

Methods for the Analysis of Matched Molecular Pairs and Chemical Space Representations

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

ANTONIO DE LA VEGA DE LEÓN

aus Madrid, Spanien

Bonn, 2016

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath
 2. Gutachter: Jun.-Prof. Dr. rer. nat. Thomas Schultz
- Tag der Promotion: 27. September 2016
Erscheinungsjahr: 2016

*Para mis padres,
por todo el amor y el apoyo que siempre me han dado*

Abstract

Compound optimization is a complex process where different properties are optimized to increase the biological activity and therapeutic effects of a molecule. Frequently, the structure of molecules is modified in order to improve their property values. Therefore, computational analysis of the effects of structure modifications on property values is of great importance for the drug discovery process. It is also essential to analyze chemical space, i.e., the set of all chemically feasible molecules, in order to find subsets of molecules that display favorable property values. This thesis aims to expand the computational repertoire to analyze the effect of structure alterations and visualize chemical space.

Matched molecular pairs are defined as pairs of compounds that share a large common substructure and only differ by a small chemical transformation. They have been frequently used to study property changes caused by structure modifications. These analyses are expanded in this thesis by studying the effect of chemical transformations on the ionization state and ligand efficiency, both measures of great importance in drug design. Additionally, novel matched molecular pairs based on retrosynthetic rules are developed to increase their utility for prospective use of chemical transformations in compound optimization. Further, new methods based on matched molecular pairs are described to obtain preliminary SAR information of screening hit compounds and predict the potency change caused by a chemical transformation.

Visualizations of chemical space are introduced to aid compound optimization efforts. First, principal component plots are used to rationalize a matched molecular pair based multi-objective compound optimization procedure. Then, star coordinate and parallel coordinate plots are introduced to analyze drug-like subspaces, where compounds with favorable property values can be found. Finally, a novel network-based visualization of high-dimensional property space is developed. Concluding, the applications developed in this thesis expand the methodological spectrum of computer-aided compound optimization.

Acknowledgements

This thesis is the culmination of many years of work and I want to personally thank here those that helped me along the way.

To my parents, thank you for always pushing me to become the best at what I wanted to be. Without your continued support throughout the years, this thesis would not have been possible. Thank you for always being by my side.

To my supervisor, Prof. Dr. Jürgen Bajorath, thank you very much for all that you have done for me. Your teachings, first as my masters professor and then as my PhD supervisor, have instilled in me a passion for chemoinformatics that I will keep with me going forward. From you I have learned how to become a better scientist and a better person.

To Prof. Dr. Thomas Schultz, thank you for your help and insight in the study where we collaborated and for acceding to be the second reviewer of this thesis. To Prof. Dr. Andreas Weber and Prof. Dr. Evi Kostenis, thank you for being in my defense committee.

To the LSI group, a heartfelt thank you. Together, you made my stay as a PhD student one I will not forget. To Norbert Furtmann, thank you for your companionship throughout the years, for the friendship we have shared, and for everything I learned from you. To Kathrin Heikamp, thank you for the good times we shared, and for your support in my procrastination. To Jenny Balfer, thank you for all you taught me (both the science and the german customs), and for your cheerful spirit. Thank you all three for the insightful comments and useful feedback that improved this thesis. To Dagmar Stumpfe, thank you for your very special way to motivate and encourage me. To Dilyana Dimova and Shilva Kayastha, thank you for the work, the fun, and the sweat we shared. To everyone else, thank you for making my stay so memorable.

Contents

1	Introduction	1
2	Target-based analysis of ionization states of bioactive compounds	23
3	Formation of activity cliffs is accompanied by systematic increases in ligand efficiency from lowly to highly potent compounds	29
4	Matched molecular pairs derived by retrosynthetic fragmentation	41
5	Systematic identification of matching molecular series and mapping of screening hits	49
6	Prediction of compound potency changes in matched molecular pairs using support vector regression	61
7	Compound optimization through data set-dependent chemical transformations	75
8	Visualization of multi-property landscapes for compound selection and optimization	89
9	Chemical space visualization: transforming multi-dimensional chemical spaces into similarity-based molecular networks	105
10	Conclusion	117
	Bibliography	119

1 Introduction

1.1 Drug discovery

Drug discovery is the process of identifying small molecules that treat a specific disease and developing them to market approval. The entire process generally takes 10 to 15 years¹ and costs billions of dollars². It can be separated into several stages that will be described below and are shown schematically in Figure 1.³

The first stage involves finding a target, usually a protein, whose inhibition (or activation) treats the symptoms or cures the disease.⁴ An assay is developed in order to easily test the activity of compounds in a high-throughput fashion. Although target-based drug discovery has been very popular in the era of molecular sciences, alternative approaches like phenotypic assays are used when the target is unknown.⁵ In this case, an assay is prepared whose readout is the alteration of molecular markers related to the disease. The molecular basis of the effect of the compound may not be known but can be neglected as long as the assay is a good representation of the biology associated with the disease condition.

