JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

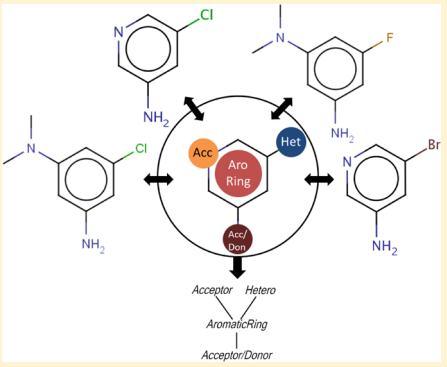
pubs.acs.org/jcim

Fuzzy Matched Pairs: A Means To Determine the Pharmacophore Impact on Molecular Interaction

Tim Geppert* and Bernd Beck

Department of Lead Identification and Optimization Support, Boehringer-Ingelheim Pharma GmbH & Co. KG, Birkendorferstrasse 65, 88397 Biberach an der Riss, Germany

ABSTRACT: Within this work, a methodological extension of the matched molecular pair analysis is presented. The method is based on a pharmacophore retyping of the molecular graph and a consecutive matched molecular pair analysis. The features of the new methodology are exemplified using a large data set on CYP inhibition. We show that Fuzzy Matched Pairs can be used to extract activity and selectivity determining pharmacophoric features. Based on the fuzzy pharmacophore description, the method clusters molecular transfers and offers new opportunities for the combination of data from different sources, namely public and industry datasets.



INTRODUCTION

The knowledge about bioisosteric and nonbioisosteric transformations is helpful in all phases of the drug discovery pipeline, but it is especially useful during lead optimization.[1,2] In this phase, researchers face the challenge to optimize molecules in a multiparameter space. Transformations that follow the principle of bioisosterism in a subset of parameter space are desired, and they should form activity cliffs in the rest of the space.[3] For example, it is important to maintain target activity, while improving on a variety of ADMET properties, or vice versa.

One of the methods often used in this context is the Quantitative Structure Activity Relationship (QSAR) analysis.[4] QSAR tries to establish a relation between the molecular structure and a property of interest by using machine learning and regression. There are several possibilities to describe a molecular structure in this context. Fragment-based QSAR relates molecule fragments to a property, while descriptor-based QSAR approaches incorporate calculated molecular properties to determine the relationship between an activity and a structure.[5,6]

Pharmacophoric features and molecular fields, as a description of molecular structures, also have a long tradition in the area of QSAR analyses, as well as virtual screening.[7−9]

The recently introduced Matched Molecular Pairs (MMP) method is also seen as a useful approach in this area.[10,11] A MMP is formed between two molecules if they share a constant "core" structure while one other part of the molecule is changed. The constant part is called the key structure, while the variable part is called the value transfer of this MMP. An algorithm for MMP generation was proposed by Hussain and Rea.[12] Based on this algorithm, several ADMET-related end points have been analyzed to determine informative value transformations. There are published studies on hERG, P450, solubility, and permeability by Gleeson et al.,[13] as well as studies on plasma protein binding and others in a landmark study by Leach et al.[14]

The MMP approach can be seen as a fragment-based QSAR method, which describes the potential impact of exchanging one part of the molecule. It suggests a substructure that has shown better properties in the past and can have a positive impact on the regarded property of a new "core" structure. This is why there has been great interest in the literature for MMPs.[15] In addition, it could be shown that, by retrospective analysis of in-house and external datasets, a large knowledge base of value transformations can be extracted.[16] A prototype for such a knowledge base is the so-called SwissBioisostere database, which is based on public domain transformations.[17]

The initial works on MMPs have been focused on substructure transfers without taking the local environment of this transfer into account.[14] The results are based on the implicit assumption that changes on an observed molecular parameter can be only attributed to the molecular properties of the substituted substructure. Recent studies showed that this simplification often does not hold true.[18,19] Papadatos et al.,[20] as well as Warner et al.,[21] evaluated the influence of the local environment on MMPs for several properties. Their results show that, in many cases, the environment indeed has a pronounced influence on MMP transfers.[20,21] The studies suggest that the local environment should be included in the MMP generation process.

There are different approaches available to include the environment of the MMP transfer. First of all, the exact

Received: November 25, 2013
Published: March 4, 2014

dx.doi.org/10.1021/ci400694q | J. Chem. Inf. Model. 2014, 54, 1093−1102

environment can be described using the molecular graph and its corresponding smiles for the comparison, as it has been done in the WizePairs approach.[22] A more abstract form of the description of the environment, such as the molecular pharmacophore concept, can also be used. The pharamcophore is well-known in the QSAR and virtual screening area but, so far, has been rarely used for MMP approaches.

It has to be noted that the inclusion of environment information reduces the number of available transfers with a statistically sound number of occurrences, which is the drawback of such an approach. Different abstraction levels, from an atom-based (low abstraction) description to a pharmacophore-based (high abstraction) description, will reduce the number of transfers to different amounts. This highlights that the number of underlying observations is generally one of the challenges of the MMP method.[22]

In this study, we will introduce an extension to the MMP framework, called the fuzzy matched pair (FMP) approach. The FMP method combines the favorable capabilities of the classical MMP approach with a less-rigid molecular description, using a reduced molecular graph, which is based on pharmacophore feature types. We believe that FMPs can aggregate similar transfers and thus show important molecular features that are responsible for a corresponding molecular property change. In this work, we will describe the methodological blueprint of the fuzzy matched pair framework and show how they can be applied to a prototypic activity dataset to extract pharmacophoric features that determine molecular activity. We are convinced that molecular recognition is well-described by pharmacophoric interaction points, as the broad literature on pharmacophores in the area of drug design illustrates.[23]

METHODS

MMP Algorithm. For the MMP generation, we followed the method published by Hussain and Rea.[12] The algorithm consists of the following steps. First, the molecules are standardized using OEChem.[24] In a second step, the molecules are cut at all exocyclic single bonds. Based on these fragmentations, single-, double-, and triple-cut MMPs are generated. They are stored in a table structure with the larger (key) part building the constant part and the smaller part the variable (value) part. The matched pairs are formed by simple string matching on the unique smiles of the constant part. The variable parts of the matched constant parts form the matched pair transfers.

Nearest-Neighbor Environment. Based on the described fragmentation, the nearest-neighbor environment is calculated as follows. All R-Group anchors of the constant fragment are determined. Starting with these atoms, a breadth first search algorithm on the molecular graph is executed. The depth of this search is equal to the number of nearest neighbors that are included within the environment information. The calculations are implemented using the OEChem graph structure and the Java language.[24] The method is also applied for the calculation of the pharmacophore environment. For these calculations, the depth is set to 1.

