, PubChem, and ChemBank databases. These models were able to highlight the structural trends associated with various off-target effects of marketed drugs, including cases where other structural similarity metrics would not have detected an off-target effect. These results indicate the usefulness of the QSEA approach, particularly whenever applicable with public databases, in providing a new means, beyond a simple similarity between ligand structures, to capture SAR trends and thereby contribute to success in drug discovery.



INTRODUCTION

In the target-based approach of lead discovery, molecular targets are usually identified and validated before lead discovery starts; assays and screens are then used to identify lead compounds against the target of interest.[1] Progressing from target selection through lead discovery to lead optimization, knowledge is generated about the primary target and also a few secondary targets. However, the biologically local scope of such in vitro-derived knowledge, even supplemented by information conventionally extracted from research and patent literature,[2] may prove altogether insufficient with the first in vivo experiments. When exposed to whole cell or animal models, the lead compound is no longer exposed to just a few selected targets. Rather, the candidate compound can now interact with a large number of mostly unknown targets, possibly yielding a polypharmacological network.[3,4]

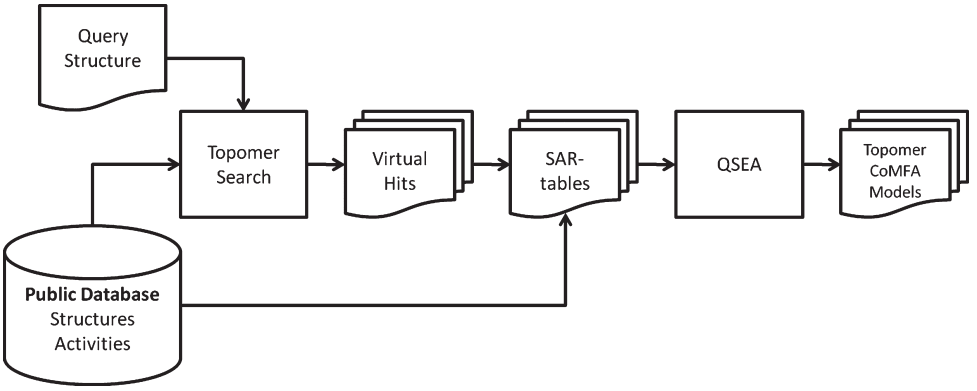
In the phenotype-based approach, lead molecules are identified that produce a desired phenotype. Hereby, usually high-content screening methods are applied to identify, characterize, and optimize the lead compounds for their capacity to have the desired impact on a cell or subcellular function. However, for clinical evaluation of lead compounds, it is required that their mechanism of action is elucidated.[5] Yet, target deconvolution, used to unveil the molecular targets that underlie the observed phenotypic effects, still represents a challenging and resource intensive adventure.[6,7] Thus, irrespective of the discovery approach taken, a detailed understanding of the targets for a particular lead compound is often crucial for successful drug discovery.[8]

Screening multiple candidate compounds against the thousands of accessible off-targets does not seem to be a practical option. Therefore in silico approaches for prediction of off-target effects have become more and more important.[9−11] Several attempts have been reported where cheminformatic methods have been applied to relate protein and ligand structures. From the protein perspective, as one example, Kuhn et al.[12] generate mappings of biological and chemical spaces by quantifying the probability of two proteins to bind the same ligand as a function of their sequence similarity. Campillo[13] uses side effects shared by two molecules to link them to a common target. From the ligand perspective, Jenkins[14] reviewed different approaches that use chemical similarities of small molecules to identify a biological target for a ligand structure. Mestres[15] takes into consideration the available ligand−target interaction matrices for prediction of target profiles.

Although individual mappings between chemical and biological spaces[16] represent useful starting points, it reduces chemical biology information to merely links between a ligand and a target.[17] Each of these mappings raises new questions: What is the significance of the reported activity? What is the underlying mechanism of action? Unexpectedly uncovering a possible link of the current lead compound to an unwanted off-target effect can substantially redirect a lead optimization project by triggering

Received: July 13, 2010
Published: March 16, 2011

Scheme 1. Schematic View of the Workflow



new studies with the goal of determining "What within our structure is causing this effect?"

Although these questions are challenging, with the advent of the chemogenomics era, vast amounts of structure−activity information have been generated and put into public use.[18] Activities such as the NIH-roadmap initiative[19] (PubChem) or the NCI-sponsored activity within the Initiative for Chemical Genetics[20] (ChemBank) gave birth to huge public source databases containing millions of structure−activity data points. The ChEMBL[21] database is another large chemical biology repository being maintained and amended by the EBI. In addition, smaller databases such as BindingDB,[22] Binding-MOAD,[23] and K_i-Database[24] exist that contain manually curated information extracted from the scientific literature. However, before these huge efforts to generate and curate experimental bioassay data can productively influence any individual discovery project, cheminformatic approaches that enable their analysis and mining[25] must be developed and deployed.

The main question that our studies intended to address was this: "Given a particular compound, we want to investigate the general usefulness of these public source databases for the extraction of relevant structure−activity data, in particular for off-target prediction". Therefore, the recently reported method of Quantitative Series Enrichment Analysis (QSEA)[26] was embedded into a workflow for mining of structure−activity relationships. A selection of 250 marketed drugs was used to represent a diverse pool of lead-like chemical structures. Each of these structures was used as a query to conduct a topomer search[27] in public source databases to find shape- and pharmacophore-similar structures with associated activity data against one or more targets. For each query, the results from the search were assembled into target- or assay-specific SAR tables. To each of these SAR tables, the QSEA method was applied, seeking any topomer CoMFA[28] models with sufficient statistical significance. The resulting topomer CoMFA models were finally analyzed as to whether they are able to reveal chemically tractable structure−activity relationships. In the context of the query structure, this is a necessary property to usefully help guide lead optimization.