Once a reliable assay has been developed, a large screen of small molecules is carried out.⁶ Active compounds found in the assays are called “hits”. These hits are small molecules for which activity against the desired target has been confirmed.⁷ The structure of these compounds is optimized in design-make-test cycles where new alterations are tested to determine their effect on molecular properties. These modifications normally introduce chemical groups that create energetically favorable contacts with the target, while reducing the possibility of binding to undesired targets. At the same time, the absorption, distribution, metabolism,

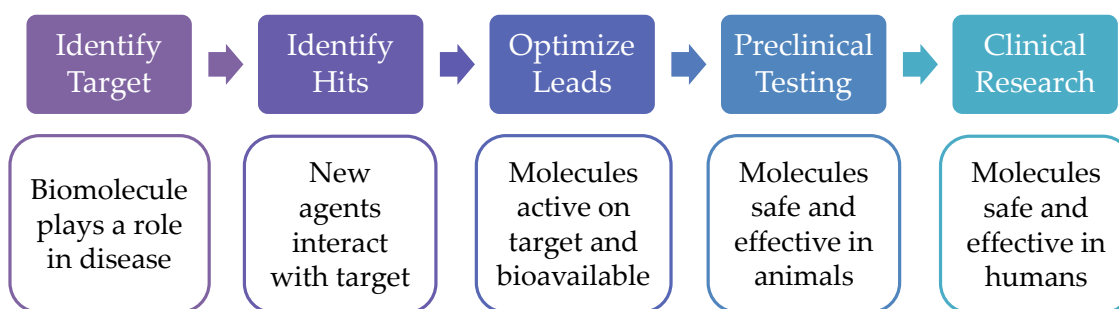


Figure 1: Drug discovery process. A schematic representation of the drug discovery process is shown. The figure has been adapted from [3].

excretion, and toxicity (ADMET) of the molecules *in vivo* is modeled through biochemical assays or computational prediction algorithms.⁸ If the ADMET properties are not managed correctly, the molecules will not be able to reach the target (i.e., they will not be bioavailable) and will not show pharmacological effect.⁹ They could also develop toxic properties causing severe side effects. Therefore, compound optimization is a multi-objective process where activity, bioavailability, toxicity, and chemical synthesizability are improved together through structure modifications.⁷ This process generates “lead” compounds that can be optimized further to obtain drug candidates.

Preclinical studies begin once a candidate molecule is found that is safe, active, and bioavailable. The compound is tested *in vivo* or *in vitro* to obtain an approximation of dosage for “first-in-man” studies.¹⁰ The information obtained in this step will serve to design clinical trials, the final step in the drug discovery process. In clinical trials the compound will be tested in humans against a control treatment or a placebo.¹¹ The results from these trials should provide evidence of the safety and effectiveness of the candidate molecule. They will be submitted to regulatory agencies as a prerequisite to bring the drug to the market.

The drug discovery process is a long and costly procedure. The cost of developing new drugs has increased exponentially for the past decades and the number of new drugs introduced each year has remained constant.¹ In order to increase the efficiency of drug discovery, chemoinformatic approaches are used to assist at different stages.¹² Chemoinformatics has been defined as “*the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the arena of drug lead identification and optimization*”.¹³ A central theme of this discipline is under-

standing how structure modifications affect the properties of molecules. With this information, compound optimization can be more focused, resulting in a faster and more efficient process. In the next chapter, the analysis of the relation between structure and property is introduced.

1.2 Structure-property relationship

There are many properties that need to be considered during compound optimization. Activity is the most important one to take into account because without it the disease condition is not treated. However, activity alone is not sufficient to develop a drug and many other properties need to be examined. A common example is lipophilicity, i.e., the tendency of a compound to dissolve in a nonpolar medium like octanol.¹⁴ If a compound is too lipophilic, it will become trapped in the cellular bilayer. If it is not lipophilic enough, it will not cross the membrane of cells and not be absorbed from the gastrointestinal tract. Therefore, for lipophilicity and many other properties some value ranges are considered favorable and others unfavorable. Another important property is the ionization state of a molecule, i.e., the presence or absence of ionized chemical groups.¹⁵ The presence of charged groups affects many other ADMET properties such as bioavailability or lipophilicity.

Activity has been the main focus of chemoinformatics analyses because of its importance in drug discovery. Structure-activity relationship (SAR) analysis aims to determine how alterations in the structure of molecules affect their binding properties to a specific target.¹⁶ SAR analysis is frequently carried out on structurally related compounds such as those forming analog series. If small modifications in the molecular structure result in small differences in activity, the set of compounds has continuous SAR character (Figure 2 top). By contrast, when small structural changes lead to large potency differences, the SAR character is discontinuous (Figure 2 bottom).

Discontinuous SAR often result in the presence of activity cliffs. Activity cliffs are pairs of compounds that are structurally similar but have a large potency difference.¹⁷ They have been extensively studied and characterized in various ways.¹⁸ Alternative molecular representations and similarity measures (*vide infra*) may create large variability in the distribution of activity cliffs for the same set of compounds.¹⁹ The large potency difference between activity cliff partners has also been rationalized on the basis of protein-ligand three-dimensional crystal