Pharmacophore Retyping. To retype the molecule to its pharmacophore representation, we use pharmacophoric types based on the Lipinski SMARTS pattern definitions taken from RDKit.[25] As an input to the pharmacophoric retyping, a molecule or molecule substructure is needed. Based on this structure, all pharmacophoric types are determined in a top-down method. At first, all donor/acceptor features are determined. The remaining atoms are checked for donor or acceptor features. Remaining noncarbon atoms then get the heteroatom type. By using the ring detection feature of OEChem, all ring systems are determined. All aromatic ring systems are identified with the OEChem aromaticity detection. Figure 1 shows an example retyping of a structure from ChEMBL version 16.[26]

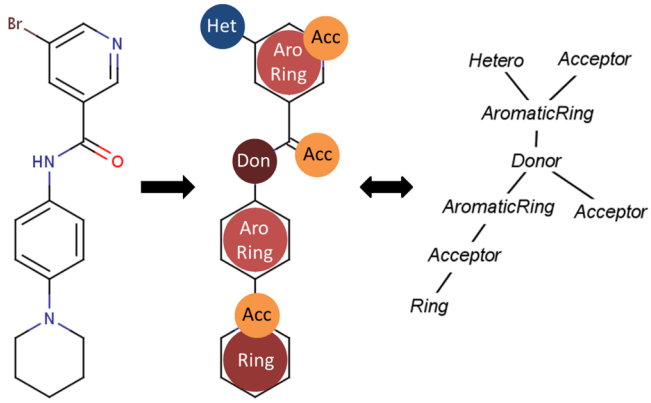


Figure 1. Retyping of a ChEMBL structure (ID: CHEMBL 1333282).[26] The pharmacophore patterns are determined, and a new molecular graph that consists of pharmacophore types is generated.

Dataset. We used a public data set on CYP inhibition to show and explain the different MMP approaches within this publication. ChEMBL version 16 was downloaded from the online repository and installed on a MySQL server. We extracted the dataset by Veith et al.[27] from the database. The dataset contains more than 12k compounds with activity data for the different CYP isoforms (CYP2C19, CYP2C9, CYP2D6, and CYP3A4). All data points of the dataset that are annotated as active or inactive are retained. This resulted in 12 902 data points for CYP2C19, of which 5837 are marked active and 7065 are marked inactive. It presents a typical example for lead optimization. In this case, we optimize against an antitarget. Veith et al.[27] stated that, based on assay limitations, it is not possible to discriminate between inhibitors and substrates in the dataset. Therefore, we will refer to active and inactive molecules, to be consistent with the classification used by Veith et al.[27] All structures were standardized and neutralized using the JChem Standardizer.[28]

Data Aggregation. The MMP generation results in environment-independent exchanges of molecular substituents. Based on the aggregation of MMPs with the same substitution, we also aggregate the biological data. The biological data are thereby classified into one of the following classes. An exchange from active to inactive is defined as a decrease, an exchange from inactive to active is defined as an increase, and we define all remaining substitutions as neutral. The complete MMP transfers can be further grouped based on their local environment. To determine differences between the environment subgroups and the complete transfer distribution, we use a bootstrap method, as described below. Based on bootstrap sampling, we test whether an occurrence of a sample distribution is within the 95% confidence interval of the background distribution. If a sample is outside this interval, we mark it as different.

Statistical Evaluation. To assess the statistical relevance of the different matched pair methods, we use the resampling

dx.doi.org/10.1021/ci400694q | J. Chem. Inf. Model. 2014, 54, 1093−1102

method by Efron.[29] In summary, the method is based on the following steps. Based on a distribution Ω, 1000 samples Y are drawn with replacement. The mean of each sampled Y is recorded and consecutively ordered in decreasing order for the 1000 results. The value at position 25 and 975 are defined as the confidence interval (CI) borders of the sample mean. This results in a 95% CI. To use the method for the assessment of the relevance of a single matched pair, we apply it as follows. We define a pair $A \rightarrow B_1$ as the path from A to B_1. For each A that forms paths to $B_1, ..., B_n$ with $n \geq 3$, we determine the background distribution for taking any path from A to any B_i $\{i = 1, ..., n\}$ by calculation of eq 1:

$$\Omega = \sum_{i=1}^{n} A \rightarrow B_i \tag{1}$$

We then sample the number of occurrences of a specific path $A \rightarrow B_i$ from the background distribution Ω and compare the distribution for the increase, decrease, and neutral class with the distribution in the background distribution, using the above-described resampling method. We define a result as different from the background distribution if it is outside the 95% CI of the background distribution. It is important to note that this approach is independent of any assumption about the background distribution. For the statistical evaluation, we use the R framework.[30]

All the above-described algorithms are implemented as in-house KNIME nodes for the workflow environment KNIME version 2.7.[31] The modular-node-based workflow can be seen as analogous to the columns in Figure 2.

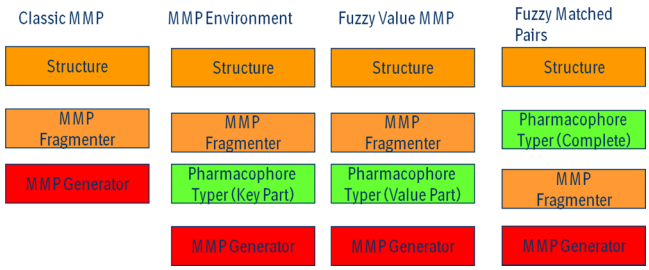


Figure 2. Overview about the transition between the generation of matched molecular pairs and fuzzy matched pairs. Each box encapsulates a part of the algorithm and has been implemented in the workflow tool KNIME.[31]

RESULTS

We will now describe how the different MMP approaches can be applied within lead optimization. In Figure 2 the relation between the MMP approaches is shown with the classic MMP approach at one end and the novel FMP method on the other end.

To present the fuzzy matched pairs approach, we use the previously described dataset from Veith et al.[27] The dataset is selected for our FMP analysis because there is broad literature on the four CYP isoforms, which enables us to assess the results of our FMP pharmacophore hypothesis. Based on the multitarget information within the dataset, we can also evaluate if we are able to determine features that are responsible for selectivity between the different isoforms of CYP. Since Boehringer-Ingelheim has an in-house panel of measured CYP inhibition, we can assess whether the results based on the literature data correlate to findings based on in-house measurements and, thus, can be added to the knowledgebase about CYP inhibition.

Classical Matched Molecular Pairs (MMP). In the following paragraph, we will describe results based on the classical MMP approach. The 12 902 molecules with annotated CYP2C19 activity classification were fragmented into 141 205 transfers, 1075 of which had more than 20 occurrences, which is widely used as the lower threshold for the extraction of statistically meaningful transfers.[14] Table 1A presents six transfers observed with a high frequency, in which more than 90% of the underlying transfers do not change potency against CYP2C19. They can be seen as bioisosteric replacements, with regard to this end point. In Table 1B, the six most frequently occurring transformations that result in a change from active to inactive (decrease) or inactive to active (increase) for more than 30% of the underlying molecule pairs are listed. We also indicated the transfers that have a relevant difference to the background distribution with arrows.