METHODS

All modeling work was done using SYBYL 8.0[29] and a nonreleased version of SYBYL 7.1 with topomer routines implemented. Automation of all routines was achieved through Python scripts accessing a special API to the topomer routines. Topomer CoMFA models and predictions were automated by means of SPL (SYBYL Programming Language).

Ligand Data Set. The ligand data set used in the present study comprises a selection of 250 marketed drugs as collected by Cleves and Jain[9] (see also Table S1 of the Supporting Information). These drugs are ligands active against one of 22 different therapeutic targets, thus representing a diverse selection of pharmacology and biology. All the query structures have been standardized using the program dbtranslate with ionization rules defined in csp_standard.defs[30] followed by 3-D coordinate generation using the program Concord (version 6.1).[31]

Public Source Databases. Structures and associated activity information were downloaded from seven publicly available databases during October 2008 and February 2009 for the ChEMBL database in February 2010 (Table 1).

For a subset of the PubChem compound database, biological activity information is available. Both the structural part (in SD format) and the activity information (CSV format) were downloaded from the PubChem ftp-site. The structures and associated activity information of the Chembank database was obtained by making use of the Chembank SOAP interface. For the ChEMBL database, structures were downloaded in SD format, and biological activity information was downloaded by selecting all available biological activities for all the resulting structures from the similarity searches (as provided by the Web interface of the ChEMBL database). The structures and activity data of the other databases were downloaded in SMILES and text formats.

All the structures in these candidate data sets were converted into SLN format followed by standardization of functional groups using the program dbtranslate with ionization rules defined in csp_standard.defs. The program Concord (version 6.1) was used to generate 3-D coordinates.

Virtual Screening. The 250 structures of the Ligand data set were used as queries in a topomer similarity search against the downloaded and prepared databases using the program dbtop.[32] As cutoff criterion, a maximum topomer distance of 180 (default value) was applied. For comparison purposes a conventional "Tanimoto" similarity search was also done, using the program dbsearch and the UNITY fingerprint descriptor. Here, the similarity cutoff was chosen at a Tanimoto coefficient of 0.85.

SAR Table Construction. From the structures identified by the topomer similarity search, SAR tables were constructed for

Table 1. Public Source Databases Used in the Current Study

database	SAR data points	assays/targets	source	date of download
PubChem	39,323,334	1273 (assays)	http://pubchem.ncbi.nlm.nih.gov	20.10.2008
ChemBank	8,795,367	1744 (assays)	http://chembank.broad.harvard.edu	17.02.2009
ChEMBL	2,400,000	7192 (targets)	http://www.ebi.ac.uk/ChEMBLdb	19.02.2010
Binding-DB	46,723	618 (enzymes)	http://www.bindingdb.org	17.12.2008
K_i-Database	27,743	546 (enzymes)	http://pdsp.med.unc.edu	09.12.2008
Binding MOAD	3,410	426 (EC numbers)	http://www.bindingmoad.org	08.12.2008
Affinity-DB	748	474 (PDB ids)	http://pc1664.pharmazie.uni-marburg.de/affinity	08.12.2008

every combination of query structure and assay-id/target name within a database. This was achieved by mapping the biological activities to the corresponding structures using the database's internal compound identifiers. No attempts were made to transform the reported biological activities into a form most suitable for QSAR analysis. Instead, in the case of the PubChem database, the PubChem activity score was used; this usually ranges from 0 to 100. For activities from the ChemBank database, the reported composite Z-scores were used. For all other databases, the reported IC_{50} or K_i value was chosen as activity value. Construction of SAR tables from the ChEMBL databases needed a special procedure because for a given target a mixture of assay types, units, functions, and organisms could be expected.

Many of the public databases also store assay data that is not target related (i.e., fluorescence measurements). Although potentially helpful to reduce the number of SAR tables, no filtering was applied to separate target-related assay data from the rest.

Quantitative Series Enrichment Analysis (QSEA). The topomer CoMFA methodology on which QSEA depends differs from conventional CoMFA only in using topomers as the requisitely aligned 3-D structures. For QSEA, the particular fragment pair then used to represent each molecular structure is automatically chosen as the one providing the lowest mean topomer dissimilarity to the other structures in the SAR table.

All SAR tables with 10 or more structures were processed through the QSEA protocol. A brief summary of the QSEA method is given as follows (for a more detailed description see ref 26). QSEA begins with designation of a centroid compound within the data set on the basis of its similarity distance matrix constructed from the topomer descriptor/distance between the compounds. The structure of the centroid compound is then split into all possible two-piece fragmentations using available single bonds. The resulting fragments were not allowed to be smaller than three heavy atoms. For all resulting two-piece fragmentations, the other structures of the SAR table are checked for their best matching two-piece fragmentation. These sets of corresponding two-piece fragmentations are compared on the basis of their overall mean topomeric distance. The set with the smallest topomeric distance mean is chosen as input for the Topomer CoMFA routine. A spreadsheet is created in SYBYL from the SAR table, and for each structure, the two fragments from the chosen set of two-piece fragmentations are imported. Their resulting topomer conformations are used to calculate steric and electrostatic topomeric fields. In addition, the corresponding topomeric distances between the centroid and the remainder compounds are loaded into the spreadsheet and are used to sort the table relative to this distance in ascending order. Starting from the top of the spreadsheet, partial least-squares (PLS) is applied to the centroid compound and its two closest neighbors to seek a correlation of the topomeric field information with the reported biologic activities. The resulting topomer CoMFA model is then used to predict the activities of the other structures in the SAR table. Then, in an iterative process, the next closest neighbor compound is added to the training set, again followed by PLS analysis and activity predictions for the remaining structures, until all structures have been included in the training set. Statistical parameters such as q^2, r^2, number of components, predictive r^2, and standard errors of prediction are recorded at each step.

