Chemical Transformations That Yield Compounds with Distinct Activity Profiles

Ye Hu and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

S *Supporting Information*

ABSTRACT: We have systematically searched for chemical changes that generate compounds with distinct biological activity profiles. For this purpose, activity profiles were generated for ~42000 compounds active against human targets. Unique activity profiles involving multiple target proteins were determined, and all possible matched molecular pairs (MMPs) were identified for compounds representing these profiles. An MMP is defined as a pair of compounds that are distinguished from each other only at a single site such as an R group or ring system. For example, in an MMP, a hydroxyl group might be replaced by a halogen atom or a benzene ring by an amide group. From ~37500 MMPs, more than 300 nonredundant chemical transformations were isolated that yielded compounds with distinct activity profiles. None of these transformations was found in pairs of compounds with overlapping activity profiles. These transformations were ranked according to the number of MMPs, the number of activity profiles, and the total number of targets that they covered. In many instances, prioritized transformations involved ring systems of varying complexity. All transformations that were found to switch activity profiles are provided to enable further analysis and aid in compound design efforts.

KEYWORDS: Active compounds, target annotations, activity profiles, profile analysis, matched molecular pairs, chemical transformations
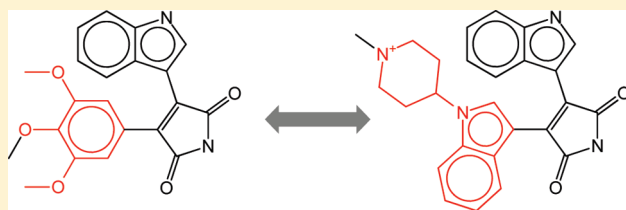
Exploring structural determinants of specific biological activities of small molecules is generally of high interest in medicinal chemistry. Such investigations can be carried out at different levels, for example, by analyzing chemical neighborhood behavior,[1] studying compound series following the classical quantitative structure−activity relationship (QSAR) paradigm,[2] or exploring different types of activity landscape models[3] including conventional single-target[3,4] and selectivity landscapes[4] or multitarget activity landscape representations.[5] Statistical surveys of substituents that affect compound potency have also been reported.[6,7] Typically, such studies require the application of a canonical definition of molecular frameworks and substituents, for which several alternatives exist. Another way to generalize chemical modifications in a consistent manner is the utilization of the matched molecules pair (MMP) formalism.[8] An MMP is defined as a pair of compounds that are distinguished from each other only at a single site (such as an R group or ring system) or, in other words, that are related by a specific chemical "transformation", that is, the exchange of one group with another. In the context of MMP analysis, the term transformation is utilized to generalize chemical changes but not to refer to reaction information. Hence, chemical changes in MMPs are algorithmically defined and generalized, as further explained below, but they are not as the result of specific chemical reactions.

The MMP concept has recently been applied to a number of medicinal chemistry or drug discovery relevant questions. For example, MMPs have been systematically generated and analyzed for bioactive compounds to identify substitutions that form activity cliffs across different compound classes.[9] Furthermore, MMPs have been utilized to compare compounds with primary target and antitarget annotations to predict chemical changes that might affect antitarget activity.[10] In addition, the way in which physicochemical parameters of compounds change as a consequence of MMP transformations has been investigated.[10] To support such data mining and prediction efforts, an efficient algorithm has been introduced to generate MMPs on a large scale,[11] as discussed in the Experimental Procedures.

The major goal of our study has been to analyze whether chemical transformations exist that produce compounds with distinct (nonoverlapping) activity profiles. Therefore, on the basis of currently available public domain data, we have first generated activity profiles for all qualifying compounds and then, utilizing the MMP formalism, systematically searched for chemical transformations that met our activity profile criteria. Methodological details are provided in the Experimental Procedures.