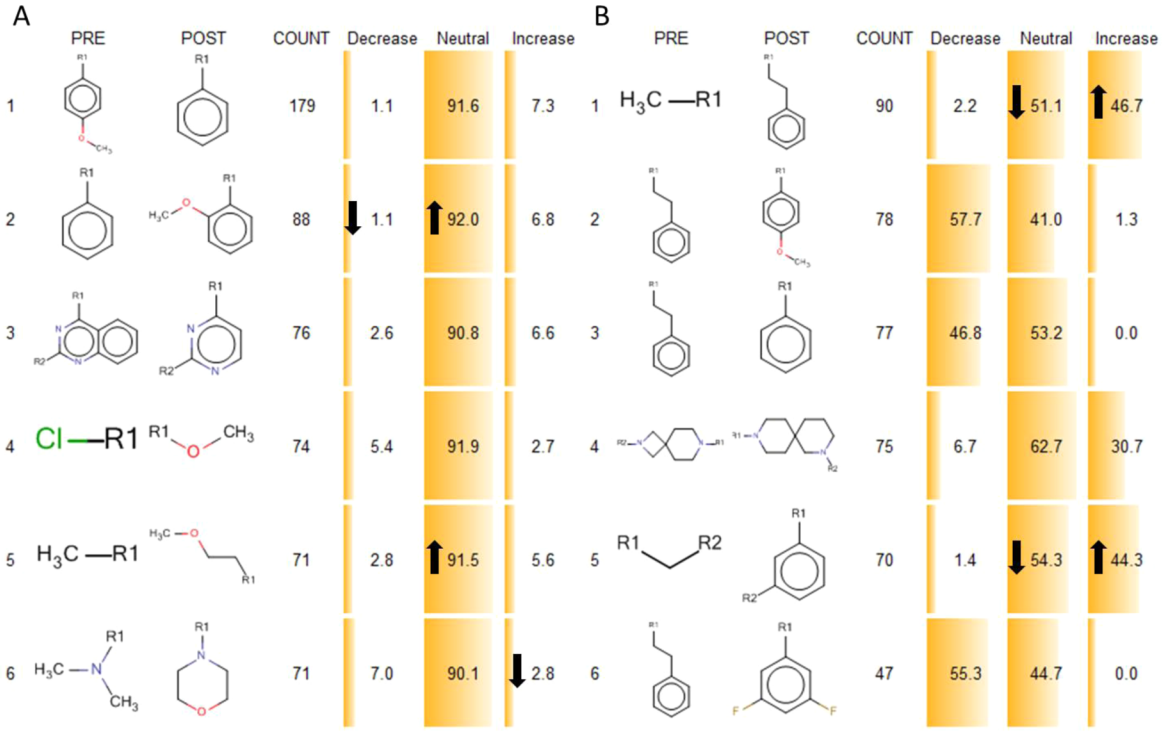
Introduction of a methoxy group to a benzyl ring in the para- and ortho- position has no influence on potency (Table 1A, rows 1 and 2). Also, the replacement of a chloro by a methoxy group is neutral (Table 1A, row 4). The introduction of aromatic ringsystems increases the activity in 46% of the observed transfers (Table 1B, row 1). This transfer is also outside the 95% CI of exchanging methyl with any other R-Group. This underlines that this transfer is statistically relevant and improves activity. The same is true for the introduction of a linking aromatic ring system (row 5). The observation is in accordance with a recent publication by Ritchie et al.[32] There, it is stated that the CYP2C19 activity is positively impacted by the number of carbo-aromatics.

MMP Environment Approach. Following the outline of Figure 2, we now test the pharmacophore environment surrounding the replacement structure (value) of the MMP transfer. Figure 3 shows how the retyped pharmacophore graph can represent a set of similar substructures together. The clustering combines also features within ring systems and outside ring systems.

Figure 4 presents the environment influence on the activity profile for selected transfers shown in Table 1. In Figure 4, the neutral exchange of methoxy benzene to phenyl with its underlying pharmacophore environments is depicted. It can be seen that the environment in the case of the introduction of the methoxy group to the benzene ring has no influence on the property. Therefore, this is a bioisoster exchange in all observed environments. A possible explanation is that the environment of the exchange, in this case, is already the aromatic ringsystem to which the methoxy moiety is attached. As outlined in Figure 4B, the introduction of a metasubstituted aromatic ring as linker increases the activity, irrespective of the environment, which again supports the notion that increasing the number of carbo-aromatic atoms increases activity. The exchange of an ethyl with a methyl linker next to an aromatic ring has variable effects on activity (Figure 4C). If it is next to an acceptor, it increases the number of active compounds, whereas next to an acceptor/donor feature, the opposite is observed. The observations for the decreasing group are also outside the 95% CI of the complete background. This transfer exchanges the distance between the pharmacophore feature and the aromatic ring system which could explain the pronounced influence on the CYP2C19 behavior. As mentioned previously, the introduction of aromatic ringsystems results in an increase

dx.doi.org/10.1021/ci400694q | J. Chem. Inf. Model. 2014, 54, 1093−1102

Table 1. Frequently Occurring (A) Bioisosteric (>90% Neutral) and (B) Nonbioisosteric (>30% Decrease or >30% Increase) Transformations[a]



[a]Transfers are indicated by the start substructure (PRE) and transferred substructure (POST). The number of occurrences (COUNT), as well as the percentage of transfers that reduce activity (Decrease), increase potency (Increase), and have no influence (Neutral) is indicated. The table is sorted by frequency in decreasing order. Bold black arrows in the Decrease, Neutral, and Increase columns indicate that this group is enriched (up) or reduced (down), in comparison to the background distribution.

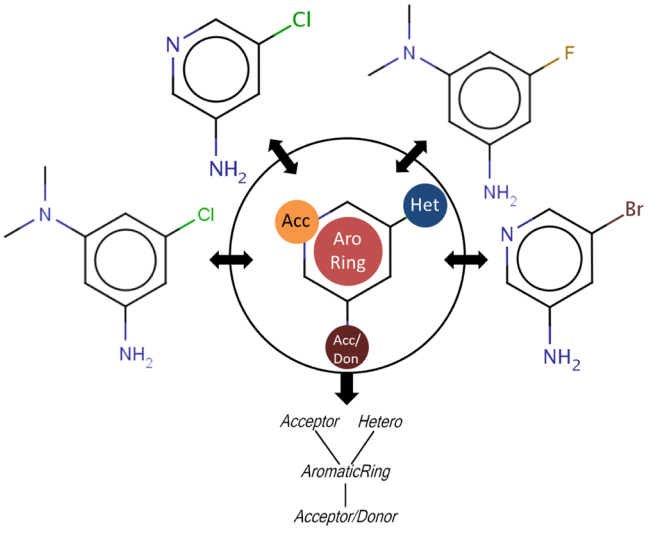


Figure 3. Overview about different molecular graphs that result in the same retyped pharmacophore graph.

in potency against CYP2C19 in several cases (see Figure 4B, as well as Table 1B).