RESULTS AND DISCUSSION

Virtual Screening. As expected, the number of hits obtained from a search in a particular database strongly depends on database size and composition. There is a clear distinction between the size of larger databases such as ChEMBL, PubChem, and ChemBank with millions of data points giving rise to virtual hits of more than 39,000, 30,000, and 10,000 compound hits, respectively, and smaller databases such as Binding DB and K_i-Bank with data points in the tens of thousands and virtual hits between 400 and 6000 (detailed numerical results can be found in Table S2 of the Supporting Information), where virtual hits mean the number of shape-similar compounds for each query summed up over all query structures.

Comparing Topomers to the Tanimoto Similarity Metric. The outcome of virtual screening using topomer similarity as the descriptor was compared to the outcome of virtual screening using the Tanimoto coefficient similarity on the basis of the UNITY fingerprint. Fingerprint-based similarity searching is a popular method for virtual screening when only a single active structure is available.[33] The choice of similarity thresholds is still a matter of debate. Earlier studies suggested a threshold value of 0.85,[34,35] and this value has been recognized as a conservative default.[36,37] Other authors advocate using lower threshold values of 0.7 and even 0.6.[38−40] The use of low similarity thresholds is typical for lead identification stages of a project where sensitivity is privileged over specificity,[41] and the intention is to capture as many true actives as possible at the expense of increasing the number of false positives. Because of the more resource intensive research in lead optimization, the yield in terms of active compounds needs to be higher, and the medicinal chemist must be conservative.[42] For the QSEA method in particular an increased number of false positives would have a negative impact on the determination of a centroid compound. Therefore, in this study, similarity thresholds were kept at conservative levels: 180 for the topomer and 0.85 for the Tanimoto coefficient. On the basis of a conceptual study on similarity−property principles,[43] these values were viewed as performing with comparable effectiveness in forecasting biological similarities.

dx.doi.org/10.1021/ci100270x |J. Chem. Inf. Model. 2011, 51, 843–851

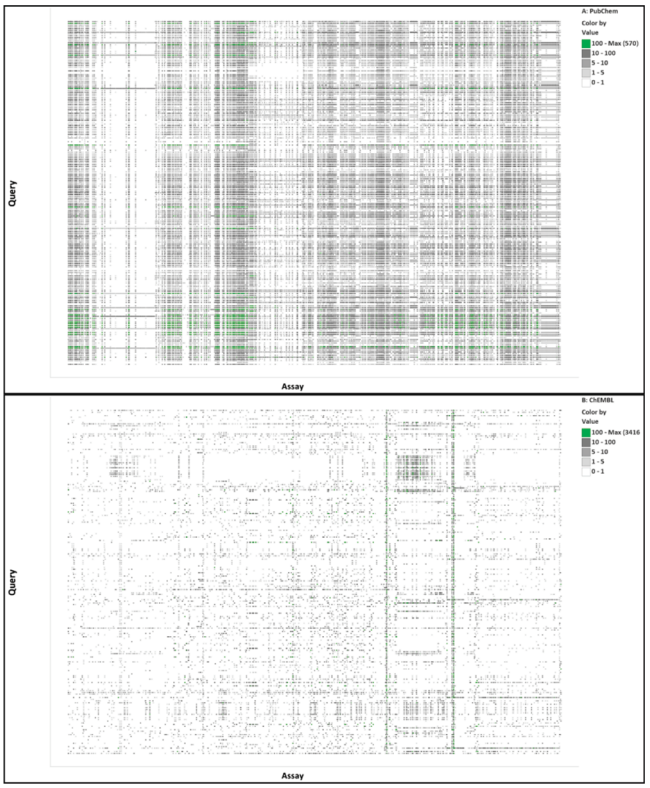


Figure 1. Heatmap representation of the results from virtually screening the PubChem (A) and Chembl (B) databases. Rows represent the 250 query structures. Columns represent the individual assays: 1274 and 2667 for PubChem and ChEMBL, respectively. Coloring of the cells represents the number of structure−activity data points and is as follows: white, 0; light-gray, 1−5; gray, 5−10; dark gray, 10−100; and green, >100.

Overall, the topomer search yielded about an order of magnitude more hits than did the fingerprint search, with the virtual hits found by the fingerprint method often being a subset of the topomer search hits (see Table S2 of the Supporting Information). The much larger number of structures retrievable by topomer similarity has surely increased the scope and very likely the number of useful SAR results. This outcome supports the particular strength of the topomer search method when assembling SAR data sets.

Construction of SAR Tables. Mapping of the corresponding biological activities to all hit structures yielded more than 1.3 million SAR data points for structures derived from the PubChem database. SAR data points from the ChemBank and ChEMBL database were in the range of 450,000 and 410,000, respectively.

The heatmaps in Figure 1 illustrate some trends involving the individual assays and virtual screening searches (taken from the PubChem and ChEMBL databases: white, 0; light-gray, 1−5; gray, 5−10; dark gray, 10−100; and green, >100 hits; a heatmap of the ChemBank result is not shown because it resembles mostly that of the PubChem database). Again as expected, heatmaps in Figure 1 show great variation, among their combination of the 250 similarity search results with the assays, in the numbers of individual assay results. In the heatmap of the PubChem result, for example, green rows represent certain simpler query structures such as acetaminophen or tolbutamide, which returned thousands of compound hits that furthermore had been broadly tested giving rise to many SAR data points. However, other query structures such as terconazole or zidovudine (white rows) yielded no hits (other than the query itself) that had been tested in any assay. This extreme variation may partially reflect the compositions of the underlying screening sets, which probably are heavily populated by high-throughput chemistries that link "building blocks" by such simple reactions as amide- and sulfonamide bond formations. More complex structures produced by longer sequences of more difficult reactions (i.e., C−C bond forming reactions) may be less well covered within these screening sets. There are certain assays giving rise to compound hits for almost any of the query structures and as well as assays for which no matching hits could be found for any of the query structures. This outcome is primarily related to the screening set size used for the particular assay. Many of the PubChem assays are confirmatory or follow-up screens of larger high-throughput campaigns. These screens have a highly reduced screening set size, and therefore, it is less likely to find compound hits against a particular query structure.