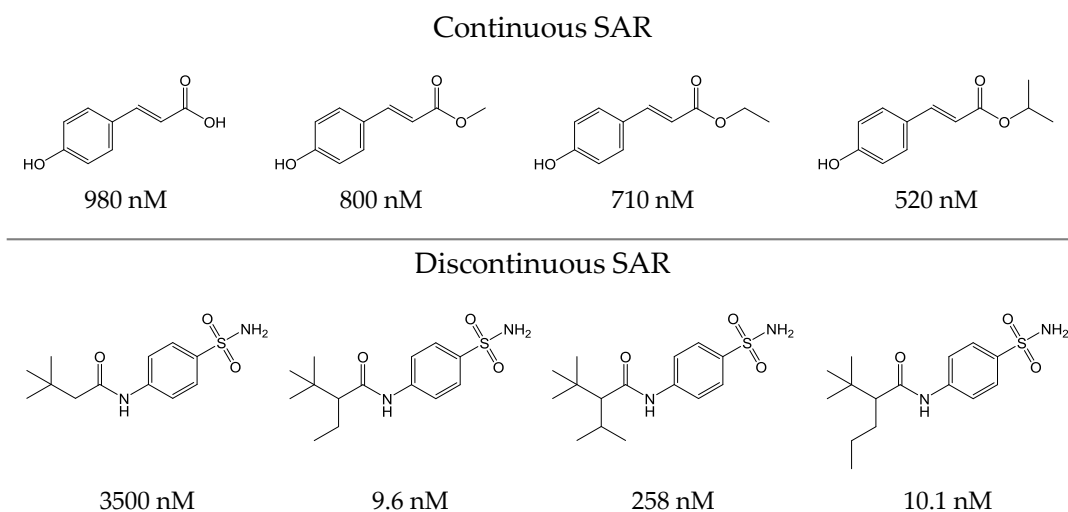


Figure 2: SAR character. Two analog series are shown that display continuous (top) or discontinuous (bottom) SAR character. Below each molecule its inhibition constant value against carbonic anhydrase II is given.

structures.²⁰ Therefore, activity cliffs highlight chemical groups critical for activity against a specific target.

Although the previous paragraphs have focused on activity, the relation of different properties to structure changes is analyzed in a similar manner. In this way, SAR analysis is expanded to structure-property relationship (SPR) analysis. In order to analyze SPRs of large molecule data sets, computational methods are applied. These methods require molecules to be in a computer-accessible format. Relations between compounds can only be studied with formally defined similarity measures. In the next section, different molecular representations and similarity measures are described. Afterwards, computational methods that analyze SPRs in order to aid compound optimization efforts are reported.

1.3 Molecular representations and similarity measures

Appropriate representations are required to handle molecules in computer code. Molecular structures are commonly modeled as annotated graphs where vertices are the atoms and edges are the bonds of the molecule.²¹ For each atom,

characteristics such as charge, hybridization state, or stereochemistry are saved. Additionally, two- or three-dimensional atom coordinates are stored if only the topology of the molecule or a specific conformation is modeled, respectively. Likewise, bond information such as order or stereochemistry is saved. This data can be encoded into connectivity table file types, such as the MOL file type.²² In addition, linear notations such as SMILES²³ and InChI²⁴ can encode the two-dimensional structure as a string.

Linear notations are very useful for fast identity comparisons of molecules. Still, if the molecules are not identical, similarity measures need to compare the two molecular graphs. This task is often performed through maximum common substructure (MCS) computation, whose time complexity is known to be NP-complete. Therefore, it is a computationally demanding calculation.²⁵ Similarity measures are important because many cheminformatics algorithms rely on the “similarity property principle”.²⁶ This principle states that two molecules that are similar should have similar properties. Accordingly, the systematic extraction of pairwise similarity relationships is crucial for many cheminformatics analyses.

1.3.1 Molecular descriptors

Molecular descriptors are mathematical models that describe the structure or the properties of molecules. They represent various characteristics of the molecule such as its topology, lipophilicity, size, or charge.²⁷ Some simple descriptors, such as atom counts and molecular weight, can be directly calculated from the molecular formula or from linear notations. These simple descriptors are frequently called 1D descriptors, because they do not require atom connectivity information. Those descriptors that require the molecular structure, such as topological indices, are considered 2D descriptors. Finally, some descriptors, such as the dipole moment, require a pre-specified three-dimensional conformation of the molecule to be computed and are called 3D descriptors. A molecule can be represented by a set of different descriptors encoded as a vector of numeric values (Figure 3).

Descriptor value vectors can be interpreted as positions in high-dimensional space. In this property space each dimension corresponds to one descriptor and molecules distribute based on their descriptor values. These high-dimensional spaces are used to represent and study chemical space. Therefore, distance in property space accounts for similarity between molecules. Any L_p -norm distance (also called Minkowski distance) can be used, but the most common distances employed are the Euclidean ($p=2$) and Manhattan (also called Hamming) distance

