Journal of
Chemical Information
and Modeling

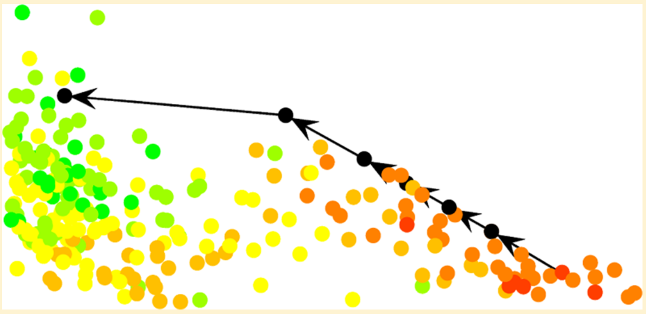
Article

pubs.acs.org/jcim

Compound Optimization through Data Set-Dependent Chemical Transformations

Antonio de la Vega de León 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



ABSTRACT: We have searched for chemical transformations that improve drug development-relevant properties within a given class of active compounds, regardless of the compounds they are applied to. For different compound data sets, varying numbers of frequently occurring data set-dependent transformations were identified that consistently induced favorable changes of selected molecular properties. Sequences of compound pairs representing such transformations were determined that formed pathways leading from unfavorable to favorable regions of property space. Data set-dependent transformations were then applied to predict a series of compounds with increasingly favorable property values. By database searching the desired biological activity was detected for several designed molecules or compounds that were very similar to these molecules. Taken together our findings indicate that data set-dependent transformations can be applied to predict compounds that map to favorable regions of molecular property space and retain their biological activity.

INTRODUCTION

Chemical modifications occurring in pharmaceutically relevant compounds can be systematically studied by molecule pair analysis.[1,2] For example, a matched molecular pair (MMP) is defined as a pair of compounds that are only distinguished by a structural change at a single site,[2] i.e., the exchange of a substructure between these compounds, which is often referred to as a chemical transformation.[3] The MMP concept is useful for many applications in medicinal chemistry.[4,5] For example, on the basis of MMP analysis, bioisosteric replacements have been identified across different compound classes[6] and also chemical changes leading to the formation of activity cliffs.[7,8] The identification of bioisosteres or activity cliff-forming transformations requires the study of potency changes that are associated with chemical transformations. In addition, the effect of transformations on other compound properties can be also assessed, which has become a popular topic in ADMET analysis.[9−12] In this context, the consequences of defined structural changes on physicochemical properties such as solubility or more complex compound characteristics such as metabolic stability or oral availability are investigated.

We have searched for transformations and transformation sequences to optimize compounds in drug development-relevant property space. A key question of our study has been whether structural modifications can be derived from data sets of known active compounds that induce favorable changes in property space and can be utilized to optimize compounds sharing the same activity. Therefore, we set out to apply the MMP concept and identify transformations that consistently improve molecular properties of known active compounds. We then attempted to use such transformations to delineate compound pathways from undesired to desired regions of property space and design new compounds.

Data set-dependent transformations (in the following referred to as set-dependent transformations) were identified in different compound sets that led to favorable changes of selected molecules properties in varying structural contexts and enable compound design. We then searched for newly designed compounds in a public domain data and identified a number of identical or very similar compounds sharing the same activity.

MATERIALS AND METHODS

Data Sets. Four sets of G protein coupled receptor (GPCR) antagonists active against the adenosine A2a (A2AR), cannabinoid CB2 (CB2), dopamine D2 (D2R), or μ-opioid receptor (MOR) were collected from ChEMBL (release 14).[13] Only compounds with high-confidence activity annotations and available K_i values were selected. If multiple K_i values were available, their geometric mean was calculated as the final compound potency. If K_i values for a compound differed by more than 1 order of magnitude, it was omitted from further consideration. The data sets contained between ∼1400 and ∼2100 compounds, as summarized in Table 1.

Descriptors and Value Ranges. For all test compounds, four descriptors were calculated using the CDK Toolkit[14] in KNIME.[15] These descriptors included molecular weight (MW), topological polar surface area (TPSA), the number of rotatable bonds (rotN), and the water/octanol partition coefficient (logP).

Received: March 19, 2013

The ADME-related property descriptor classification scheme introduced by Lobell et al.[16] was applied to distinguish between favorable (green), intermediate (yellow), and unfavorable (red) compound property descriptor value ranges. For property space analysis, the following value range combinations were defined:[16] favorable: LogP ≤ 3, MW ≤ 400, TPSA ≤ 120, rotN ≤ 7; intermediate: LogP 3–5, MW 400–500, TPSA 120–140, rotN 8–10; unfavorable: LogP > 5, MW > 500, TPSA > 140, rotN > 10.

Table 1. Compounds, MMPs, and Transformations[a]

data set	A2AR	CB2	D2R	MOR
compounds	2154	1393	1442	1415
MMPs	13791	8123	7757	7952
transformations	15640	10344	11102	9988
frequent	240	114	76	116
preferred	47	31	18	30

[a] For each data set, the total number of compounds, MMPs, corresponding transformations, and the number of frequent and preferred transformations are reported.