In contrast to the stripe pattern shown in the heatmap of the PubChem result, the heatmap of the ChEMBL result shows a more scattered pattern. The matrix of query and assay id is much less populated. The average number of SAR data points is much lower than in the heatmap of the PubChem result, reflecting not only the much lower overall number of SAR data points but also the spread of data points over more than twice as many different targets. It may also reflect the different chemical and biological compositions of the databases, whereas the PubChem compounds belong to general screening compounds that are commercially available and have been used in many screening campaigns. The compounds of the ChEMBL database mostly originate from individual medicinal chemistry projects, where only a limited number of compounds have been prepared exclusively for a limited set of experiments and, thus, are in general not available from commercial sources. The heatmap of the ChEMBL result shows only a few vertical bands, and these belong to rather generic ChEMBL targets with many structures associated to many activity end points, e.g., ADMET (CHEMBL612558) with >50,000 structures/>120,000 activity end points and MUS_MUSCULUS (CHEMBL375) with >40,000 structures/>150,000 activity end points.

Construction of SAR tables from the matrices of query and assay ids yielded 14,830 SAR tables from the PubChem database and 9340 SAR tables from the ChemBank database. All SAR tables were checked for their size and activity spread. SAR tables with less than 10 structures were rejected as being less likely to produce meaningful structure−activity relationships. In addition, SAR tables showing an activity spread of less than 50% or a spread of composite Z-score of less than 4, for tables derived from PubChem and ChemBank, respectively, were viewed as inappropriate for quantifying structure−activity relationships and were therefore rejected too. Application of these two considerations reduced the number of SAR tables to 1767 for the PubChem database and 1076 for ChemBank databases. Because of the variety of activity end points for biological data from the ChEMBL database, no filtering on the basis of activity spread was applied to those SAR tables, and all 2875 SAR tables having more than or equal to 10 structures were considered for model building with QSEA.

Perhaps surprisingly, for the three smaller databases, construction of SAR tables from the virtual screening results proved much

more challenging, partially because of difficulty in handling inconsistencies among the uniformity of the data. Often, there were many duplicate activity records found for the same structure, however, with large variations in the reported activity value. In many cases the activity range reported for the same structure and target could span more than 3 log-units. Also, the biological data in the Binding DB and Binding MOAD consisted of a mixture of reported K_i- and IC^{50} values, and activities from different source organisms were mixed. These inconsistencies prevented the automatic construction of SAR tables.

QSEA

Application of the QSEA method to these 5718 SAR tables yielded 1795 SAR tables having at least one topomer CoMFA model possessing a q^2 (leave-one-out crossvalidated r^2) greater than 0.2. Because of QSEA's systematic exploration of SAR tables, within many of the 1795 SAR tables other significant models were found involving structural subsets. SAR tables from the ChEMBL database yielded the majority of models (1337 models) reflecting the higher quality of dose−response curve data over HTS data, which yielded 217 models and 187 models from SAR tables of the PubChem and ChemBank databases, respectively.

The favorable results from the ChEMBL database seem to be in line with a comparable study published by Olah.[44] He applied an automated PLS routine for QSAR derivation on 1632 series extracted from the WOMBAT[45] database. For more than 700 series he could find a QSAR equation having a q^2 greater than 0.2 using 2-D descriptors and leave-one-out cross-validation. The WOMBAT database as well as the ChEMBL database contains curated biological activity data taken from dose−response experiments. Apart from the different structural compositions, this is the major difference to the other two databases, PubChem and ChemBank, where biological data are recorded from double- or triple-point measurements at a single concentration. For success of automated model creation, the quality of biological activity data seems crucial.

Usually, the original query molecule was a member of the training set (requiring an experimentally measured biological data point). Often the reported activity for the query structures was low enough that no off-target effect had been determined. Arguably, such a result provides an independent proof of the safety of the corresponding drugs, which of course have usually been on the market for quite some time. Furthermore, when the query structure was not part of a training set, the predicted activity was also often below the threshold, implying that such an off-target effect would not be experimentally observed. In both of these situations, the QSEA method would be applied precautionary, seeking to avoid the kinds of structural change to the query structure that would be most likely to induce an off-target effect.

However, whenever the experimental activity of the query structure was high enough to suggest a (presumably undesirable) off-target effect, the SAR table then represents the project-critical situation of a lead structure being threatened by an experimentally observed off-target effect. In this specific situation, the model derived from the SAR table can provide structural interpretations that can help to circumvent the undesired effect.

Four application examples will now be described in enough detail to indicate the potentially useful results from a QSEA application. Please note that because topomer CoMFA models are based on the topomer representation of each molecular structure, i.e., as a pair of directly bonded fragments, the contour plots that summarize topomer CoMFA models consist of several pictures (for each of the two fragments, the results for the steric and electrostatic contours are depicted separately).

Fentanyl. The topomer search in the ChEMBL database using the drug molecule fentanyl as the query yielded a SAR table with 18 topomer-similar compounds having measured experimental affinities for the (undesirable) binding to the hERG voltage-gated IKr potassium channel, with affinities ranging from 150 to 7143 nM. Fentanyl itself, an opioid analgesic, was not part of the data set, but it has a reported affinity to hERG of 5.74 μM.[46] Among the various QSEA-developed topomer CoMFA models, a model with 15 compounds provided a crossvalidated r^2 value of 0.48. The contour plots summarizing this topomer CoMFA model are shown in Figure 2. Variations in the right-hand side of the structures (ethyl-benzene) are rather limited. Therefore, this side did not contribute significantly to the overall model, and only the left-hand side of fentanyl is shown overlaid with the left-hand side of the most potent structure of the data set: CHEBI-ID: 344262. This compound had a reported affinity of 150 nM. The contour plots of the model allow a self-consistent interpretation of how structural changes at the molecules will affect their affinity to the hERG channel. Please note that the target variable is used as is, so high values represent low affinities and low values high affinity. Because high values/low affinities are desired for an off-target activity, the CoMFA contours can be interpreted in the traditional way:

1 *Para* substituents on the aromatic ring (yellow contour) are disfavored.