For most of the MMP transfers in this example, the local environment has no influence on the potency distribution, in comparison to the analysis without the environment information. It underlines that the transfers are stable within most of the environments observed.

Value Fuzzy Matched Pairs. The next paragraph will introduce the MMP extension called Value Fuzzy Matched Pairs (VFMP). For VFMP, the classical MMP generation is followed by a retyping of the substituent. The transfers are then calculated for the pharmacophore-type variable parts. The VFMP approach can cluster similar structures together and highlights pharmacophoric features that are prevalent in a variety of different MMP transfers as presented in Figure 3. In Table 2A, the top six neutral VFMP transfers sorted by the number of occurrences are presented. The exchange of an acceptor (Acc) to an acceptor/donor (Acc/Don) feature is neutral, with respect to CYP2C19 activity, in more than 90% of the observed transfers, as is the introduction of a second heteroatom at an aromatic ring (Table 2A, second row). Table 2B presents the analysis for transfers, which leads to a change in activity classification for more than 30% of the underlying molecules. The removal of an aromatic ring leads to a reduction of activity for more than 30% of all analyzed transfers (Table 2B, second row), while the exchange of an aromatic ring to an acceptor reduces the number of active molecules by 38% (Table 2B, first row), which is consistent with the observations of the classical MMP analysis. The exchange of an aromatic to an aliphatic ring is disfavored in 40% of observed cases (Table 2B, row 5). This observation is again in agreement with the observation of Ritchie et al.[32] The exchange of an acceptor next to an aromatic ring with a donor feature increases the inhibitory potential by 34% (Table 2B, row 9), which is supported by a publication of Sun et al.[33] They also observed that additional donor features increase the activity to CYP2C19 by a feature extraction method on the same dataset.[33] It is of importance to note that all the transfers in Table 2B exhibit a relevant

dx.doi.org/10.1021/ci400694q | J. Chem. Inf. Model. 2014, 54, 1093−1102

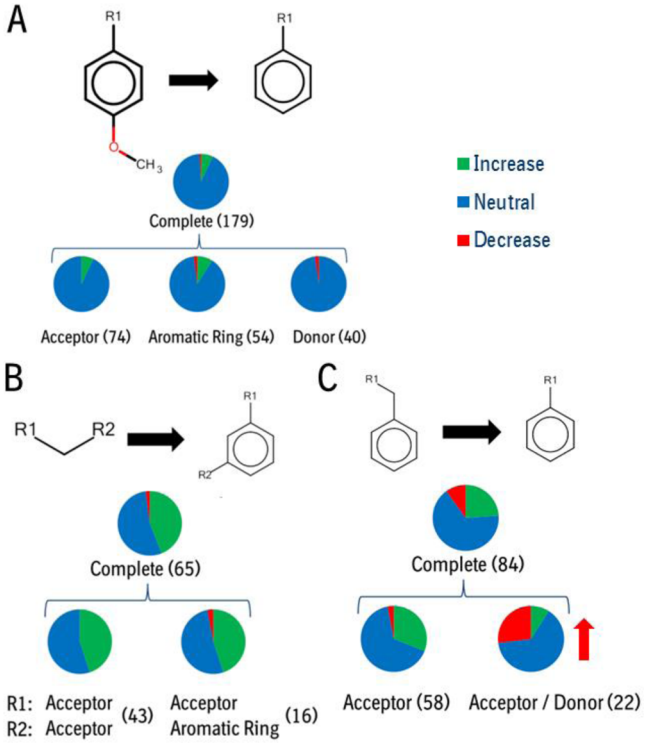


Figure 4. Overview about the environment influence on the potency of selected MMP transfers. Complete denotes the transfer distribution of the entire class, while the subclasses denote the transfer properties based on the first pharmacophore type at the attachment point. The number of underlying transfers is annotated in brackets. The red arrow indicates that the decreasing fraction of transfers is enriched (up), in comparison to the transfer without taking the environment into account.

difference to the background distribution, which, in most cases, is explained by a reduction of the neutral class.

Again, the clusters can be split into groups based on the underlying environment information. The corresponding results are presented in Figure 5.

The reduction of an aromatic ring system linker to a R1-C-R2 linkage (Figure 5A) is reducing the number of active compounds. In combination with an acceptor at R2 and an acceptor and heteroatom at R1, this transfer is neutral and, therefore, outside the 95% CI of the complete transfer group. Introduction of an aromatic ring increases the activity in all environments (Figure 5C) and is most pronounced in the proximity of a donor/acceptor feature. For this environment, the number of neutral transfers is reduced and the number of transfers that lead to enhanced activity is enriched in comparison to the complete group. Introduction of an aromatic ring with two acceptors is favorable in all environments (Figure 5B). Replacing an aromatic ring system connected to two acceptors with an aliphatic ring system connected to two acceptors decreases the number of active compounds in all environments (Figure 5D). In this case, the size of the molecule is not changed and we have a direct comparison of the influence of aromaticity on the inhibition. In addition to the environment groups, we can further break down the pharmacophore transfers based on the underlying substructures. An example is shown for the pronounced exchange of aromatic to aliphatic ring systems with two acceptors (Figure 6).

The complete data are divided into two transfer structures, both of which present an exchange to a morpholino moiety. This heteroaliphatic ring system has lower activity than the aromatic ring and the heteroaromatic ring. The highest difference is observed for the transfer from an aromatic to a heteroaliphatic ring system. This result is again consistent with the finding of Ritchie et al.[32]

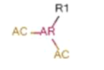
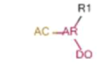
Fuzzy Matched Pairs. In the previous part, we used the pharmacophore retyping as a post-processing step, consecutive to the MMP generation. For the Fuzzy Matched Pairs (FMP) method, we will use the pharmacophore typing before the

Table 2. Frequently Occurring (A) Bioisosteric (>90% Neutral) and (B) Nonbioisosteric (>30% Decrease or >30% Increase) VFMP Transformations[a]