Our approach is outlined in Figure 1. For preselected compounds (see the Experimental Procedures), activity profiles were generated by assembling all available target annotations. Then, all unique activity profiles were determined, and compounds displaying these activity profiles were collected. In the next step, all possible profile pairs were assembled. Pairs formed between single targets were removed, and the remaining profile pairs were

Figure 1

Activity Profiles

Compound	active against	Targets
1		A, B
2		A, C
3		D, E, F
4		A, B
5		A, C

Unique Profiles

Compounds	Profile
1,4	AB
2,5	AC
3	DEF

Profile Pairs

	overlapping	
AB		AC
	distinct	
AB		DEF
AC		DEF

Corresponding Compound Pairs

1-2; 1-5; 4-2; 4-5	AB —— AC	
1-3; 4-3	AB —— DEF	
2-3; 5-3	AC —— DEF	

Figure 1. Methodological summary. An outline of the approach to identify activity profile-switching chemical transformations is presented. In the first step, an activity profile was generated for each compound from reported target annotations. In the second step, all unique activity profiles were determined, and compounds were organized according to their activity profiles. In the third step, all possible profile pairs were generated and classified as "overlapping" profile pairs (i.e., consisting of activity profiles that had at least one target in common) or "distinct" profile pairs (i.e., consisting of profiles that did not have any target in common). In the fourth step, all compound pairs "corresponding" to distinct or overlapping profile pairs were identified (i.e., for each profile pair, all pairs of compounds were systematically generated that formed this profile pair). For example, compounds **1** and **4** display activity profile AB, and compound **3** displays profile DEF. Thus, two compound pairs can be generated in this case, 1-3 and 4-3, which "correspond" to the distinct profile pair AB-DEF. In the fifth step, MMPs were systematically generated from all compound pairs, and transformations were determined for each activity profile pair.

classified as pairs consisting of distinct or overlapping profiles. Then, all compound pairs representing distinct or overlapping profile pairs were identified. From these compound pairs, MMPs were systematically generated, and transformations were determined.