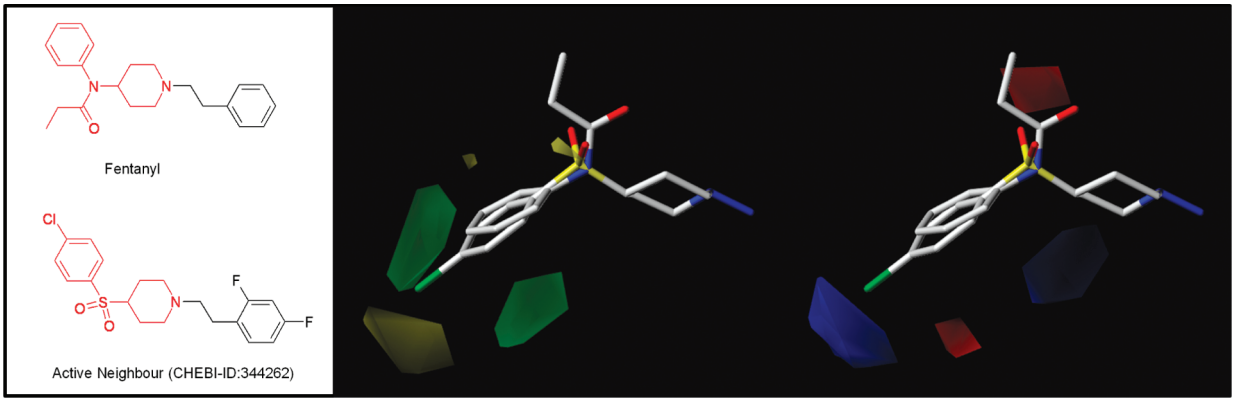


Figure 2. Topomer CoMFA model of fentanyl/hERG binding. Green and yellow contours highlight favored and disfavored steric interactions, whereas red and blue contours highlight favored negative and positive electrostatic interactions (disfavored positive and negative electrostatic interactions).

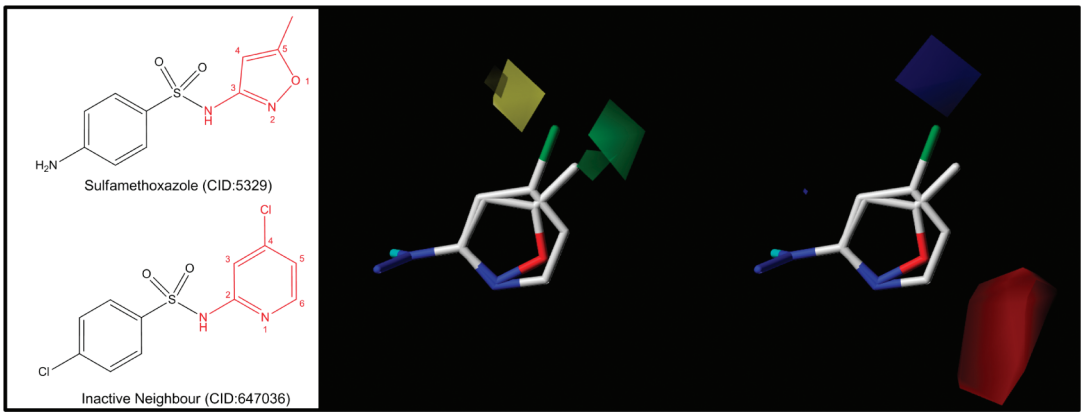


Figure 3. Topomer CoMFA model of sulfamethoxazole/estrogen receptor SAR table. The amino-part of sulfamethoxazole and an inactive neighbor compound (CID: 647036) are overlaid.

2 Because of the different linkage between the aromatic and piperidine ring, the orientation of the aromatic rings in the two structures are slightly different, and the steric bulk of the aromatic ring with the amino-linkage of fentanyl is favored.
3 Electronegative substitution in para position of the aromatic ring (blue contour) is disfavored.
4 Electronegative substitution on the carbon next to the linker atom of the aromatic and piperidine ring (red contour) is favored.

For the "most active" structure shown (CHEBI-ID: 344262), each of the positions is occupied by a physicochemically appropriate structural group. The chlorine in *para* position is filling the sterically disfavored space and is also pointing toward a blue contour where electronegative interactions are disfavored, while both sulfonyl-oxygens are directed away from the red contour where electronegative interactions are favored. The topomer CoMFA model for this table explains the low affinity predicted for the query molecule, fentanyl, in terms of positive contributions from the different linkage between the piperidine and aromatic ring, the lacking of an electronegative *para* substituent on the aromatic ring and with its carbonyl group pointing toward the electronegative contour.

Sulfamethoxazole. Topomer similarity searching of the PubChem database with sulfamethoxazole as the query yielded a SAR table containing 127 compounds with activities reported for undesirable estrogen receptor activation. Sulfamethoxazole (CID: 5329), a bacteriostatic inhibiting dihydrofolate synthetase, is part of the training set and had a reported PubChem score of 117. From QSEA, a training set size of only 16 compounds generated the topomer CoMFA model having the highest q^2 value of 0.46, with variation in an amino fragment having the greater influence on the biological activities. Figure 3 superimposes the amino fragments from sulfamethoxazole and an inactive neighbor (CID: 647036) onto the contour plots for the amino fragment. There appear to be three main contributions to the activity differences:

1 Electronegative substitution is disfavored at the *meta* position (red contour).
2 Steric bulk is favored at the *para* position (green contour).
3 Electronegative substituent is favored in other *meta* positions (red contour).