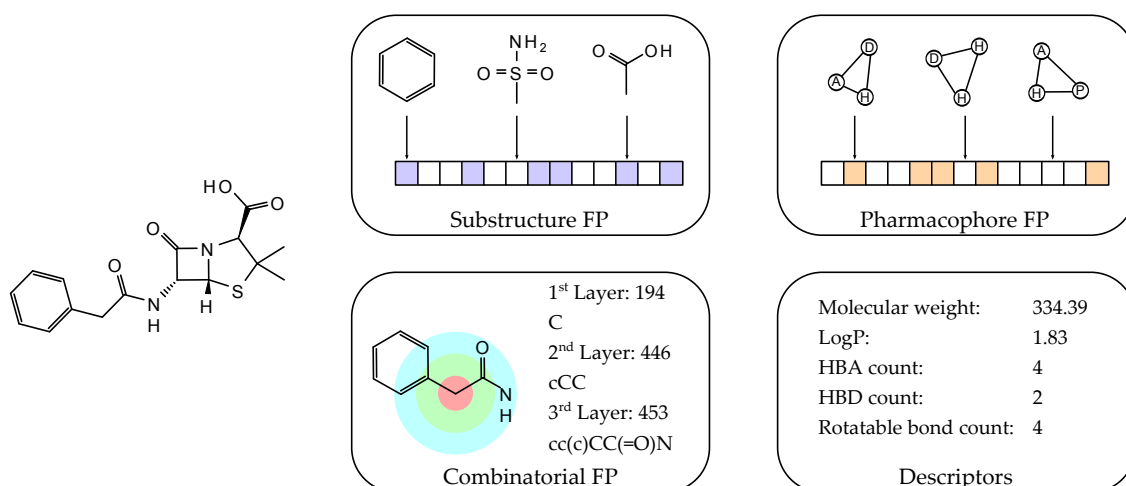


Figure 3: Molecular descriptors and fingerprints. Penicillin G and the computation of fingerprints and descriptors are depicted schematically. For substructure and pharmacophore fingerprints (FPs) the bit string is represented as a set of cells. Filled cells denote chemical patterns present in the molecule. Unfilled cells represent absent patterns. For the combinatorial FP one atom is taken as an example and all substructures rooted on this atom with maximum radius of 2 are depicted with concentric layers. For each layer, the substructure is given as a SMILES string and the hash value is reported. Finally, different molecular descriptors are given.

($p=1$).²⁸ For two molecules, x and y , represented by n descriptor values, the Minkowski distance follows the equation:

$$\text{distance}(x, y, p) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

where x_i and y_i are the values of molecules x and y for descriptor i .

1.3.2 Fingerprints

Fingerprints represent a molecule with a set of Boolean values encoded as a bit string. Each position in the bit string represents one chemical pattern and if this pattern is present in the molecule the corresponding position will be set on. There are different kinds of fingerprints depending on the types of chemical

patterns they use. Substructure fingerprints use pre-specified molecular fragments as patterns (Figure 3). One example is the Molecular ACCess System (MACCS) keys consisting of a set of 166 substructures.²⁹ Pharmacophore fingerprints define combinations of two, three, or four pharmacophore centers with different distances between their members (Figure 3).³⁰ These pharmacophore centers are abstractions of the molecular structure where different chemical groups that share common characteristics are considered together as the same pharmacophore type. Common pharmacophore types include hydrogen bond donors, hydrogen bond acceptors, and hydrophobic centers. Combinatorial fingerprints, in contrast to the previous two, do not have a fixed size as they do not use pre-specified patterns (Figure 3). For each molecule, all possible subgraphs up to a specific size are enumerated and hashed into numbers. The set of numbers represent all possible substructures present in the molecule. A commonly used combinatorial fingerprint is the Extended Connectivity FingerPrint (ECFP).³¹

Similarity indices permit the calculation of chemical similarity based on bit strings. They compare the overlap in chemical patterns present in each molecule to obtain a similarity value. As with distances, there are many similarity indices that can be employed. However, the Tanimoto coefficient (also called Jaccard coefficient) is the most commonly used similarity index in chemoinformatics.³² It is a measure of the intersection divided by the union of the two pattern sets.²⁸ Therefore, it can be understood as the percentage of shared chemical patterns between two molecules. For a molecule x with a chemical patterns and a molecule y with b chemical patterns, c of which are also present on molecule x , the Tanimoto coefficient (Tc) is described as:

$$Tc(x, y) = \frac{c}{a + b - c}$$

1.3.3 Scaffolds

Molecular similarity can be quickly computed as distances on property space or as similarity indices based on binary bit strings. In comparison to graph-based similarity measures, these methods trade a one-time penalty (calculation of the fingerprint/descriptor vector) for a much faster similarity assessment. However, it is sometimes not intuitive why two molecules have high similarity with these methods when comparing their chemical structures.³³ There have been many attempts to provide similarity relationships based on chemical structures that are not too computationally demanding. One approach is based on scaffolds.

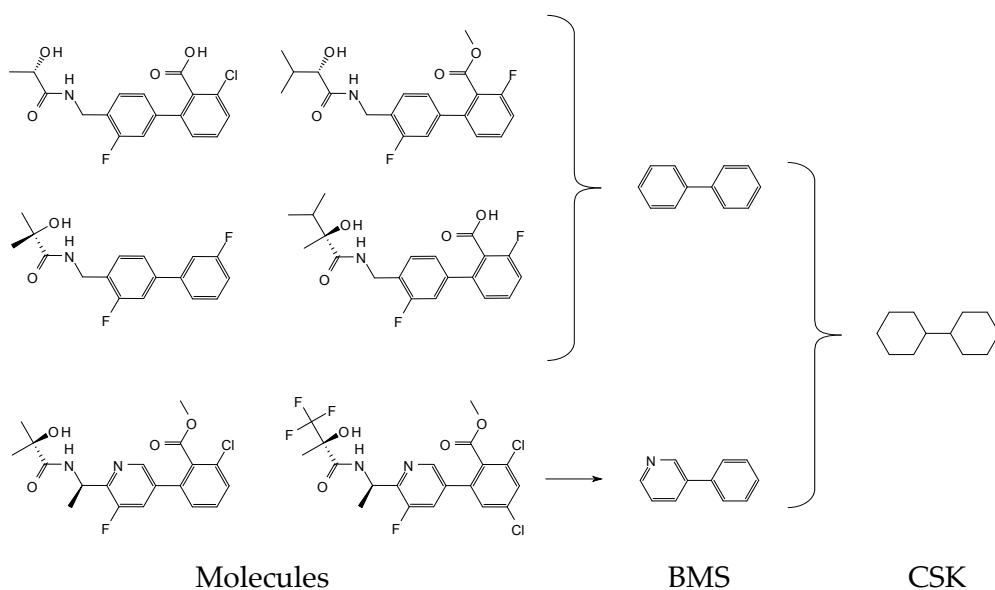


Figure 4: Scaffolds. The hierarchical organization of chemical structures is schematically represented. The six molecules shown yield two different Bemis-Murcko scaffolds (BMSs). These two BMSs both share the same cyclic skeleton (CSK).