The results of this analysis are summarized in Figure 2. We identified 41801 compounds that were active against a total of 754 different target proteins with at least 10 μM potency. These compounds had 1778 unique activity profiles. From these profiles, we then enumerated all possible profiles pairs, yielding ∼1.5 million pairs. More than 1.4 million of these consisted of distinct activity profiles, whereas only ∼37000 pairs contained overlapping profiles, that is, activity profiles that had at least one target in common. Then, all possible compound pairs were assigned to distinct and overlapping profile pairs. From these compound pairs, all possible MMPs were generated, and from these MMPs, chemical transformations were systematically extracted. Importantly, a single MMP might yield several overlapping transformations. A total of 37527 MMPs with distinct activity profiles were obtained yielding 80592 transformations. A much larger number of 82635 MMPs with overlapping activity profiles was obtained, yielding 510046 transformations. It should be noted that the same transformation might in principle be
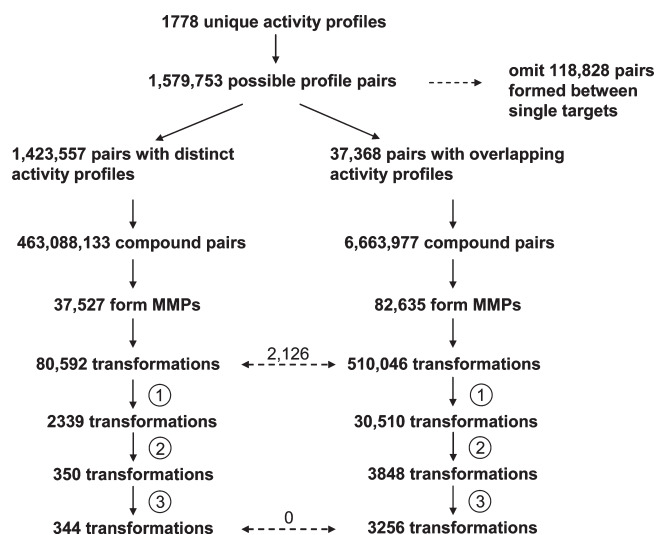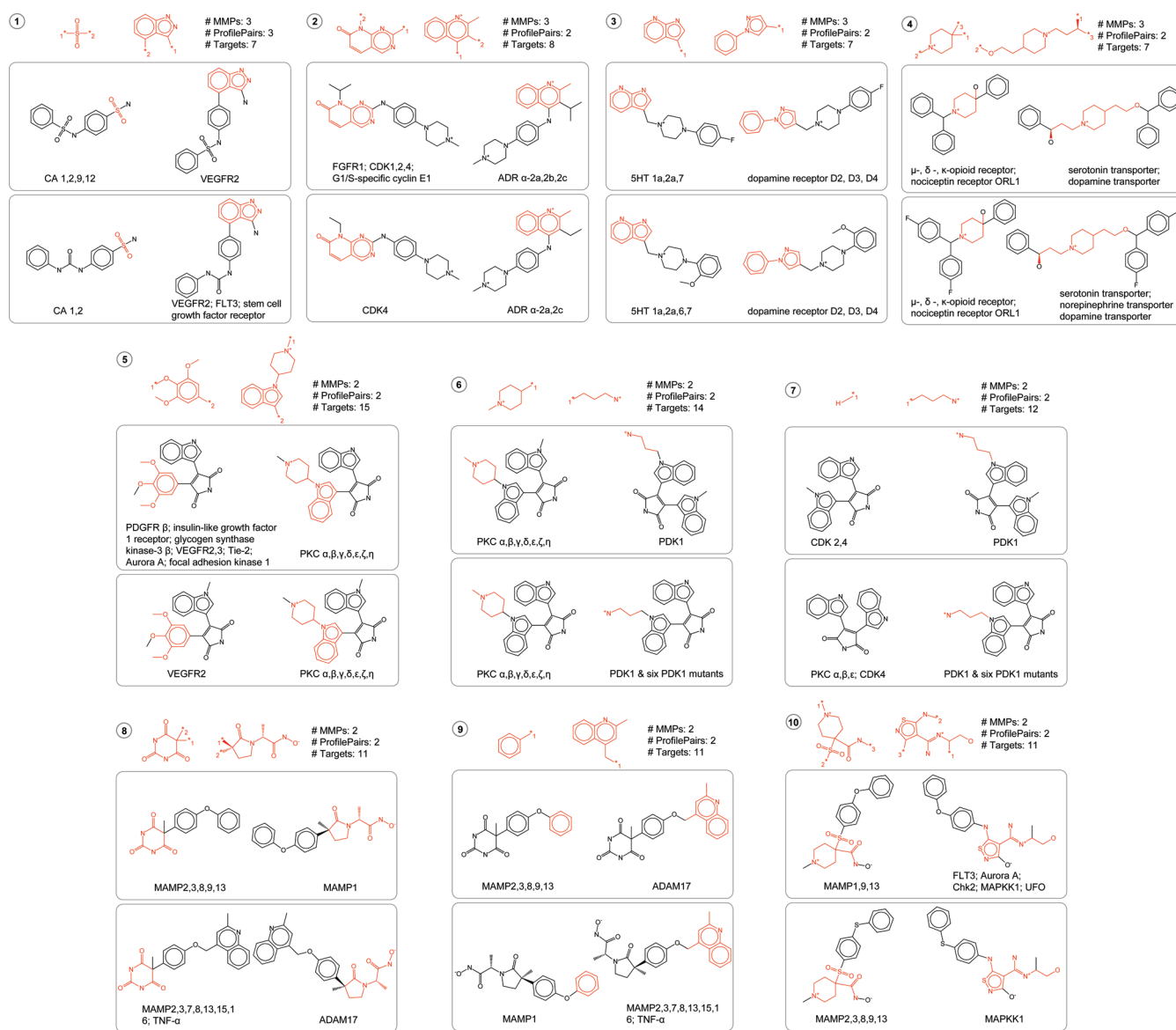
Figure 2

Figure 2. Statistical analysis. Shown is a flowchart summarizing the profile, compound, MMP, and transformation statistics at different stages of our analysis. Three rules (1−3) are applied to filter transformations, as described in the text.

encoded by MMP(s) with distinct or overlapping activity profiles, although this might not be very likely. Therefore, we also determined the overlap between transformations defined by MMPs with distinct and overlapping activity profiles. However, the overlap between these sets of transformations was very small, with only 2126 shared transformations. These transformations were omitted from further consideration.