Sulfamethoxazole is unsubstituted at the 2-position, possesses a methyl substituent at position 5 filling the sterically favored region, and its electronegative oxygen within the isoxazole ring points toward the electronegative region at position 2. The structure of the inactive neighbor has a chlorine substituent in position 4 and is unsubstituted in positions 5 and 6. These structural differences both explain the observed activity of sulfamethoxazole and also indicate how the structure of sulfamethoxazole might be changed to circumvent its undesired activation of the estrogen receptor.

Acetohexamide. A third example is derived again from a search of the PubChem database using acetohexamide as the query, generating an SAR table with 39 molecules having reported cytochrome P450 2C19 inhibitory activities. The query structure itself, a hypoglycemic agent acting on the ATP-dependent K^+ channel, was not part of this SAR table. Because the structurally closest neighbor of acetohexamide (CID: 751041) is inactive, a prediction based merely on the structural similarity that acetohexamide itself would not be active would have been incorrect. Therefore, it is quite interesting that the structural interpretation of the topomer CoMFA model from this SAR table instead correctly suggests the opposite prediction.

Figure 4 shows the contour plots contributed to this model by the sulfamido fragment. The green contour above the plane of the aromatic ring at its *para* position is occupied by the acetyl substituent of acetohexamide, with its somewhat electropositive hydrogens further pointing toward the blue contours and its carbonyl oxygen close to the electronegative red contours. Its inactive neighbor instead employs a methyl substituent at the *para* position, which does not reach the sterically favored green region and positions its electropositive hydrogens closer to the electronegative red contours.

Figure 5 shows how the contour plots contributed to the model and the remaining fragments accentuate the biological difference between acetohexamide and its neighbor. Here, the orientation of the rear ends of the cyclohexyl ring (acetoheximide) and piperidyl ring (inactive neighbor), respectively, point into sterically favored (green contours) above and sterically disfavored regions (yellow contours) below the rings. Thus, the topomer CoMFA model suggests a known liability for this marketed drug, one which would have been hidden if only structural similarity were used for off-target prediction. The structural differences between acetohexamide and its inactive neighbor, though indeed small in overall magnitude, are apparently critical to their respective interactions with cytochrome 2C19.

Trifluridine. Trifluridine is an antiviral pyrimidine nucleoside interfering with DNA polymerase. Topomer similarity searching

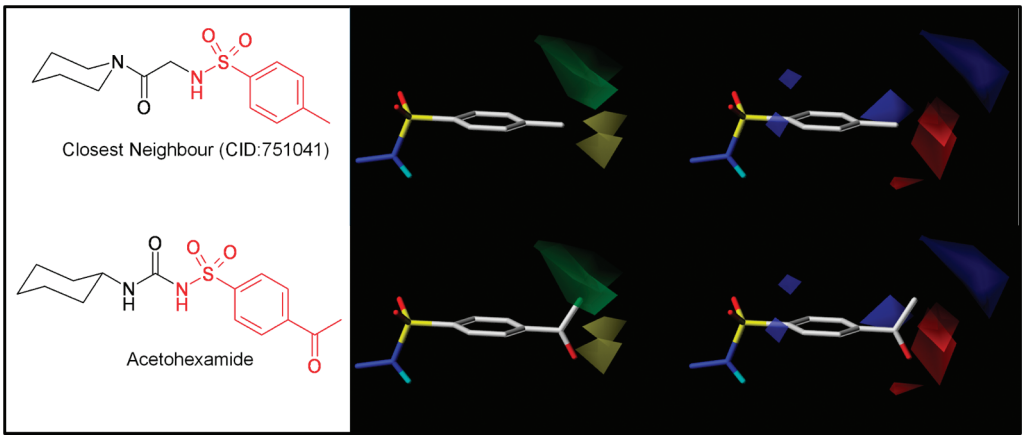


Figure 4. Topomer CoMFA model of acetohexamide/CYP450 3C19 SAR table. The sulfamido part of acetohexamide and an inactive neighbor compound (CID: 751041) are overlaid.

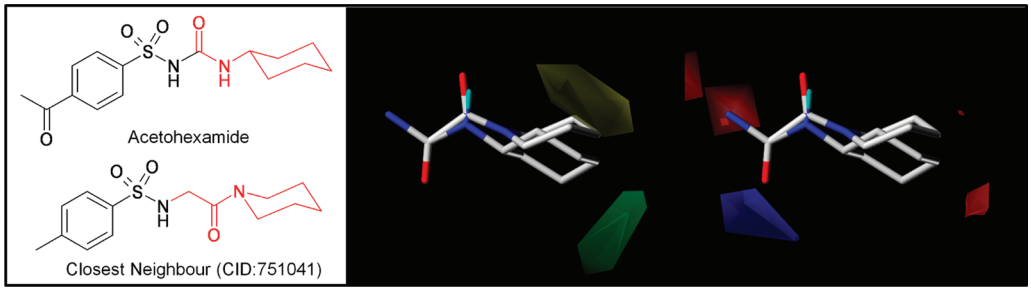


Figure 5. Topomer CoMFA model of acetohexamide/CYP450 3C19 SAR table. The amido part of acetohexamide and an inactive neighbor compound (CID: 751041) are overlaid.