A

	PRE	POST	COUNT	Decrease	Neutral	Increase
1	R1—AC—R2	R1—AD—R2	222	↓ 6.8	↑ 91.9	↓ 1.4
2		R1—AR—Het	190	5.3	92.1	2.6
3		R1—AR—AC	133	1.5	91.0	7.5
4		R1—AR—R2	121	6.6	90.1	3.3
5	AC—HR—AC—R1	R1—AD—CH_3	115	0.9	↑ 91.3	↓ 7.8
6	AC—R1	R1—AD—CH_3	112	↓ 4.5	↑ 92.9	↓ 2.7

B

	PRE	POST	COUNT	Decrease	Neutral	Increase
1	AR—R1	AC—R1	313	↑ 38.3	↓ 60.4	↓ 1.3
2	R1—AR—R2	R2—R1	274	↑ 30.3	↓ 67.2	2.6
3	R1—AR—AC		246	↑ 30.1	↓ 64.6	5.3
4			215	↓ 0.9	↓ 64.2	↑ 34.9
5	AC—HR—AC—R1	AC—AR—AD—R1	187	3.2	↓ 56.7	↑ 40.1
6	R2—AR—AC—R1		133	↑ 42.1	↓ 54.9	3.0

[a]Transfers are indicated by the start substructure (PRE) and transferred substructure (PSOT). The number of occurrences (COUNT), as well as the percentage of transfers that reduce activity (Decrease), increase potency (Increase), and have no influence (Neutral) is indicated. The table is sorted by frequency in decreasing order. Legend: AC, acceptor; AD, acceptor/donor; DO, donor; AR, aromatic ring; HR, aliphatic ring; and Het, heteroatom. Black arrows in the decrease, neutral, and increase columns indicate that this group is enriched (up) or reduced (down), in comparison to the background distribution.

dx.doi.org/10.1021/ci400694q | J. Chem. Inf. Model. 2014, 54, 1093−1102

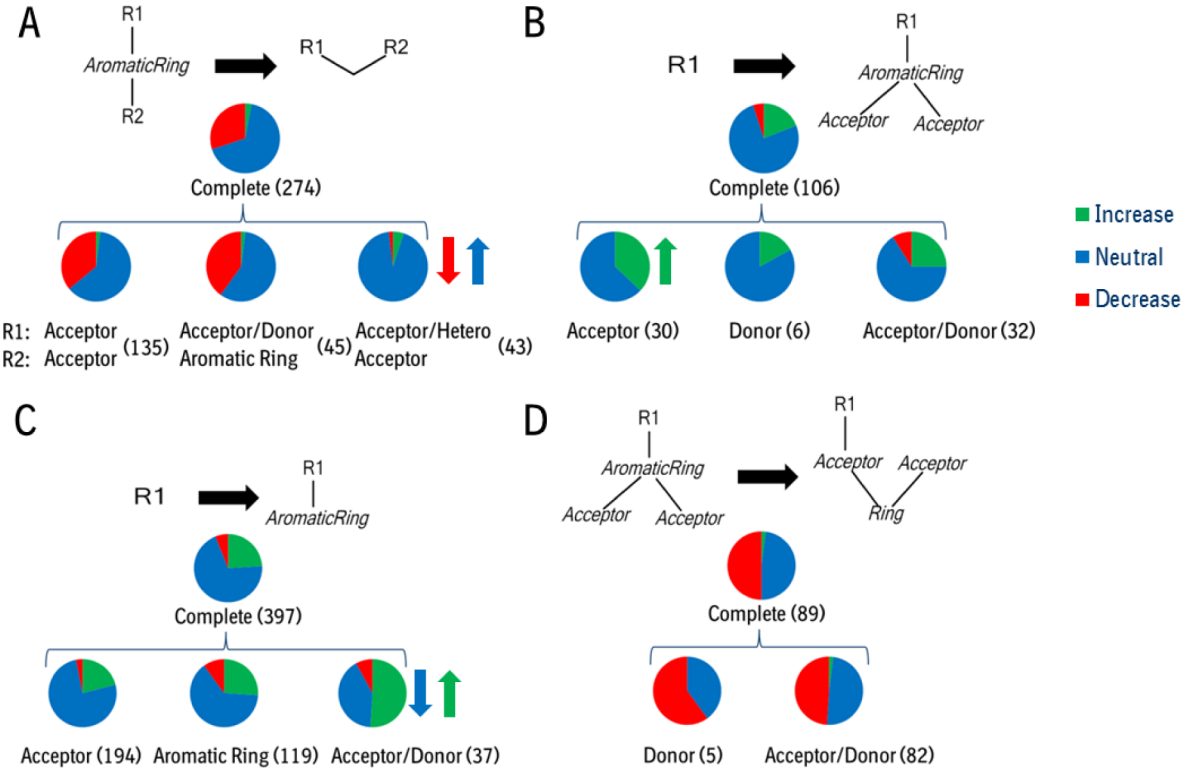


Figure 5. Overview about the environment influence on the inhibitory activity of selected VFMP transfers. Complete denotes the transfer distribution of the entire class, while the subclasses denote the transfer properties based on the first pharmacophore type at the attachment point. The number of underlying transfers is annotated in brackets. The red, blue, and green arrows indicate the decreasing (red), neutral (blue), and increasing (green) fraction of transfers, which is enriched (up) or reduced (down), in comparison to the transfer without taking the environment into account.

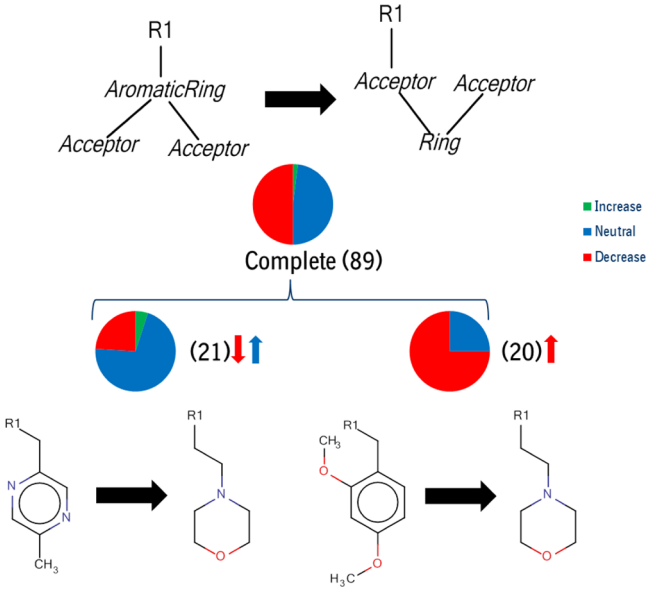


Figure 6. Underlying substructures for selected Value Fuzzy Matched Pairs (VFMP). The number of underlying transfers is annotated in brackets. The red, blue, and green arrows indicate the decreasing (red), neutral (blue), and increasing (green) fraction of transfers, which is enriched (up) or reduced (down), in comparison to the complete VFMP transfer.

matched molecular pairs are generated. In the first step of FMP, all molecules are processed using the pharmacophore retyping as described previously (see Figure 1). Based on the retyped molecular graphs, we use the method of Hussain and Rea to generate all possible fragments of size 1−30 with one, two, and three cuts. Based on the retyping, heteroatom-containing ring systems can be fragmented into a ring part and a heteroatom part. This feature of the FMP method is also highlighted in Figure 3. It opens the possibility to extract pharmacophoric features also if they are part of a ring system. The new FMP approach also reduces the number of possible matched pairs. This result is presented in Figure 7 and is based on the following observation.