Chemical Transformations. Transformation size-restricted MMPs were calculated as described previously[8] using a variant of the algorithm by Hussain and Rea.[3] The size and size difference between fragments exchanged between compounds forming an MMP were limited to maximally 13 and 8 non-hydrogen atoms, respectively, to focus transformations on chemically meaningful replacements.[8] All size-restricted MMPs representing the same chemical transformation were identified. Transformations were classified as *frequent transformations* if they occurred in at least 10 different MMPs. If several possible transformations existed for a given MMP, the smallest transformation was selected. In contrast to previous MMP applications, in our current analysis each MMP [A,B] defined two direction-dependent transformations, i.e., A→B and B→A. This was done because transformations were

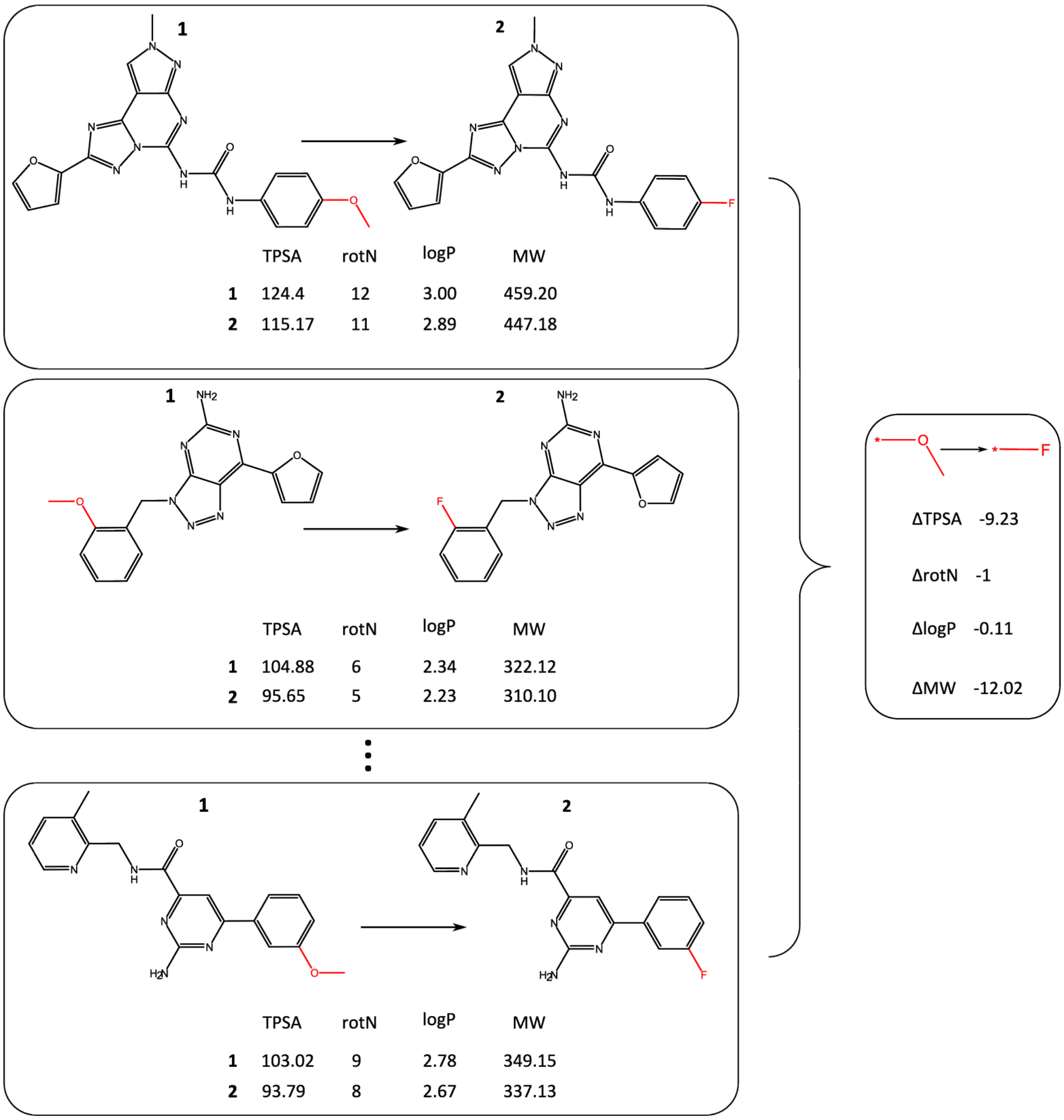


Figure 1. Transformation evaluation. To assess property changes as a consequence of a frequently occurring transformation, MMPs representing the transformation are analyzed. The descriptor value differences between compounds **2** and **1** forming each MMP are calculated and averaged over the MMPs.

associated with specific changes in descriptor values for each compound, which might be favorable in one direction and unfavorable in the other. Due to the consideration of direction-dependent transformations, the total number of unique transformations exceeded the number of MMPs, as reported in Table 1. Depending on the compound data set, between ~7,500 and ~14,000 MMPs were obtained that yielded ~10,000 to ~15,500 unique transformations.

Set-Dependent and Preferred Transformations. All frequent transformations identified for a compound set were classified *as data set-dependent transformations*. For each accepted transformation, the difference in descriptor values between compounds forming each MMP representing this transformation in the compound set was determined, and the values were averaged over all MMPs, as schematically illustrated in Figure 1. Transformations were classified as *preferred* (with respect to a given data set) if they consistently moved the values of all descriptors in a favorable direction (i.e., from red to green) or if values of one or more descriptors changed in a favorable way, while values of the others remained constant. *Preferred transformations* were not permitted to change any descriptor value in an unfavorable manner.

Transformation-Dependent Descriptor Value Changes. Descriptor value changes induced by preferred transformations were systematically assessed and predicted. For each qualifying transformation, the corresponding MMP set was 10 times randomly divided into half. For 50% of the MMPs (training set), the transformation-dependent descriptor value changes were calculated and used to predict descriptor values for the test set (i.e., the remaining 50% of the MMPs). For the latter, the actual values were then determined, and the coefficient of determination R^2 for the predicted and observed values was calculated for each of the four descriptors for the 10 independent predictions.

Visualization. For the display of compound sets and pathways, descriptor values of compounds were subjected to scaled principal component analysis (PCA) using R.[17] For each compound, the values for the first and second principal component were calculated as the x- and the y-coordinates, respectively, to obtain a 2D projection. The two first principal components accounted for 81% (CB2) to 94% (D2R) of the overall variance of the descriptor values. Compounds were represented as dots and color-coded using a continuous spectrum from green (all descriptor values were favorable) over yellow (partly unfavorable values) to red (all descriptor values were unfavorable). Pathways were delineated by connecting compounds forming MMPs with directed edges.