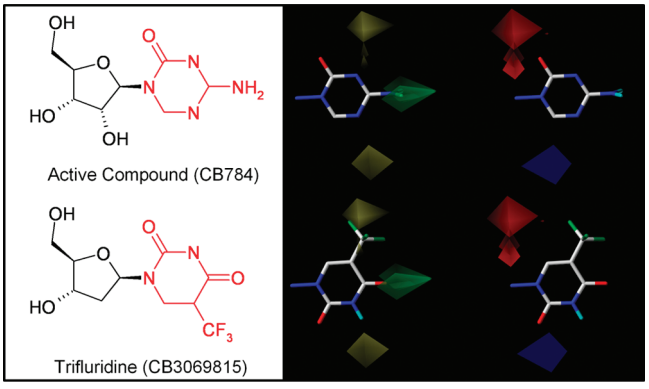


Figure 6. Topomer CoMFA model of trifluridine/estrogen acceptor activation. The pyrimidine part of trifluridine and an active neighbor compound are overlaid.

of the ChemBank database using trifluridine as the query yielded a SAR table with 10 neighbor compounds having experimental activities for inhibition of the Wnt pathway, reported as Z-scores ranging from 0.3 to 9.2. In ChemBank, an active structure is conventionally defined by a Z-score above 8.53. Therefore the reported activity of 0.32 for trifluridine (Chembank-ID: CB3069815) suggests no affect on the Wnt pathway. QSEA identifies two similar topomer CoMFA models at training set sizes of 9 and 10, with $q2$ values of 0.44 and 0.32, respectively. The contour plots of one of the topomer CoMFA models superimposed onto the least active (trifluridine) and most active compounds within the data set are shown in Figure 6.

Structural interpretation of this model suggests three major variations of the pyrimidine moiety that increase activity:

1 carbonyl in its *ortho* position
2 steric bulk in its *meta* position
3 electropositive substituent in its *para* position

Electronegative substitution in the *ortho* position raises an interesting ambiguity because both the most and least active compound have a carbonyl group in this position. However, in the most active compound (top), the pyrimidine moiety is rotated so that the carbonyl group favorably points directly into the red contoured region, while in the inactive compound (bottom), the carbonyl group is instead disfavorably directed toward the blue contoured region. This difference in topomer poses is caused by an emphasis on steric influences (actually heavy atom counts) during the structural canonicalizations that produce them. Here, the addition of the bulky CF3 group into the *meta* position flips the topomer pose of the pyridine ring by a rotation of 180 degrees. Meanwhile, the direct contributions from the accompanying change in disposition of *meta* substituents are negligible because the symmetry-related regions are both sterically disfavored. Otherwise, Wnt pathway inhibition is increased by bulky electropositive substitution in the *para* position. Thus, a favorable *para* amino group in the most active compound may be contrasted with an unfavorable *para* electronegative carbonyl group in the least active.

dx.doi.org/10.1021/ci100270x |*J. Chem. Inf. Model.* 2011, 51, 843–851

There are no major contributions to the model from the furan side chain because variations in that part of the structure do not consistently correlate with variations in activity.

The somewhat artifactual character of the topomer poses that automatically differentiated trifluridine from the most active inhibitor of the Wnt pathway requires further discussion. As a practical matter, it happens that the entire model has little consequence because the drug trifluridine does not itself exhibit the off-target effect. Nevertheless, this example highlights both the limitations and the strengths of such rule-based pose generation. On the one hand, it is generating alignments that could be viewed as ambiguous and potentially could have been avoided if the alignment were to be generated manually. On the other hand, by consistently applying the topomer rules to generate a consistent conformation, a SAR trend was automatically detected that unambiguously captured the key effects from *para* substitution that any model based on manually generated poses would also have reproduced.

CONCLUSIONS

Currently, opportunities for applying chemical biology to the drug discovery process are mostly limited to the initial biological decision of target selection. The numerous discovery decisions remaining to be made within a discovery project are predominantly chemical and structural in character. In principal, these decisions too should be facilitated by chemical biology, drawing on the public chemogenomic space that ChEMBL, PubChem, and ChemBank are beginning to populate, to supplement the much smaller SAR table-sized spaces on which an individual discovery organization or project focuses. However, existing methods for analyzing SAR tables are severely challenged by the massiveness of the data already available from ChEMBL, PubChem, and ChemBank.

Therefore, the encouraging results we have presented, from applying the new QSEA and topomer CoMFA methods to data sets from exhaustive topomer similarity searching of ChEMBL, PubChem, and ChemBank, and the high apparent relevance of these results to actual drug discovery decision making, should be of general interest to chemical biologists and medicinal chemists. Particularly noteworthy are the following:

- The entirely automatic character of performing multiple query searches in million compound databases with (necessarily) modest effort, enabling its possible implementation within an alert system
- A facility that goes beyond "this compound may exhibit this off-target effect" to "based on the available SAR data, here are the kinds of structural changes that are most likely to reduce the risk of this off-target effect"
- Among the four example applications, one that dramatizes how considering all the available SAR data can reverse a decision based only on structural similarity

ASSOCIATED CONTENT

Supporting Information. Table S1 with all 250 query structures in SMILES format; Table S2 giving numerical results of the virtual screening for each query structure and database screened (ChEMBL, PubChem, ChemBank, BindingDB, Binding-MOAD, K_i-Database); structure−activity tables for fentanyl/hERG binding, sulfamethazole/estrogen receptor activation, acetohexamide/CYP450 2C19 inhibition, and trifluridine/Wnt-pathway inhibition. This information is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bwendt@embl.de.

Present Addresses

[§]Elara Pharmaceuticals GmbH, Boxbergring 107, Heidelberg, D-69126 Germany. Phone: +49 6221 387 8172. Fax: +49 6221 387 8850. E-mail: b.wendt@elarapharma.com.

ACKNOWLEDGMENT

This study was supported by the BMBF (Grant 0315418). For carefully reading the manuscript and for providing helpful suggestions, we are grateful to Richard Cramer as well as to Brian Masek, both of Tripos. The reviewers are thanked for their constructive suggestions.

REFERENCES

(1) Sams-Dodd, F. Target-based drug discovery: Is something wrong? *Drug Discovery Today* **2005**, *10*, 139–137.

(2) Cheung, C. H.; Coumar, M. S.; Hsieh, H. P.; Chang, J. Y. Aurora kinase inhibitors in preclinical and clinical testing. *Expert Opin. Invest. Drugs* **2009**, *18*, 379–398.

(3) 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.

(4) Hopkins, A. L. Network pharmacology: The next paradigm in drug discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690.