Scaffolds represent chemical abstractions that describe the core structure of a molecule.³⁴ Two molecules that generate the same scaffold share a common substructure. Although the term “scaffold” has been used frequently in medicinal chemistry, authors apply it with different meanings and implementations.³⁵ However, many applications have chosen a consistent definition proposed by Bemis and Murcko.³⁶

Bemis and Murcko divided compounds into rings, linkers, and sidechains. Linkers are atom chains that do not belong to any ring, but connect two or more rings. Sidechains, by process of elimination, are terminal atom chains connected to a ring or a linker. According to the original definition, the combination of rings and linkers creates a “framework”. These frameworks have emerged as the basis for standard representations of scaffolds.

Two different scaffold representations can be derived from a framework. The first one maintains the chemical information of the atoms and bonds that form the framework. These are often called Bemis-Murcko scaffolds (BMSs). For the second, all atoms in the framework are transformed to carbons and all bonds are considered as single bonds.³⁷ In this manner, chemical information is abstracted but the connectivity information is preserved. These scaffolds are called cyclic

skeletons (CSKs). There is an increasing abstraction of chemical information going from molecules to BMSs to CSKs (Figure 4).

Although BMSs and CSKs are widely applied, they have some limitations. In drug discovery, small rings are typically among the different chemical groups used to optimize compounds. However, the addition of a ring modifies the framework and, therefore, these molecules produce different BMSs and CSKs.³⁸ This can make a data set appear more diverse. There have been several attempts to organize scaffold information on the basis of substructure relationships to identify sets of structurally similar scaffolds.^{39,40}

Two compounds that generate the same scaffold can be considered similar because they have a common substructure. However, the size of the shared substructure varies widely between different molecules. Additionally, molecules that only differ by a small ring would be considered different. Therefore, scaffolds have not frequently been used for pairwise similarity assessment but rather to cluster heterogeneous compound data sets according to their chemical structure.

1.3.4 Matched molecular pairs

Matched molecular pairs (MMPs) are pairs of compounds that have a large common substructure and differ only at one site.⁴¹ Similar to scaffolds, the two molecules that form an MMP share a common substructure. Nonetheless, MMPs can only have one site of variation while scaffolds can have many. The substructure that is shared between MMP partners is called the key fragment. The substructures that are different are called value fragments and together they define a chemical transformation.⁴²

After they were introduced, the calculation of MMPs was mainly done by MCS computation between all pairs of compounds in a data set.^{43,44} This made MMP generation very time consuming. An alternative method consisted of formally defining a set of chemical transformations that were matched against all compound pairs.^{45,46} MMPs were extracted from compounds matching any predefined substitution. Still, this method restricted the chemical variety of MMPs to study.

These methods were followed by a fragmentation-based algorithm published in 2010.⁴² This algorithm consists of two steps: molecule fragmentation and MMP generation. In the fragmentation, each molecule is split along single bonds that are

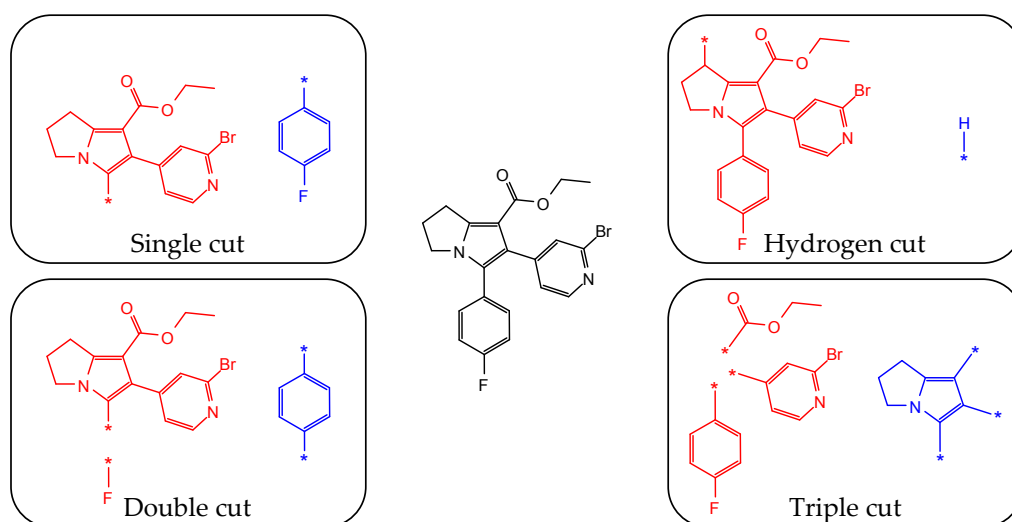


Figure 5: Fragmentation. The four types of cuts performed during fragmentation are schematically represented. The fragmented molecule is shown in the center. For each cut, the key fragment is colored red and the value fragment is colored blue.

not part of a ring. When a bond is broken, an attachment point is added to each fragment to keep connectivity information. Fragmentation can occur at one, two, or three bonds at the same time generating single, double, or triple cuts, respectively. In single cuts, the molecule is split into two fragments where the bigger one is considered the key fragment (Figure 5 top left). In double cuts, three fragments are generated. The two fragments with a single attachment point together constitute the key fragment (Figure 5 bottom left). In triple cuts, four fragments are created but only those cuts that generate three fragments with one attachment point and one with three attachment points are considered valid. Similar to double cuts, all fragments with a single attachment point are pooled together to generate the key fragment (Figure 5 bottom right). Hydrogens are not usually present as atoms in the molecule but as properties of other atoms. Therefore, the cuts described so far cannot identify chemical transformations involving hydrogen atoms, and hydrogen cuts are required to find those transformations. Hydrogen cuts are obtained by taking every key fragment from single cuts, substituting the attachment point for a hydrogen atom, and comparing the resulting structure to all data set molecules. If any molecule matches, a new fragmentation is generated where the key fragment is the same as in the single cut and the value fragment is a hydrogen atom (Figure 5 top right).