The overall number of possible matched pairs within a dataset depends on the size of the alphabet, which is defined by the number of atom types that are observed within this dataset. We focus here only on the element type of an atom and do not include ring atom or aromatic atom types. There are seven commonly occurring atom types in typical medicinal chemistry datasets, namely, the heavy atoms C, N, S, O, F, Cl, and Br. The seven elements compare to four pharmacophore types within the FMP method, namely, the acceptor/donor, acceptor, donor and heteroatom type. Based on this observation, we can define two alphabets with size 7 and 4, respectively. If we calculate the number of possible fragments of size 1−*N*, using both alphabets, we observe already at length 5, a 10-fold higher number of MMP fragments, in comparison to FMP fragments. At length 9, the MMP fragmentation is already 100-fold larger (Figure 7). Since we are interested in the number of hypothetical transfers, we also calculated the number of possible transfers using both alphabets. The number of MMP transfers is already at length 3, 10-fold higher than the possible number of FMP transfers, and at length 5, both differ by 100-

dx.doi.org/10.1021/ci400694q | J. Chem. Inf. Model. 2014, 54, 1093−1102

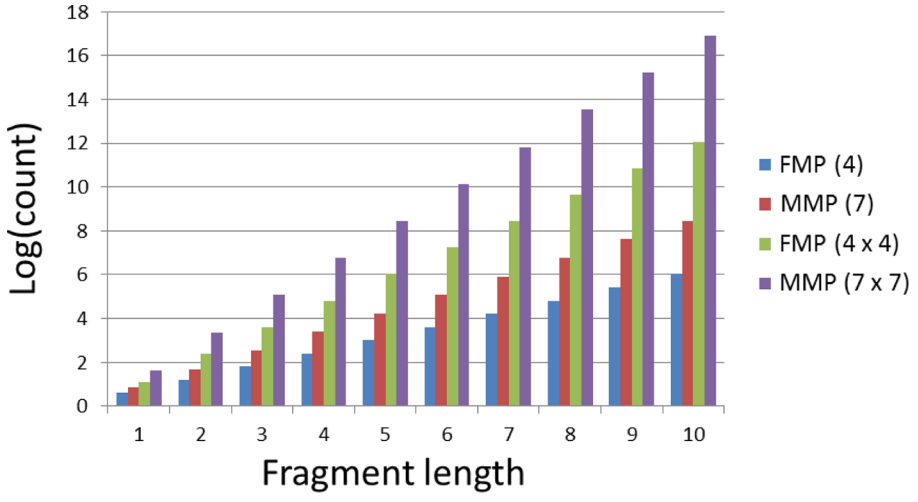
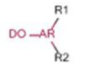
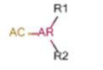


Figure 7. Number of possible fragments (4, 7) and fragment transfers (4 × 4, 7 × 7) based on the two alphabets: Fuzzy Matched Pairs (FMP) and Matched Molecular Pairs (MMP).

Table 3. Frequently Occurring (A) Bioisosteric (>90% Neutral) and (B) Nonbioisosteric (>30% Decrease or >30% Increase) FMP Transformations[a]

A

	PRE	POST	COUNT	Decrease	Neutral	Increase
1		R1—AR—Het	87	↓ 1.1	↑ 95.4	3.4
2	AC—HR—AC—R1	AD—R1	51	5.9	↑ 92.2	↓ 2.0
3		R2—AR—AC—R1	45	↓ 2.2	↑ 93.3	4.4
4	AR—R1		36	↓ 2.8	↑ 94.4	2.8
5		R1—AR—R2	36	↓ 2.8	↑ 94.4	2.8
6			35	8.6	↑ 91.4	↓ 0.0

B

	PRE	POST	COUNT	Decrease	Neutral	Increase
1	A	R1—AR—Het	245	↓ 7.8	↓ 59.2	↑ 33.1
2	AR—R1	DO—R1	136	↑ 42.6	↓ 52.2	5.1
3	R1—AR—Het	AC—R1	108	↑ 36.1	↓ 57.4	6.5
4	AR—R1	R1—AC—DO	100	↑ 52.0	↓ 48.0	↓ 0.0
5	A	Het—AR—AC—R1	94	↓ 7.4	↓ 48.9	↑ 43.6
6	AC—R1	Het—AR—AC—R1	93	↓ 3.2	↓ 59.1	↑ 37.6

[a]Transfers are indicated by the start substructure (PRE) and transferred substructure (POST). The number of occurrences (COUNT), as well as the percentage of transfers that reduce activity (Decrease), increase potency (Increase), and have no influence (Neutral) is indicated. The table is sorted by frequency in decreasing order. Legend: AC, acceptor; AD, acceptor/donor; DO, donor; AR, aromatic ring; HR, aliphatic ring; and Het, heteroatom. Bold black arrows below the Decrease, Neutral, and Increase columns indicate that this group is enriched (up) or reduced (down), in comparison to the background distribution.

fold. This observation underlines that the probability of observing multiple times the same transfers is much higher for the FMP approach in comparison to the MMP approach because the FMP method groups the transfers together. The results for the FMP method applied to the CYP2C19 isoform activity are presented in Table 3.

As already observed in Table 2, it can be seen that changing the number of heteroatoms at an aromatic ring has no influence on the activity (Table 3A, row 1). The removal of a donor feature next to an aromatic ring also has no influence (Table 3A, row 5). All six transfers with a high number of occurrences contain transfers between an aromatic system and a nonaromatic type (see Table 3B). They are, in all cases, decreasing the CYP2C19 activity. Again, the FMP transfers can be further subdivided using the underlying environment, as shown in Figure 8. It is worth mentioning that the list of all FMP transfers presented in Table 3 are highly relevant transfers because they are outside their respective background distribution.

In addition to the statistical analysis of the transfers in Table 3, we also want to highlight that the overall number of statistical relevant pairs is increased by using the FMP approach. The number of all transfers determined using the MMP and FMP methods respectively differs by 10% for transfers that occur more than 10 times. For MMPs and FMPs with a high number of occurrence (more than 80), 90% are outside the 95% CI of the background distribution. However, the FMP approach results in three times as many relevant transfers as the MMP approach. This underlines that we can increase the number of relevant transfers using the FMP approach.