Compound Pathways. Within each data set, MMP sequence pathways between compounds in unfavorable and favorable regions of descriptor space were identified. Therefore, as a pathway start and end point, a compound in the unfavorable and favorable region was selected, respectively, and the shortest path between these compounds was determined. A pathway consisted of compound pairs forming overlapping MMPs, e.g., path A–B–C was formed by MMPs [A,B] and [B,C]. Hence, these pathways were defined by a series of chemical transformations that generated compounds with increasingly favorable descriptor values.

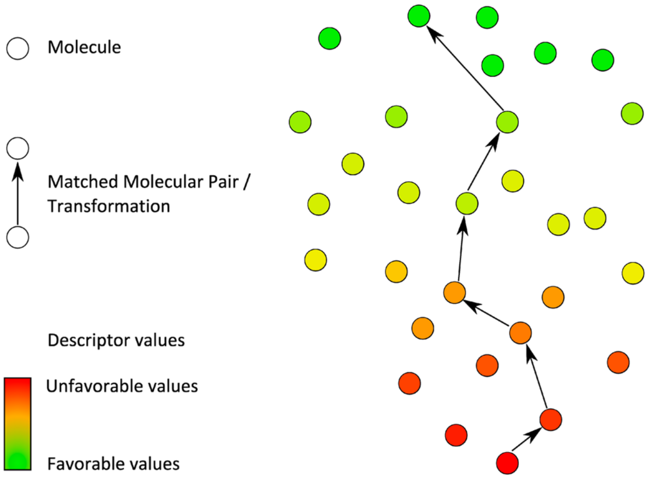
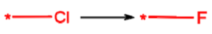
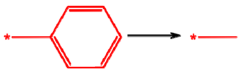
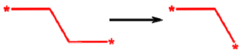
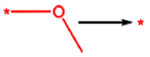
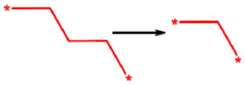
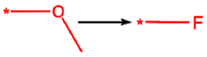


Figure 2. Compound pathway visualization. A compound pathway is delineated using arrows in the PCA projection of a hypothetical data set. Molecules are represented as nodes that are color-coded using a spectrum ranging from unfavorable to favorable descriptor values. Compounds connected by an arrow form an MMP and are related to each other by the corresponding transformation. Hence, the pathway follows a sequence of overlapping MMPs from an unfavorable to a favorable region of descriptor space.

Table 2. Conserved Transformations[a]

Transformation	Descriptor	A2AR	CB2	D2R	MOR
	MW	-15.97	-15.97	-15.97	-15.97
	rotN	0.00	0.00	0.00	0.00
	TPSA	0.00	0.00	0.00	0.00
	logP	0.00	0.00	0.00	0.00
	MW	-62.02	-62.02	-62.02	-62.02
	rotN	0.00	0.00	0.00	0.00
	TPSA	0.00	0.00	0.00	0.00
	logP	-0.55	-0.55	-0.55	-0.55
	MW	-14.02	-14.02	-14.02	-14.02
	rotN	-1.00	-1.00	-1.00	-1.00
	TPSA	0.00	0.00	0.00	0.00
	logP	-0.11	-0.11	-0.11	-0.11
	MW	-30.01	-30.01	-30.01	-30.01
	rotN	-2.00	-2.00	-2.00	-2.00
	TPSA	-8.85	-9.23	-9.23	-9.23
	logP	0.00	0.00	0.00	0.00
	MW	-28.03	-28.03	-28.03	-28.03
	rotN	-2.00	-2.00	-2.00	-2.00
	TPSA	0.00	0.00	0.00	0.00
	logP	-0.22	-0.22	-0.22	-0.22
	MW	-12.02	-12.02	-12.02	-12.02
	rotN	-1.00	-1.00	-1.00	-1.00
	TPSA	-9.23	-9.23	-9.23	-9.23
	logP	-0.11	-0.11	-0.11	-0.11

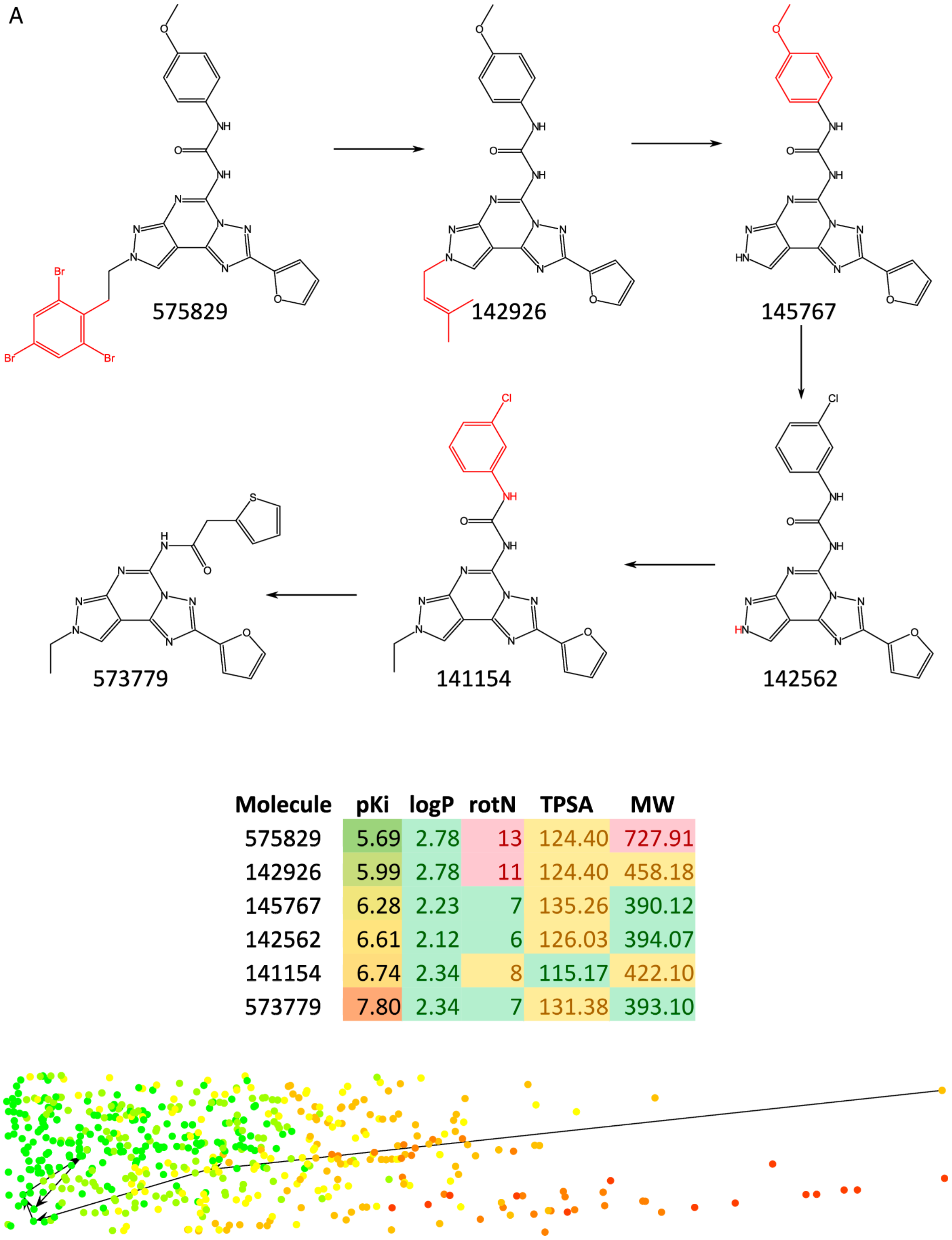
[a] Preferred transformations are listed that consistently occurred in all four compound data sets together with their average descriptor value changes.