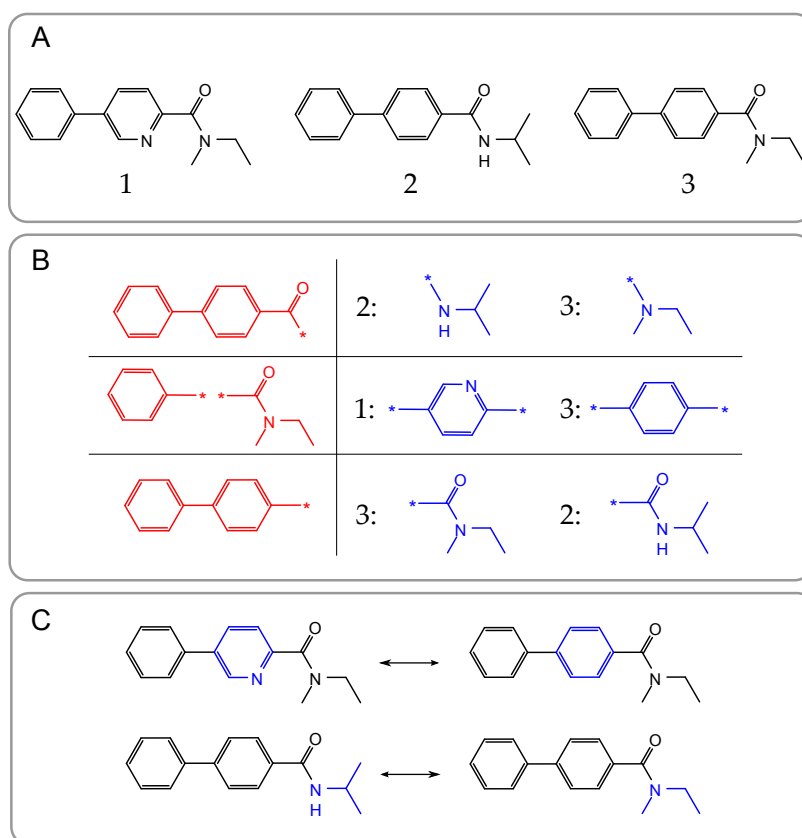


Figure 6: MMP generation. (A) Three exemplary compound structures with IDs are shown. (B) A small portion of the index table created after fragmentation of all compounds is displayed. On the left are the key fragments (colored red) and on the right are a list of value fragments (colored blue) associated to the molecule ID. (C) Two MMPs are shown. The exchanged substructure is colored in blue.

MMP generation begins after all molecules have been systematically fragmented. The resulting cuts are organized in an index table (Figure 6B). Key fragments are used as indices while the value fragments are grouped according to their key fragments. Once the fragmentation of all molecules is complete, the index table is used to obtain the MMPs. For each entry, all pairs of value fragments form MMPs. If for two molecules more than one MMP can be formed, the MMP that maximizes the size of the shared substructure is chosen (Figure 6C). Size restrictions should be applied to constrain the generation of MMPs to chemically relevant analogs. A set of previously described size restrictions⁴⁷ based on the heavy atom (HA) number of the fragments is adopted throughout this thesis:

(i) the number of HAs in key fragments should be at least twice the number of HAs in their corresponding value fragments; (ii) value fragments should not exceed 13 HAs; and (iii) the size difference between the two value fragments of an MMP should not be larger than 8 HAs.

MMPs represent an intuitive measure of chemical similarity. In contrast to similarity indices and descriptor distances, there is always a structural justification for each similarity relationship. However, it is a Boolean measure as two compounds either form an MMP or they do not. Similarity indices and descriptor distances, by contrast, can distinguish between closely related and distinct structures.

1.3.5 Extensions and applications of matched molecular pairs

The concept of MMPs has been heavily studied in recent years and several extensions have been developed. One extension is fuzzy MMPs, i.e., pairs defined not on molecules but on fuzzy representations thereof.⁴⁸ Fuzzy molecular graphs are typically pharmacophore-based simplifications that group different atoms into a single element. For example, a benzene ring would be simplified to a hydrophobic element.

MMPs have also been studied on the basis of protein-ligand crystal structures.⁴⁹ The conformation of a ligand present in the crystal was used as a template to superpose MMP partners. The effect of the chemical transformation on the affinity was related to the binding pocket environment. These MMPs were published on a public database called VAMMPIRE (Virtually Aligned Matched Molecular Pairs Including Receptor Environment). This database was later used to derive a predictive tool for lead optimization named VAMMPIRE-LORD.⁵⁰ A similar methodology has been reported to describe three-dimensional matched pairs (3DMPs).⁵¹ In this case, ligand conformations were generated using templates from analogous molecules in prealigned protein-ligand crystal structures. These conformations were fragmented and those pairs of molecules that shared a large common substructure that was near in three-dimensional space were considered 3DMPs. The analysis of 3DMPs and their activity change was applied to compound design.