The remaining transformations derived from MMPs with distinct activity profiles were filtered using a three-step procedure. First, only transformations represented by multiple MMPs were selected. Second, when different transformations were obtained from identical MMP sets, only the one containing the smallest number of heavy atoms was retained. Third, transformations were omitted whose MMP set was a subset of another transformation. Criterion one was applied to focus the analysis on frequent transformations, whereas criteria two and three were applied to remove redundant transformation information. After filtering, a total of 344 transformations remained that interconverted compounds with distinct activity profiles. As a control, we also repeated the analysis with transformations obtained from MMPs with overlapping activity profiles. A total of 3256 control transformations were obtained. There was no overlap between this set and our 344 transformations derived from MMPs with distinct activity profiles.

These 344 transformations were then ranked according to the number of MMPs from which they were obtained, the number of activity profiles that they converted, and the total number of targets involved in these activity profiles. From these independent rankings, a rank fusion was calculated to further prioritize transformations according to the smallest sum of ranks. This standard rank fusion approach adds individual ranking positions from different lists and prioritizes compounds that are overall most highly ranked; that is, a consensus ranking is obtained by calculating the sum of ranks and ordering compounds according to increasing rank sum.

Figure 3 shows the top 10 transformations yielding compounds with distinct activity profiles. The final ranking was obtained on the basis of the rank fusion procedure described

Figure 3. Prioritized transformations. Shown are the 10 top-ranked chemical transformations (rank 1−10) that yield compounds with distinct activity profiles. These final rankings were obtained after rank fusion, as described in the text. In each case, the exchanged fragments (red) are shown together with two MMPs that provide the structural context of the transformation. For each transformation, the number of MMPs, activity profile pairs, and total number of targets involved in these profile pairs are reported. In addition, activity profile targets are defined for each MMP. Target abbreviations: CA, carbonic anhydrase; VEGFR, vascular endothelial growth factor receptor protein-tyrosine kinase; FLT3, FMS-like tyrosine kinase 3; FGFR, fibroblast growth factor receptor protein-tyrosine kinase; CDK, cyclin-dependent kinase; ADR, adenosine receptor; 5HT, serotonin receptor; PDGFR, platelet-derived growth factor receptor protein-tyrosine kinase; PKC, protein kinase C; PDK, 3-phosphoinositide-dependent protein kinase; MAMP, matrix metalloproteinase; TNF-α, tumor necrosis factor α; ADAM17, disintegrin and metalloproteinase domain-containing protein 17; MAPKK1, dual specificity mitogen-activated protein kinase kinase 1; Chk2, checkpoint kinase 2; and UFO, tyrosine-protein kinase receptor with unidentified protein function.

above. All 344 transformations are provided in Table S1 in the Supporting Information. All transformations in Figure 3 are represented by two or three MMPs and two or three activity profile pairs. The activity profiles contain between seven and 15 different targets. In some cases, the targets constituting an activity profile pair are closely related; in others, they belong to different protein families. The majority of targets are different G protein receptors (GPCRs), tyrosine or serine/threonine kinases, and proteases. The top 10 transformations in Figure 3 mirror the overall distribution of transformations in Table S1. Seven of 10 prioritized transformations (rank 2−5 and 8−10 in

Figure 3) involve the exchange of substituted aromatic and/or aliphatic ring systems of varying complexity. For all transformations provided in Table S1 in the Supporting Information, such ring-to-ring transformations are also most frequently observed. The three remaining transformations in Figure 3 involve the exchange of a functional group or aliphatic substituent with a ring (rank 1 and 6) or the introduction of an aliphatic substituent (rank 7). These types of transformations are also much less frequently observed in Table S1 in the Supporting Information. Thus, transformations that yield compounds with distinct activity profiles rarely involve conventional R group replacements.