Compound Optimization. Starting from compounds located in unfavorable property space, a series of set-dependent transformations were applied to predict new compounds. During each step, a transformation was randomly selected among those that modified descriptor values toward favorable regions. Transformation-based optimization was terminated if no favorable descriptor value changes were observed during subsequent iterations or when designed compounds entered favorable regions of descriptor space. For each compound, 20 independent optimization trials were carried out. Each trial was permitted to include a maximum of 20 steps. From all trials for a given compound, the one yielding the highest proportion of predicted compounds with database matches relative to the total number of designed compounds per trial was prioritized, as further discussed below.

Searching for Predicted Compounds. Each predicted compound was searched in ChEMBL. If the designed compound was not detected, a near neighbor search was carried out for database molecules having MACCS key[18] Tanimoto similarity[19] >0.9. Activity annotations of matched compounds or near neighbors were analyzed. If candidate molecules were found to have the same receptor antagonist annotation as the start compound, they were selected and their potency values were recorded to monitor potency progression among matches during optimization.

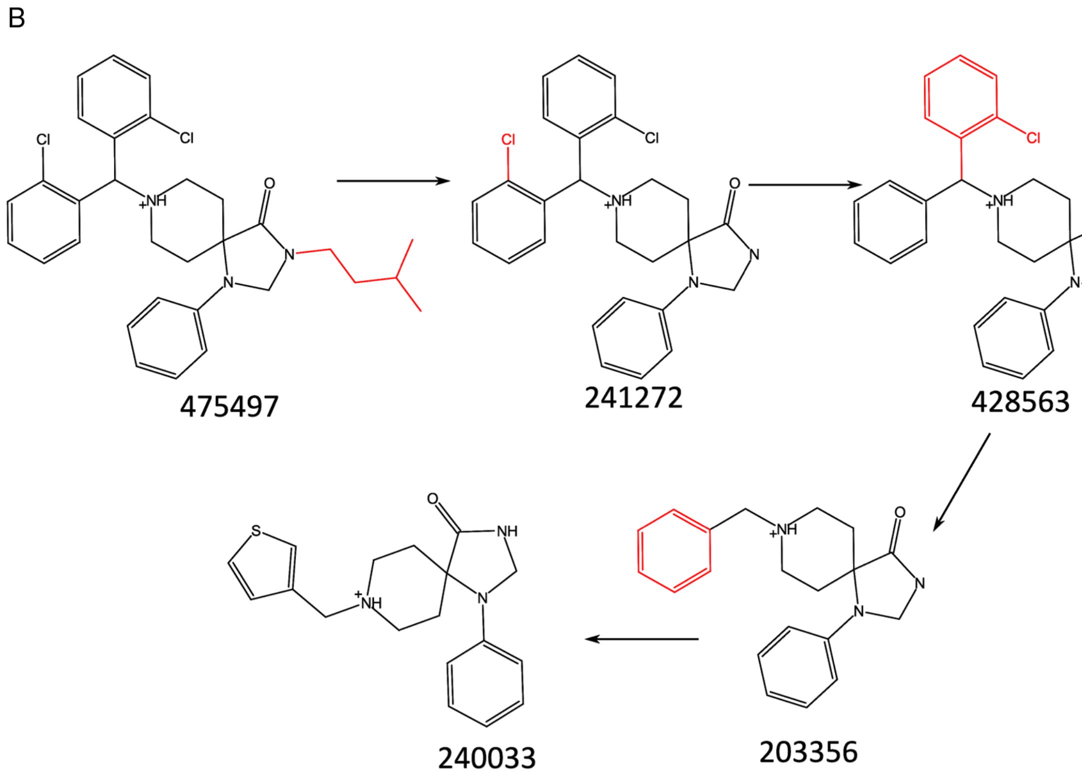
■ RESULTS AND DISCUSSION

Study Concept. We have been interested in investigating how to systematically optimize chemical properties of active compounds and "move" them through structural modifications into favorable regions of property space. We have selected four widely considered features (descriptors) that are known to account for drug development-relevant properties and for which unfavorable, intermediate, and favorable value ranges have been determined.[16] The selected properties included molecular size (MW),