(5) Hart, C. P. Finding the target after screening the phenotype. *Drug Discovery Today* **2005**, *10*, 513–519.

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

(7) Huang, S. M. A.; Mishina, Y. M.; Liu, S.; Cheung, A.; Stegmeier, F.; Michaud, G. A.; Charlat, O.; Wiellette, E.; Zhang, Y.; Wiessner, S.; Hild, M.; Shi, X.; Wilson, C.; J., Mickanin, C.; Myer, V.; Fazal, A.; Tomlinson, R.; Serluca, F.; Shao, W.; Cheng, H.; Shultz, M.; Rau, C.; Schirle, M.; Schlegl, J.; Ghidelli, S.; Fawell, S.; Lu, C.; Curtis, D.; Kirschner, M. W.; Lengauer, C.; Finan, P. M.; Tallarico, J. A.; Bouwmeester, T.; Porter, J. A.; Bauer, A.; Cong, F. Tankyrase inhibition stabilizes axin and antagonizes Wnt signalling. *Nature* **2009**, *461*, 614–20.

(8) Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **2006**, *16*, 127–136.

(9) Cleves, A. E.; Jain, A. N. Robust ligand-based modeling of the biological targets of known drugs. *J. Med. Chem.* **2006**, *49*, 2921–2938.

(10) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A,I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175–181.

(11) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2007**, *2*, 861–873.

(12) Kuhn, M.; Campillos, M.; González, P.; Jensen, L. J.; Bork, P. Large-scale prediction of drug-target relationships. *FEBS Lett.* **2008**, *582*, 1283–1290.

(13) Campillo, M.; Kuhn, M.; Gavin, A. C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263–266.

(14) Jenkins, J.; Bender, A.; Davies, J. W. In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today: Technol.* **2006**, *3*, 413–421.
(15) Garcia-Serna, R.; Mestres, J. Anticipating drug side effects by comparative pharmacology. *Expert Opin. Drug Metab. Toxicol.* **2010**, *6*, 1253–1263.
(16) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: "Target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49*, 6802–6810.
(17) Peterson, R. T. Chemical biology and the limits of reductionism. *Nat. Chem. Biol.* **2008**, *4*, 635–638.
(18) Senger, S.; Leach, A. SAR Knowledge Bases in Drug Discovery. *Ann. Rep. Comp. Chem.* **2008**, *4*, 203–216.
(19) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated platform of small molecules and biological activities. *Ann. Rep. Comp. Chem.* **2008**, *4*, 217–241.
(20) Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S.; Brudz, S.; Sullivan, J. P.; Muhlich, J.; Serrano, M.; Ferraiolo, P.; Tolliday, N. J.; Schreiber, S. L.; Clemons, P. A. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucl. Acid. Res.* **2008**, *36*, 351–359.
(21) Bender, A. Compound bioactivities go public. *Nat. Chem. Biol.* **2010**, *6*, 309.
(22) Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R. N.; Gilson, M. K. BindingDB: A Web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, 198–201.
(23) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins* **2005**, *60*, 333–340.
(24) Data was generously provided by the National Institute of Mental Health's Psychoactive Drug Screening Program, Contract HHSN-271-2008-00025-C (NIMH PDSP).
(25) Oprea, T.; Tropsha, A.; Faulon, J.-L.; Rintoul, M. D. Systems chemical biology. *Nat. Chem. Biol.* **2007**, *3*, 447–450.
(26) Wendt, B.; Cramer, R. D. Quantitative Series Enrichment Analysis (QSEA): A novel procedure for 3D-QSAR analysis. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 541–551.
(27) Cramer, R. D.; Jilek, R. J.; Andrews, K. M. dbtop: Topomer similarity searching of conventional structure databases. *J. Mol. Graphics Modell.* **2002**, *20*, 447–462.
(28) Cramer, R. D.; Topomer CoMFA, A. Design methodology for rapid lead optimization. *J. Med. Chem.* **2003**, *46*, 374–388.
(29) SYBYL, Tripos International, 1699 South Hanley Rd., St. Louis, Missouri 63144, USA.
(30) csp_standard.defs contains a set of SLN-rules (SYBYL Line Notation) to convert ionized structures into their neutral forms.
(31) Jilek, R. J.; Cramer, R. D. Topomers: A validated protocol for their self-consistent generation. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1221–1227.
(32) Pearlman, R. S. Concord; Tripos International, St. Louis, MO.
(33) Hert, J.; Willett, P.; Wilton, D. J. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(34) Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219–1229.
(35) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.
(36) Boecker, A. Toward an improved clustering of large data sets using maximum common substructures and tolological fingerprints. *J. Chem. Inf. Model.* **2008**, *48*, 2097–2107.
(37) Xie, X. Q.; Chen, J. Z. Data mining a small molecule drug screening representative subset from NIH PubChem. *J. Chem. Inf. Model.* **2008**, *48*, 465–475.
(38) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.
(39) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 394–401.
(40) Kogej, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multi-fingerprint based similarity searches for targeted class compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 1201–1213.
(41) Freitas, R. F.; Bauab, R. L.; Montanari, C. A. Novel application of 2D and 3D-similarity searches to identify substrates among cytochrome P450 2C9, 2D6, and 3A4. *J. Chem. Inf. Model.* **2010**, *50*, 97–109.
(42) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: A perspective. *J. Med. Chem.* **2010**, *53*, 3862–3886.
(43) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.
(44) Olah, M.; Bologa, C.; Oprea, T. I. An automated PLS search for biologically relevant QSAR descriptors. *J. Comput.- Aided Mol. Des.* **2004**, *18*, 437–449.
(45) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea., T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*; Oprea, T.I., Ed.; Wiley-VCH: New York, 2004, pp 223−239.
(46) Tobita, M.; Nishikawaa, T.; Nagashimaa, R. A discriminant model constructed by the support vector machine method for HERG potassium channel inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2886–2890.