525

dx.doi.org/10.1021/ml2000609 |*ACS Med. Chem. Lett.* 2011, 2, 523–527

It follows that these results could not have been obtained by conventional R group decomposition or scaffold/R group analysis. These findings have implications for compound design, as further discussed in the following. The prevalence of ring-based transformations suggests that defined changes in core structures, which might often represent characteristic binding motifs for given targets, lead to distinct activity profiles, rather than differences in substitution patterns. For example, this is apparent for transformations that change compound activity from tyrosine kinase inhibitors to adenosine receptor ligands (rank 2), from serotonin to dopamine receptor ligands (rank 3), or from receptor tyrosine kinase to serine/threonine kinase inhibitors (rank 5). Thus, partial replacements of core ring structures in active compounds would represent a preferred compound design strategy to facilitate activity profile switches.

In summary, through systematic compound data mining, we have demonstrated that chemical transformations exist that generate compounds with distinct biological activity profiles, which comprise all target annotations of a compound with potency values higher than a threshold value. We have identified more than 300 MMP transformations that convert compounds with distinct profiles consisting of multiple targets into each other. Our analysis was designed to be compound- rather than compound class-centric. Accordingly, the MMP concept was applied to represent transformations because it generalizes them in a consistent manner and does not require the organization of molecules into scaffolds and R groups or the inclusion of synthetic criteria. On this basis, we have been able to determine that partial replacements of core ring structures most frequently yield compounds with distinct activity profiles, a finding that was not anticipated. Despite general compound data sparseness, more than 300 transformations have been identified that were present in more than one MMP and triggered switches between multiple distinct activity profiles. As more activity measurements become available, additional transformations will likely be identified, and some of the transformations reported herein will probably be observed at higher frequencies, thus providing a further improved basis for the use of activity profile-switching chemical transformations in compound design. Clearly, not all transformations might be equally amenable to compound design because they are derived on the basis of the MMP concept that generalizes chemical changes without including synthetic or pharmacophore information. The top-ranked transformations in Figure 3 mostly describe exchanges of ring systems that generate compounds with distinct activity profiles. Thus, ring systems that are indicated by MMP analysis to be a signature of an activity profile involving members of a particular target family might be incorporated in compounds directed at these targets to achieve target selectivity. Prime examples for such design efforts would be the ring systems involving the third and fourth ranked transformations in Figure 3 that switch compound activity between different GPCRs or GPCRs and transporters.

EXPERIMENTAL PROCEDURES

From BindingDB[12] and ChEMBL,[13] compounds active against multiple human targets with at least 10 μM potency were extracted. Compounds were represented as 2D molecular graphs. Only compounds with unique 2D graphs were included in the analysis, and in relevant cases, tautomeric states were utilized as indicated in the original compound records. Only K_i and IC_{50} values were considered as potency measurements. Because it was not necessary to compare different types

of potency values for the generation of activity profiles, the consideration of both K_i and IC_{50} was appropriate in this case. For compounds with multiple potency measurements against the same target, the geometric mean was calculated as the final potency value. For the generation of activity profiles, the major criterion has been to set a lower threshold for target potency (i.e., 10 μM) such that weakly potent compounds were excluded from activity profile generation. Most of the BindingDB and ChEMBL compounds subjected to our analysis are active in the nanomolar range. The potency distribution is reported in Figure S1 of the Supporting Information. For the activity profiles that we have generated here, absolute potency values were not relevant, because they were built on the basis of target annotations. For all qualifying compounds, activity profiles were generated and compared. An activity profile consisted of all target annotations reported for a compound. MMPs were generated using an adaptation[9] of the Hussain and Rea algorithm.[11] Our analysis was carried out with in-house generated Perl, Java, and Pipeline Pilot[14] programs.