Molecule	pKi	logP	rotN	TPSA	MW
575829	5.69	2.78	13	124.40	727.91
142926	5.99	2.78	11	124.40	458.18
145767	6.28	2.23	7	135.26	390.12
142562	6.61	2.12	6	126.03	394.07
141154	6.74	2.34	8	115.17	422.10
573779	7.80	2.34	7	131.38	393.10

Figure 3. continued



Molecule	pKi	logP	rotN	TPSA	MW
475497	4.89	4.21	11	27.99	536.22
241272	5.70	3.66	6	36.78	466.15
428563	6.11	3.77	5	36.78	432.18
203356	6.30	3.22	3	36.78	322.19
240033	6.88	2.89	3	65.02	328.15

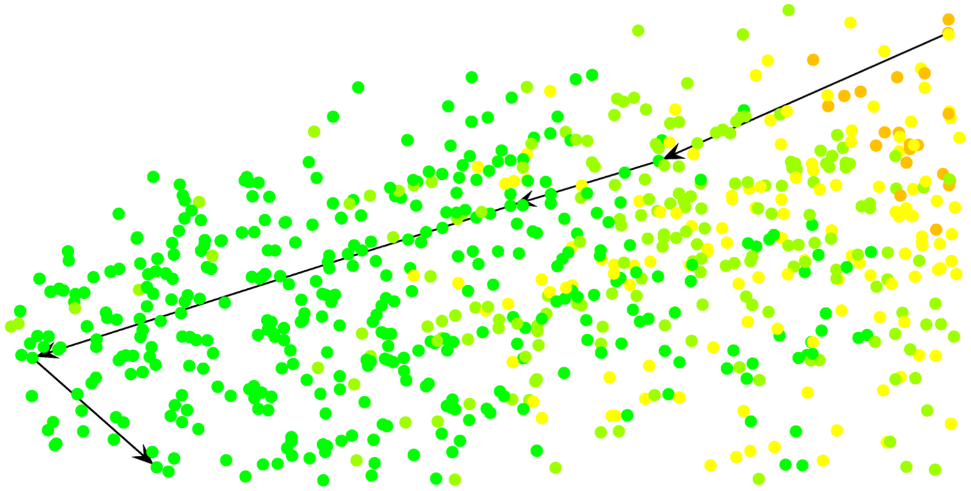


Figure 3. Compounds pathways. In (**A**) and (**B**), representative pathways from the A2AR and MOR data sets are shown, respectively. At the top of each representation, the compounds forming the pathway are shown and labeled with their ChEMBL ID. Substructures exchanged along the path are colored red. Below the structures color-coded property descriptor values are listed for each molecule. In addition, compound potencies (pK_i values) are reported using an analogous color code from green (lowest potency within the data set) over yellow to red (highest potency). At the bottom, the compound pathway is delineated in the corresponding section of the PCA projection of the data set.

polar surface area (TPSA), flexibility (rotN), and lipophilicity (logP). In contrast to the original classification scheme of Lobell et al., we did not include aqueous solubility in our analysis because solubility models available to us did not produce consistently accurate values. For evaluating sequences of structural changes, the chosen property space was suitable, especially because unfavorable and favorable regions in this space could be clearly distinguished for the compound sets under study (and separated in PCA projections).

A key question of our analysis has been whether it might be possible to derive structural modifications from data sets of known active compounds that display a general tendency to induce favorable changes in property space and that could then be applied to optimize compounds sharing the same activity. To these ends, we have applied the MMP concept to systematically identify chemical transformations and search for set-dependent transformations that occurred in different structural environments (i.e., different MMPs) and the subset of preferred transformation that consistently changed property values in a favorable manner.

Transformation Analysis. We first determined all MMPs and direction-dependent transformations in each of the four compound sets under study. Then, we identified transformations that were represented by at least 10 different MMPs, which dramatically reduced the number of candidate transformations, as reported in Table 1. The number of these frequent transformations ranged from 76 (D2R) to 240 (A2AR). For each qualifying transformation, average descriptor value changes were calculated for all corresponding MMPs per class, as illustrated in Figure 1. Next, we searched for frequent transformations that consistently moved compounds toward preferred regions of property space. The possibility to identify such transformations was *a priori* not unlikely. For example, considering the simplest case, a given transformation always changes molecular weight in a defined manner, regardless of the compound it is applied to, and if a transformation reduces molecular weight, it would generally be considered favorable.

In order to address transformation generality within a given set of active compounds, we searched for preferred transformations. As reported in Table 1, the majority of set-dependent transformations did not yield consistently favorable property changes. However, preferred transformations were identified in each set. For A2AR, CB2, MOR, and D2R, the number of preferred transformations was 47, 31, 30, and 18, respectively. Only six preferred transformations were conserved in all four data sets, as reported in Table 2. Because the set-dependent transformations were derived from compounds sharing the same activity, they are likely to retain activity if applied to an active compound. This is an important aspect bridging between data mining and compound design.

After identifying preferred transformations for each compound set, we next assessed their predictive capacity. Therefore, the set of MMPs representing each transformation was 10 times divided in half. For each training set, transformation-dependent descriptor value changes were determined and used to predict descriptor values of test set compounds, which were then compared with calculated test set values via 10-fold cross validation. These predictions were found to be highly accurate for all four data sets (more so than we might have expected), yielding R^2 values of 0.96 for D2R and 0.99 for A2AR, CB2, and MOR. Thus, preferred transformations yielded nearly identical changes in descriptor values toward favorable property space, regardless of the structural environment they occurred in, which reflected desired set-dependent generality. Previously, the potential structural context dependence of MMP-associated effects has been pointed out.[12] The high R^2 values obtained in our analysis indicated that there was relatively little context-dependence of MMP-based property effects for the compound sets we studied.