Matching molecular series (MMS) are sets of three or more molecules that share a common substructure.⁵² This straightforward extension allows the study of several closely related structures at the same time. They can represent analog series for SPR analysis. MMS are obtained from the index table generated to

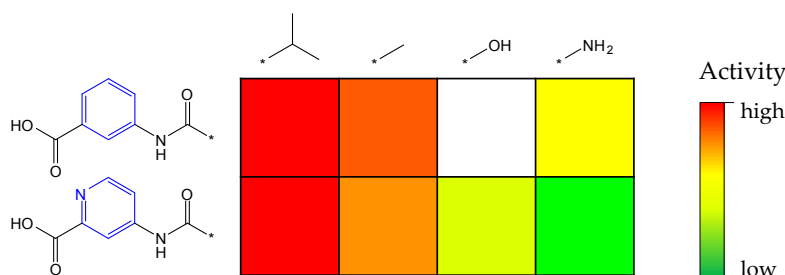


Figure 7: SAR Matrix. An exemplary SAR Matrix is shown. Colored cells represent known molecules and white cells denote virtual compounds. Each row corresponds to one MMS. The key fragments are shown left of the rows and the different value fragments are shown above the columns. The structure difference between the key fragments is colored blue. Activity is displayed by a color scheme going from red (high activity) through yellow to green (weak activity).

create MMPs. Each entry in the table that contains at least three different value fragments generates one series. MMS have been used to study SAR transfer.⁵³ In SAR transfer, the potency progression of two chemically related MMS that share several substitutions is analyzed. It not only allows the transfer of insights obtained for one chemical series to another, but can also provide novel compound suggestions. The SAR transfer methodology was extended to create a predictive algorithm that suggested possible chemical modifications to improve activity based on statistical analysis of large numbers of potency-ordered MMS.⁵⁴

Additionally, the concept of chemically related MMS was used to develop a novel visualization called the SAR Matrix (Figure 7).⁵⁵ The set of SAR Matrices of a data set organizes all structural relationships between compounds and enumerates virtual molecules that represent analogous compounds not present in the data set. Each compound is annotated with its activity. The SAR Matrix concept has been extended to create Free-Wilson-type models that can predict the activity of virtual compounds.⁵⁶ It has also been adapted to analyze screening data and predict novel hit compounds.⁵⁷

Activity cliffs (*vide supra*) have been analyzed on the basis of MMP relationships.⁴⁷ These are called MMP-cliffs and they have become the preferred activity cliff representation because of their intuitive nature. MMPs represent a conservative measure of similarity compared to fingerprint-based similarity indices.¹⁹ They have not only been studied as isolated compound pairs. If any compound forms more than one MMP-cliff to different partner molecules, networks of coordinated MMP-cliffs emerge and their topology has been analyzed.⁵⁸ A method was

later developed to prioritize clusters of coordinated MMP-cliffs and extract SAR information.⁵⁹

MMPs are frequently used to obtain chemical transformations from sets of compounds. In MMP analysis (MMPA), transformations are collected by grouping MMPs with identical structure modifications and their effect on different molecular properties is analyzed.⁶⁰ In some cases, these transformations conserve the values of different chemical properties such as activity and are considered bioisosteric. In others, small modifications consistently cause large changes in property values. Bioisosteric replacements have been intensively studied on the basis of MMPs⁶¹, as well as chemical transformations that are frequently found between activity cliff partners⁶². Recently, a study expanded on this idea analyzing the effect of transformations on both potency and ADMET descriptors.⁶³

Fuzzy MMPs are useful for MMPA on small compound data sets. By abstracting chemical information, each different transformation will be more frequent. It has been shown that a large number of MMPs per transformation is needed to statistically characterize a large change in activity from heterogeneous data sources (such as publicly available databases).⁶⁴ Therefore, the effect of chemical transformations can be determined with greater statistical rigor.

1.4 Compound optimization

Compound optimization is a multi-objective procedure that takes many properties such as potency, selectivity, bioavailability, and toxicity into account. These properties generate a high-dimensional space where compounds distribute based on their property values. The high-dimensional property space is used to represent and study the chemical space, the set of all chemically feasible molecules. Because of the large number of possible molecules (estimated to be 10^{33})⁶⁵, the chemical space is very large and a systematic exploration is complicated. However, drug compounds are thought to be present in specific drug-like subspaces rather than distributed over the whole space. Therefore, the search of characteristic property value combinations that better determine drug-like character is a central part of compound optimization. Additionally, the structure of the molecule determines its property values and modifications of the structure affect several properties at the same time. Hence, SPR analysis is crucial to understand which chemical modifications lead to favorable value changes in order for compound optimization to be a targeted process.

Compound optimization can be approached with different methodologies. Prediction methods and visualizations give insights into how chemical modifications affect property values. Composite measures are frequently used to drive compound optimization or to filter out compounds thought to have unfavorable properties. Moreover, multi-objective optimization algorithms are used to find the most suitable compounds that best balance different properties.