To generate MMPs using our implementation of the Hussain and Rea algorithm, all selected compounds are initially fragmented as follows: All nonring single bonds between two nonhydrogen atoms in a compound are marked, which is followed by systematic deletion of these bonds, producing so-called single cuts, and their two- and three-bond combinations, that is, double cuts and triple cuts, respectively. A single cut results in two fragments F1 and F2 that are added to an index list. Then, two "key-value" pairs are built as follows: Fragment F1 is added as a "key" to the index using F2 as the corresponding "value", and the same is done for F2 as the key, and F1 is the value. Double cuts produce a core and two terminal fragments. Here, the core is used as the value and the combination of the two terminal fragments as the key. Furthermore, only those triple cuts are retained that result in a single core and three terminal fragments. Here, the core is also stored as the value, and the three terminal fragments together are stored as the key. Connectivity information of all fragments is retained. Stereochemical criteria are considered. For all generated key−value pairs, their origin is also stored, that is, the source compound information. The pairwise combination of all compounds sharing a particular key (and hence the corresponding fragment) then yields MMPs, and the two value fragments define the structural transformation for each of these MMPs. We consider a combination of two compounds only as an MMP if the heavy (nonhydrogen) atom counts of their distinguishing fragments differs by maximal eight atoms such that compounds forming MMPs do not significantly differ in size. Importantly, by utilizing single and multiple cuts, transformations involving both R groups and core structures are obtained, thus generalizing the transformation scheme. Hence, the algorithm is not limited to R group exchanges and might yield multiple, differently sized fragments that define transformations.

■ ASSOCIATED CONTENT

S **Supporting Information.** All chemical transformations that generate compounds with distinct activity profiles (Table S1) and potency distribution of active compounds subjected to MMP analysis (Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
*Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

■ ACKNOWLEDGMENT

■ ABBREVIATIONS

ADR, adenosine receptor; CA, carbonic anhydrase; CDK, cyclin-dependent kinase; CHK2, checkpoint kinase 2; FGFR, fibroblast growth factor receptor; FMS, Friend murine leukemia virus integration site; FLT, FMS-like tyrosine kinase; MAMP, matrix metalloproteinase; MAPKK1, mitogen-activated protein kinase kinase 1; MMP, matched molecules pair; QSAR, quantitative structure—activity relationship; PDGFR, platelet-derived growth factor receptor; PDK, 3-phosphoinositide-dependent protein kinase; PKC, protein kinase C; TNF, tumor necrosis factor; UFO, unidentified protein function; VEGFR, vascular endothelial growth factor receptor

■ REFERENCES

(1) 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.

(2) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for Applying the Quantitative Structure-Activity Relationship Paradigm. *Methods Mol. Biol.* **2004**, *275*, 131–214.

(3) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure-Activity Landscapes. *Drug Discovery Today* **2009**, *14*, 698–705.

(4) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209–8223.

(5) Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Design of Multi-target Activity Landscapes that Capture Hierarchical Activity Cliff Distributions. *J. Chem. Inf. Model.* **2011**, *51*, 258–266.

(6) Raymond, J. W.; Watson, I. A.; Mahoui, A. Rationalizing Lead Optimization by Associating Quantitative Relevance with Molecular Structure Modification. *J. Chem. Inf. Model.* **2009**, *49*, 1952–1962.

(7) Hajduk, P. J.; Sauer, D. R. Statistical Analysis of the Effects of Common Chemical Substituents on Ligand Potency. *J. Med. Chem.* **2008**, *51*, 553–564.

(8) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271—285.

(9) Wassermann, A. M.; Bajorath, J. Chemical Substitutions that Introduce Activity Cliffs across Different Compound Classes and Biological Targets. *J. Chem. Inf. Model.* **2010**, *50*, 1248–1256.

(10) Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirka-manathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, A. W.; Macdonald, S. J. Lead Optimization Using Matched Molecular Pairs: Inclusion of Contextual Information for Enhanced Prediction of HERG Inhibition, Solubility, and Lipophilicity. *J. Chem. Inf. Model.* **2010**, *50*, 1872–1886.

(11) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348.

(12) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.

(13) European Bioinformatics Institute (EBI). *ChEMBL*; Cambridge, 2011; http://www.ebi.ac.uk/chembl/ (accessed January 20, 2011).

(14) *Pipeline Pilot*, Student ed.; Accelrys, Inc.: San Diego, CA, 2009.