Detection of Compound Pathways. PCA projections revealed that compounds in all four data sets were widely distributed over unfavorable and favorable regions in property space. Hence, in the next step, we systematically searched the data sets for all MMP sequence pathways leading from a compound located in unfavorable property space to a compound in favorable space that involved preferred and other transformations. A model of such a pathway is shown in Figure 2. For A2AR, D2R, CB2, and MOR, 46,029, 4569, 1200, and 672 qualifying compound pathways were detected, respectively. On average, these pathways included 11.2 (A2AR), 9.3 (D2R), 15.2 (CB2), and 9.2 (MOR) compounds. Exemplary pathways with compounds and associated property values are shown in Figure 3. Hence, in this retrospective data set analysis, many MMP sequence pathways bridging between unfavorable and favorable regions of property space were found that involved preferred transformations.

Compound Optimization. Finally, we attempted to design compounds forming optimization paths using set-dependent transformations. At each step, only transformations were accepted that generated analogs with predicted favorable value changes for one or more descriptors while keeping other descriptor values within their current ranges. As starting points for compound design, all data set compounds were selected that mapped to unfavorable regions in the PCA projections and had a pK_i value greater than 7 (i.e., property optimization was modeled for relatively potent compounds). Depending on the data set, between 27 and 56 candidate compounds were identified as starting points (Table 3). These compounds were subjected to sequences of randomly chosen set-dependent transformations (see Methods) to design a series of new analogs. For each candidate compound, it was determined whether optimization trial(s) generated new analogs that reached favorable regions of property space. If so, we searched for these analogs in ChEMBL. If an analog was not found, a near neighbor search was carried out (see Methods). The results of our optimization trial are reported in Table 3. In this table, database search results are only reported for terminal analogs of optimization paths. The optimization trials revealed that compound optimization

Table 3. Optimization Trials[a]

data set		A2AR	CB2	D2R	MOR
optimization candidate compounds		36	35	27	56
did not reach favorable space		17	0	27	34
reached favorable space	no NN	8	11	0	17
	NN	8	24	0	4
	active NN	3	0	0	1

[a] For each data set, the number of compounds subjected to optimization trials (”optimization candidate compounds”) is given, and the subsets of these compounds for which predicted analogs did not reach or reached favorable regions of property space are reported. For the final analog of an optimization trial reaching favorable space, the results of near neighbor analysis are also provided. “No NN” and “NN” means that no near neighbor and one or more near neighbors of the final analog were found in ChEMBL, respectively. In addition, “active NN” means that a near neighbor sharing the same receptor antagonist activity annotation was identified.

at least partly succeeded for three of four compound classes, with the exception of D2R. In the latter case, no analogs of any of the 27 start compounds reached favorable property space, although compounds were found to move in the right direction, albeit in too small steps. D2R also produced the overall smallest number of set-dependent and preferred transformations. By contrast, all analogs derived from all 35 CB2 starting points reached favorable space. For 24 compounds, near neighbors of

A

Cpd	logP	rotN	TPSA	MW	Transformation	Database cpd	pKi
1	1.68	14	155.72	607.03		[EM] 596135	8.30
2	1.79	11	155.72	573.01	[R1]C(F)(F)F>>[R1]Cl	[EM] 592896	8.80
3	1.90	10	155.72	555.02	[R1]F>>[R1]	[NN] 592896	8.80
4	2.01	9	155.72	537.03	[R1]F>>[R1]	[NN] 592896	8.80
5	2.12	8	155.72	519.04	[R1]F>>[R1]	[NN] 592896	8.80
6	2.23	7	155.72	485.07	[R1]Cl>>[R1]	[NN] 596135	8.30
7	2.67	7	113.85	435.68	[R1]S(=O)(=O)[R2]>>[R1]C[R2]	[NN] 596133	8.96