Figure 8 shows a direct comparison of aromatic ring systems with nonaromatic ring systems. It is obvious that the largest reduction of activity is observed for the transfer of aromatic ring systems with two acceptors into nonaromatic ring systems

dx.doi.org/10.1021/ci400694q | J. Chem. Inf. Model. 2014, 54, 1093−1102

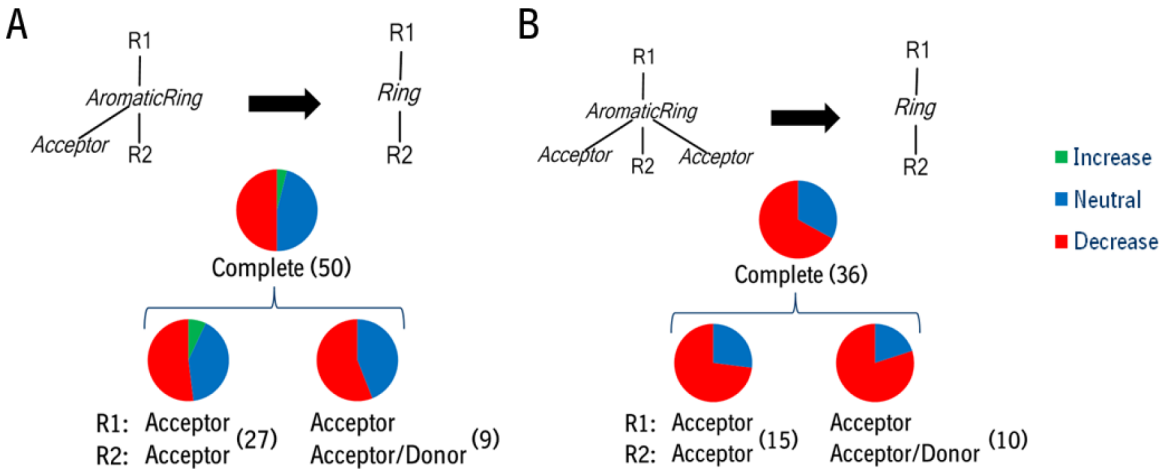


Figure 8. Overview about the environment influence on the CYP2C19 inhibitory activity of selected FMP transfers. "Complete" denotes the transfer distribution of the entire class, while the subclasses denote the transfer properties based on the pharmacophore type at the attachment point. The number of underlying transfers is annotated in brackets. The environment-based transfers are within the 95% CI of the complete group.

without an acceptor (67%). This feature is stable within all observed environments (Figure 8B). Exchange of an aromatic ring system with one acceptor to a nonaromatic ring system is not as pronounced (50%) but is still a transfer with high influence on activity in all environments (Figure 8A). This is supported by a study on extraction of possible pharmacophoric features for CYP2C19 inhibition. This analysis also found that two acceptors in an aromatic environment are an essential feature for CYP2C19 inhibition.[34] The analysis of the dataset showed that aromatic ring systems have the largest influence on CYP2C19 inhibition.

Looking into the details of the underlying environments and the transfers that have most pronounced impact, we can see that the introduction of an aromatic ring system is positive but this effect is, to a large extent, related to the introduction of the aromatic ring system next to an acceptor/donor. As a second result, the analysis revealed that the exchange of an aromatic ring with a nonaromatic ring system is highly unfavorable regarding CYP2C19 interaction. It suggests that the aromatic ring has a positive influence by itself, as we have seen in the classical MMP analysis. The positive influence is further increased if you compare molecules of similar size with an aromatic or nonaromatic ring at the same area.

In the next paragraph, we will analyze if the FMP method can identify transfers that give a selectivity handle between different end points. The ChEMBL dataset contains measurements for CYP2C19 as well as three other isoforms (namely, CYP3A4, CYP2C9, and CYP2D6).[27] Based on the FMP approach, we generated the possible transfers for all isoforms and determined the overlap between CYP2C19 and the other isoforms. CYP2C19 and CYP2C9 share 73 overlaps, followed by CYP2D6 and CYP2C19 29 overlaps and CYP2C19 and CYP3A4 with 63 overlaps. In Figure 9, all overlapping transfers of all four isoforms are shown. It can be seen that the transfer "aromatic ring with two acceptors" to "nonaromatic ring with one acceptor" reduces the activity in the case of CYP2C19. It is also reducing activity in CYP2C9 and CYP3A4, while it increases activity in CYP2D6. The comparison between the different isoform FMP shows that they are able to identify pharmacophores, which are responsible for selectivity. This transfer can be seen as a selectivity switch between the different isoforms and was also determined by the study of Ritchie et al.[32] Using the classical MMP approach, this activity switch is not detectable.

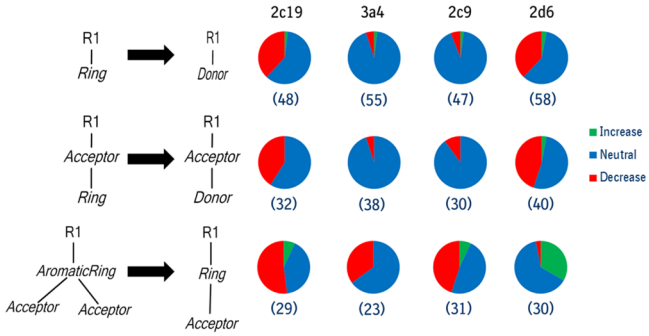


Figure 9. Schematic representation of FMP transfers and their multiparameter optimization potential. The number of underlying transfers is given in brackets.

Finally we analyzed the FMP overlap between a Boehringer-Ingelheim in-house data set for CYP2C19, containing 4000 compounds, and the ChEMBL dataset. The results are presented in Table 4. All the overlapping transfers of both datasets point in the same direction. In both neutral exchanges the number of heteroatoms at aromatic ring systems is reduced. A reduction of the activity by replacing an aromatic ring and one acceptor by an aliphatic ring system can also be found in both datasets. We undertook the same comparison using the classic MMP approach. Using this method, no overlap was observable. This result underlines the hypotheses that the FMP approach presents a useful extension to the classic MMP approach and how the novel FMP approach can be used to extract useful transformations from external data that are relevant for the in-house research process.