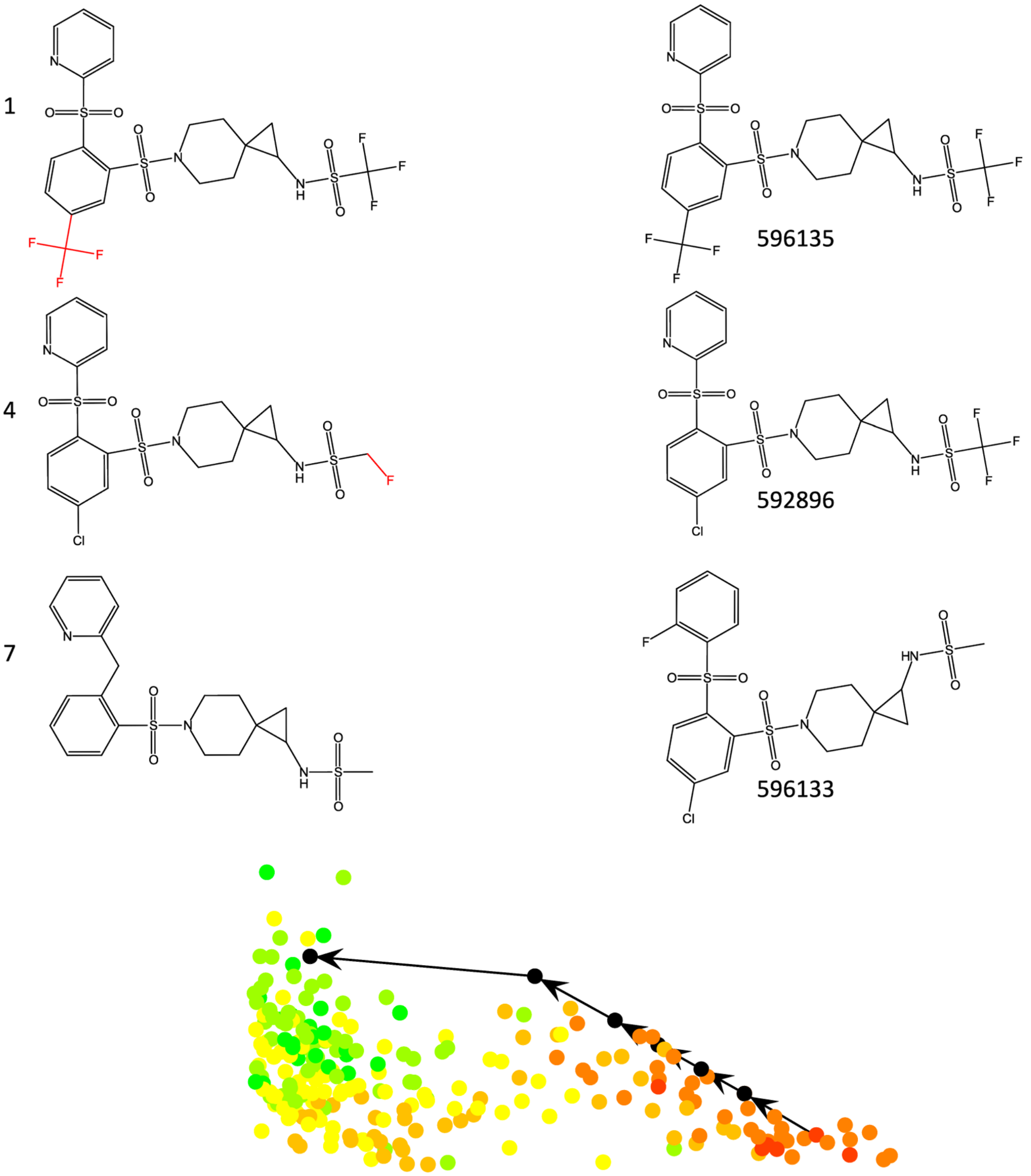


Figure 4. continued

B

Cpd	logP	rotN	TPSA	MW	Transformation	Database cpd	pKi
1	3.11	16	122.10	538.30		[EM] 473440	9.15
2	3.00	13	112.87	494.28	[R1]COC>>[R1]	[EM] 480756	8.60
3	2.89	12	112.87	480.26	[R1]CC[R2]>>[R1]C[R2]	[NN] 473439	8.82
4	2.78	11	114.45	466.22	[R1]C>>[R1]	[NN] 480749	8.02
5	2.70	10	114.50	452.20	[R1]CO[R2]>>[R1]O[R2]	[NN] 245848	10.00
6	2.60	7	105.20	408.20	[R1]COC>>[R1]	[NN] 240030	8.89

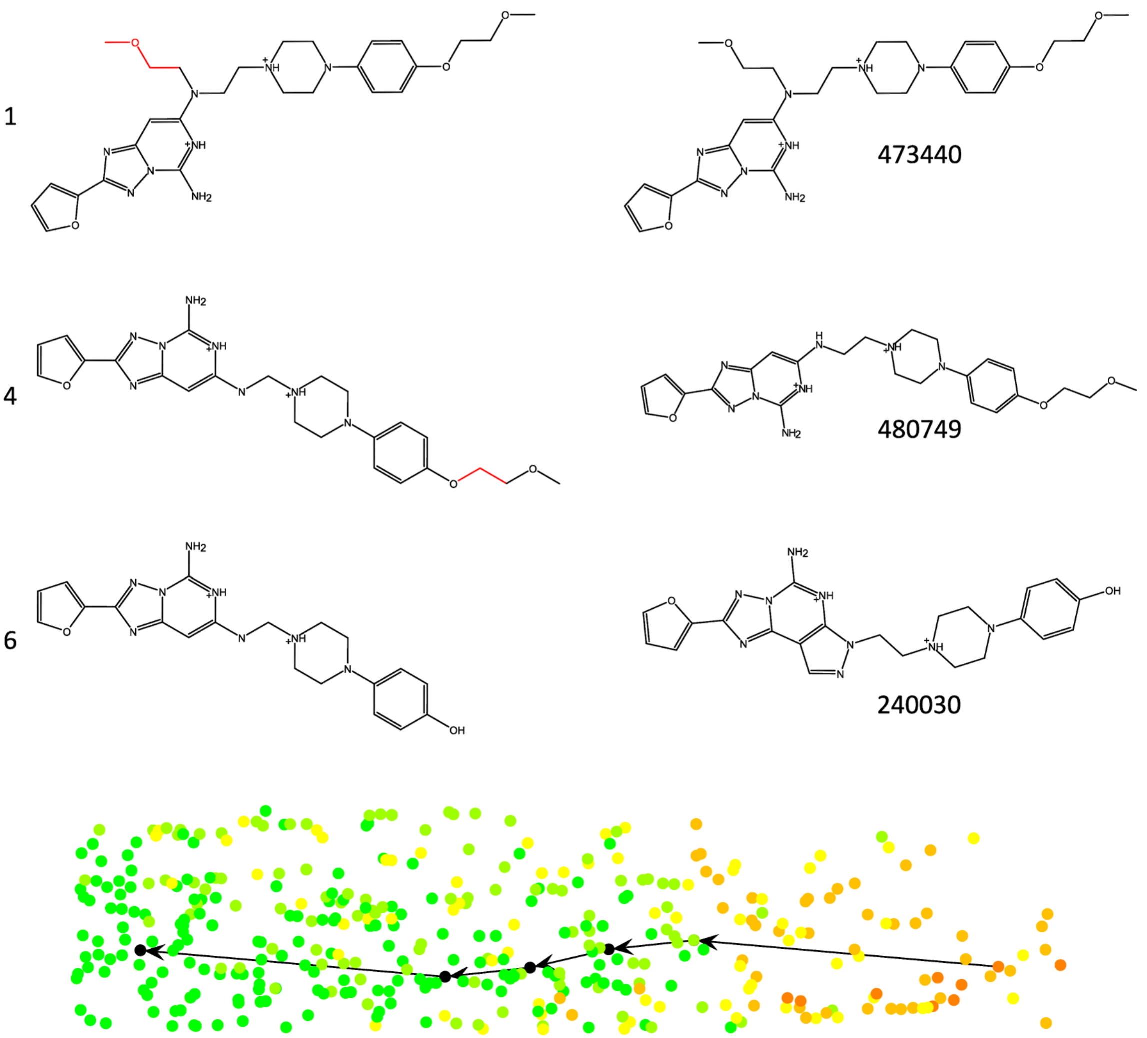


Figure 4. Compound optimization. In (A) and (B), exemplary compound optimization paths are shown originating from compound 596135 of data set CB2 and from compound 473440 (A2AR), respectively. Representation elements are according to Figure 3. Structures of designed compounds (left) and database matches (right) are shown in the middle of the figure and are numbered according to the table insert at the top. For each designed molecule, predicted descriptor values are reported in the table insert. For matches and near neighbors, potency values are given. In the table insert, database compounds that exactly matched designed compounds are designated "EM" and near neighbors of designed compounds "NN". At the bottom, designed compounds (black nodes) were mapped into the PCA projection of the data set on the basis of their descriptor values. The optimization path formed by these predicted compounds is traced.

the terminal analog of a path were identified in ChEMBL, but none of these closely related compounds was known to share CB2 activity. For MOR, trials for 22 of 56 candidates succeeded, and in five of these cases, near neighbors were identified, one of which was known to have MOR antagonist activity. Furthermore, for A2AR, 19 of 36 candidate compounds yielded derivatives that reached favorable property space. For 11 terminal compounds, near neighbors were identified, and three of these were annotated with A2AR antagonist activity.

Figure 4 shows the results of two successful optimization trials for CB2 and A2AR, respectively. In a number of successful optimization trials, intermediate pathway compounds also had exact matches or near neighbors with shared activity, which was also the case for the two exemplary trials in Figure 4. It can be seen how designed compounds approached and reached favorable property space while essentially retaining comparable potency levels. Thus, set-dependent transformations were activity-conservative, consistent with principles of the approach.

Taken together, the results in Table 3 and Figure 4 indicate that compound property optimization on the basis of set-dependent transformation is a feasible task. In light of the database search results and detected near neighbor relationships, a number of designed compounds might also be attractive candidates for experimental evaluation. Hence, the compound set-centric and transformation-based compound design strategy introduced herein should merit further investigation using different compound classes and molecular properties.

Concluding Remarks. In this study, we have addressed the question whether structural modifications can be identified for sets of compounds sharing the same activity that display a general tendency to further improve molecular properties. If so, such modifications might be applied for compound design. For the purpose of our analysis, we have adapted the MMP and transformation concepts that are suitable for the systematic identification of chemical changes within variable structural contexts, i.e., modifications that are shared by pairs of structurally distinct compounds. The MMP concept is not the only possible route to prospective compound design and optimization. For example, knowledge-based sets of structural transformation have also been utilized.[20] In our study, varying numbers of set-dependent and preferred transformations were identified for four different data sets that induced favorable molecular property changes in different compounds. In these data sets, we identified large numbers of MMP sequence pathways that led from active compounds located in unfavorable regions of property space to others in favorable regions. We then devised a compound design protocol applying randomly selected transformations to iteratively generate derivatives of compounds located in unfavorable property space and produce compound paths leading into favorable space. For three of four compound sets, many optimization trials were successful and often yielded attractive derivatives. Lead optimization is a multiparametric process that requires the improvement of druglike molecular properties alongside compound potency, consistent with the basic ideas underlying our approach. In summary, the approach introduced herein closely combines compound data mining and prospective compound design and can provide design suggestions for experimental studies. Our findings indicate that set-dependent transformations can be applied, even in a random fashion, to generate compounds with favorable molecular properties.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Sheridan, R. P. The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103−108.

(2) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271−285.

(3) 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.

(4) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739−7750.

(5) Wassermann, A. M.; Dimova, D.; Iyer, P.; Bajorath, J. Advances in Computational Medicinal Chemistry: Matched Molecular Pair Analysis. *Drug Dev. Res.* **2012**, *73*, 518−527.

(6) Wassermann, A. M.; Bajorath, J. Large-Scale Exploration of Bioisosteric Replacements on the Basis of Matched Molecular Pairs. *Future Med. Chem.* **2011**, *3*, 425−436.

(7) 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.

(8) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.

(9) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched Molecular Pairs As a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006**, *46*, 6672−6682.

(10) Keefer, C. E.; Chang, G.; Kauffman, G. W. Extraction of Tacit Knowledge from Large ADME Data Sets via Pairwise Analysis. *Bioorg. Med. Chem.* **2011**, *19*, 3739−3749.

(11) Dossetter, A. G. A Matched Molecular Pair Analysis of in Vitro Human Microsomal Metabolic Stability Measurements for Methylene Substitution or Replacements − Identification of Those Transforms More Likely To Have Beneficial Effects. *Med. Chem. Commun.* **2012**, *3*, 1518−1525.

(12) Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, 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.

(13) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(14) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111−2120.

(15) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME - the Konstanz Information Miner: Version 2.0 and Beyond. *SIGKDD Explor. Newsl.* **2009**, *11*, 26−31.

(16) Lobell, M.; Hendrix, M.; Hinzen, B.; Keldnich, J.; Meier, H.; Schmeck, C.; Schohe-Loop, R.; Wunberg, T.; Hillisch, A. In Silico ADMET Traffic Lights as Tool for the Prioritization of HTS Hits. *ChemMedChem* **2006**, *1*, 1229−1236.

(17) *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(18) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.

(19) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(20) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; Simeons, F. R. C.; Stojanovski, L.; Prat, A.; Seidah, N. G.; Constam, D. B.; Bickerton, G. R.; Read, K. D.; Wetsel, W. C.; Gilbert, I. H.; Roth, B. L.; Hopkins, A. L. Automated Design of Ligands to Polypharmacological Profiles. *Nature* **2012**, *492*, 215−220.