CONCLUSION

In this work, we have presented the new fuzzy matched pairs (FMP) method, which combines the classical Matched Molecular Pairs (MMP) approach with a pharmacophore description. The method was applied to different application scenarios and the results were compared to other MMP methods. The FMP approach turns out to be very useful for

Table 4. Overlap between Transfers within the ChEMBL Dataset and the In-House Dataset[a]

PRE	POST	ChEMBL COUNT	ChEMBL Decrease	ChEMBL Neutral	ChEMBL Increase	INHOUSE COUNT	INHOUSE Decrease	INHOUSE Neutral	INHOUSE Increase
	R1—AR—Het	87	1.1	95.4	3.4	21	4.8	85.7	9.5
	R2—AR—AC—R1	45	2.2	93.3	4.4	37	0.0	89.2	10.8
	R2—HR—R1	50	50.0	46.0	4.0	20	20.0	75.0	5.0

[a]Legend: AC, acceptor; AD, acceptor/donor; DO, donor; AR, aromatic ring; HR, aliphatic ring; and Het, heteroatom.

ligand-based design, especially during the lead optimization phase.

The main feature of the method is that the possible transfers are clustered together in groups of pharmacophore transfers. We could also highlight that, using the FMP method, the number of relevant transfers can be increased. This enables the scientist to reduce the analysis to the most relevant pharmacophore transfers. By analyzing the underlying structures, she or he can get possible alternatives with the same pharmacophoric properties. It is even possible to suggest unusual substructures that conform to the group properties. This underlines that the clustering of MMP based on the pharmacophore description enables one to suggest alternatives for a query substructure that would not be found by classical MMP. As a second feature, it seems possible to determine pharmacophores that are necessary to modulate activity of a target. We were able to show that the method can extract selectivity information for different target end points analyzed in parallel. We also want to highlight that the method enables one to extract useful transformations of external data, which are also relevant for in-house drug discovery and lead optimization. To us, this is an additional and very important aspect of the FMP approach, because the extraction of relevant exchanges from external datasets, which can be used for internal optimization, is often not possible with the classical MMP approach. Currently, we apply the FMP method to the evaluation of HTS hit series, as well as various ADMET properties. In addition, we are testing alternative "pharmacophore" descriptions of molecules to generate new types of FMP.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Tim.Geppert@boehringer-ingelheim.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Peter Haebel, Nils Weskamp, Bernd Wellenzohn, and Michael Bieler for helpful discussions. Valuable feedback on the manuscript from Jan Kriegl is gratefully acknowledged. Furthermore, we would like to thank our colleagues in information systems for technical support and our colleagues in medicinal chemistry for useful feedback.

REFERENCES

(1) Patani, G. A.; LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96* (8), 3147−3176.

(2) Wassermann, A. M.; Bajorath, J. Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. *Future Med. Chem.* **2011**, *3* (4), 425−436.

(3) 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* (5), 1138−1145.

(4) Hansch, C.; Hoekman, D.; Leo, A.; Zhang, L.; Li, P. The expanding role of quantitative structure-activity relationships (QSAR) in toxicology. *Toxicol. Lett.* **1995**, *79* (1−3), 45−53.

(5) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 868−873.

(6) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96* (3), 1027−1044.

(7) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38* (19), 2894−2896.

(8) Lower, M.; Geppert, T.; Schneider, P.; Hoy, B.; Wessler, S.; Schneider, G. Inhibitors of Helicobacter pylori protease HtrA found by "virtual ligand" screening combat bacterial invasion of epithelia. *PLoS. One* **2011**, *6* (3), No. e17986.

(9) Gussregen, S.; Matter, H.; Hessler, G.; Muller, M.; Schmidt, F.; Clark, T. 3D-QSAR based on quantum-chemical molecular fields: Toward an improved description of halogen interactions. *J. Chem. Inf. Model.* **2012**, *52* (9), 2441−2453.

(10) Wassermann, A. M.; Dimova, D.; Iyer, P.; Bajorath, J. Advances in Computational Medicinal Chemistry: Matched Molecular Pair Analysis. *Drug Dev. Res.* **2012**, *73* (8), 518−527.

(11) Zhang, B.; Wassermann, A. M.; Vogt, M.; Bajorath, J. Systematic assessment of compound series with SAR transfer potential. *J. Chem. Inf. Model.* **2012**, *52* (12), 3138−3143.

(12) Hussain, J.; Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **2010**, *50* (3), 339−348.

(13) Gleeson, P.; Bravi, G.; Modi, S.; Lowe, D. ADMET rules of thumb II: A comparison of the effects of common substituents on a range of ADMET parameters. *Bioorg. Med. Chem.* **2009**, *17* (16), 5906−5919.

(14) 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**, *49* (23), 6672−6682.

(15) Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched molecular pair analysis in drug discovery. *Drug Discovery Today* **2013**, *18* (15−16), 724−731.

(16) Haubertin, D. Y.; Bruneau, P. A database of historically-observed chemical replacements. *J. Chem. Inf. Model.* **2007**, *47* (4), 1294−1302.

(17) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. SwissBioisostere: A database of molecular replacements for ligand design. *Nucleic Acids Res.* **2013**, *41*, D1137−D1143 (Database issue).

(18) Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today* **2006**, *11* (15−16), 700−707.
(19) Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I. I.; Solov'ev, V. P. Exhaustive QSPR studies of a large diverse set of ionic liquids: How accurately can we predict melting points? *J. Chem. Inf. Model.* **2007**, *47* (3), 1111−1122.
(20) 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* (10), 1872−1886.
(21) Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. WizePairZ: A novel algorithm to identify, encode, and exploit matched molecular pairs with unspecified cores in medicinal chemistry. *J. Chem. Inf. Model.* **2010**, *50* (8), 1350−1357.
(22) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* **2011**, *54* (22), 7739−7750.
(23) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9* (4), 273−276.
(24) *OEChem TKv2012.Feb*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2012.
(25) Landrum, G. *RDKit: Open-source cheminformatics*, http://www.rdkit.org.
(26) 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 (Database issue).
(27) Veith, H.; Southall, N.; Huang, R.; James, T.; Fayne, D.; Artemenko, N.; Shen, M.; Inglese, J.; Austin, C. P.; Lloyd, D. G.; Auld, D. S. Comprehensive characterization of Cytochrome P450 isozyme selectivity across chemical libraries. *Nat. Biotechnol.* **2009**, *27* (11), 1050−1055.
(28) *JChem 5.11.5*, ChemAxon, http://www.chemaxon.com, 2013.
(29) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, 1993.
(30) R Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing (http://www.R-project.org), 2013.
(31) Konstanz Information Miner. *KNIME 2.7.2*, KNIME.com GmbH, 2011.
(32) Ritchie, T. J.; Macdonald, S. J.; Young, R. J.; Pickett, S. D. The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and -aliphatic ring types. *Drug Discovery Today* **2011**, *16* (3−4), 164−171.
(33) Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Huang, R. Predictive models for Cytochrome p450 isozymes based on quantitative high throughput screening data. *J. Chem. Inf. Model.* **2011**, *51* (10), 2474−2481.
(34) de Groot, M. J.; Ekins, S. Pharmacophore modeling of Cytochromes P450. *Adv. Drug Delivery Rev.* **2002**, *54* (3), 367−383.