1.4.1 Quantitative structure-property relationship

Quantitative structure-property relationship (QSPR) produces mathematical models that attempt to predict property values of a compound based on its structure.⁶⁶ Hansch and coworkers pioneered QSPR studies, focusing on activity prediction or quantitative structure-activity relationship (QSAR).⁶⁷ They analyzed how different substitutions of a benzene ring altered properties such as lipophilicity. Then, they correlated these changes to activity using linear functions. Along with activity, toxicity prediction has also been heavily studied and is often used by environmental government agencies and pharmaceutical companies to detect hazardous substances or identify potentially toxic drug candidates.⁶⁶

The first QSPR models aimed to provide a mechanistic interpretation of relations between chemical structure and biological properties. Recent studies increasingly focus on accurate predictions or classifications using nonlinear methods taken from the machine learning field. These models are called “black boxes” because the results they provide are difficult to rationalize chemically.⁶⁸ However, recent studies have developed visualization strategies to better understand the prediction results of these models.⁶⁹

Support vector machine (SVM)⁷⁰ is a popular machine learning method that has been used in SPR analysis. SVM searches for a hyperplane that best separates positive and negative instances in the data set. It has been used to predict activity cliffs⁷¹ and find highly potent compounds⁷². The adaptation of SVM to regression is called support vector regression (SVR).⁷³ They are methodologically very similar. SVR searches for a function that can predict values with a maximal error of ϵ . The regression function is based on the hyperplane formulation from SVM methodology and is defined as:

$$f(x) = \langle w, x \rangle + b$$

Similar to SVM, slack variables (ξ_i and ξ_i^*) are used to model penalties when the predictive function cannot fit the data with a maximal error of ε . The parameter C determines the cost of these slack variables for the optimization algorithm. SVR attempts to minimize w based on a set of constraints.

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \\ & \text{subject to} && \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned}$$

This optimization problem can be solved using the Lagrange formulation. Thereby, α_i , α_i^* , η_i , and η_i^* denote the Lagrange multipliers for each of the four conditions specified above. All four multipliers have to be non-negative. The value of the partial derivatives of the Lagrange function for w , b , ξ_i , and ξ_i^* is zero at an optimal point, allowing to reformulate the optimization problem as:

$$\begin{aligned} & \text{maximize} && \begin{cases} -\frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle \\ -\varepsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \sum_{i=1}^l y_i (\alpha_i - \alpha_i^*) \end{cases} \\ & \text{subject to} && \sum_{i=1}^l (\alpha_i + \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C] \end{aligned}$$

From this system, w can now be expressed as:

$$w = \sum_{i=1}^l (\alpha_i - \alpha_i^*) x_i$$

and the function can be then reformulated as:

$$f(x) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b$$

Therefore, w is a linear combination of the training points x_i and the prediction formula only needs to calculate dot products between data points. Only those data points for which $(\alpha_i - \alpha_i^*)$ is non-zero are taken into account for the model

and are called the support vectors. Although the regression formula described above can only model linear relations between structure and property, it can be adapted for nonlinear regression with the “kernel trick”.⁷⁴ Kernel functions replace the dot product $\langle x_i, x \rangle$ to obtain nonlinear models. The kernels are defined as: $k(x_i, x) = \langle \Phi(x_i), \Phi(x) \rangle$, where Φ is a mapping function into a higher dimensional space. By finding a suitable kernel function, an explicit mapping of the data points does not need to be calculated. However, unless the kernel function applied is symmetric and positive semi-definite there are no guarantees that the SVR optimization will find solutions.

1.4.2 Visualization

Visualizations allow the study of chemical space. They can offer insight into SPRs present among compounds of a data set. Therefore, they can be used to highlight subsets of compounds that present interesting properties and search for regions of space with drug-like compounds. Most implementations have focused first on activity analysis. Consequently, SAR visualization has received a lot of attention in the last years.³³ Several applications have been developed based on different molecular representations or for specific types of data sets.

Visualizations can be categorized as coordinate-free or coordinate-based. In coordinate-free representations, the distance between molecules in the plot is meaningless. A common example is a graph. In coordinate-based visualizations, the distance between molecules in the display correlates with their similarity. A prototypical example is a scatter plot. The first implementations did not represent molecules directly but rather molecule pairs.

Coordinate-based plots were introduced for SAR analysis with the Structure-Activity Similarity (SAS) maps.⁷⁵ SAS maps are two-dimensional scatter plots where every data point represents a compound pair. One axis measures the similarity between the two compounds of the pair. The other axis represents the potency difference. This visualization can quickly highlight activity cliffs, as they concentrate on a specific quadrant of the plot. Several other scatter plot designs have been derived from the SAS maps. They all use compound pairs as data points but differ in the information that is displayed on the axes. Dual and triple activity difference maps are two- and three-dimensional scatter plots where each axis displays potency difference for a defined target.⁷⁶ In comparison to SAS maps, compound similarity is not modeled explicitly in the represented space but can be added through data point annotation. Additionally, molecule pairs with low

similarity can be omitted to focus the visualization on selectivity switches and multi-target activity cliffs.

In the previous paragraph, only similarity values between molecules were needed because compound pairs, rather than single molecules, were studied. However, the distribution of individual compounds in high-dimensional property space cannot be visualized directly. Therefore, dimensionality reduction techniques are used to obtain representations of chemical space suitable for visualization. The more frequently used methods in chemoinformatics are principal component analysis (PCA) and multi-dimensional scaling (MDS).

PCA performs an orthogonal transformation of the original high-dimensional space.⁷⁷ It generates a set of uncorrelated principal components created as linear combinations of original descriptors and ranked based on the fraction of original variance that they conserve. Principal components can be obtained from the eigenvectors of the covariance matrix of the descriptor values and can therefore vary when the descriptor values are scaled.⁷⁸

MDS is not a transformation like PCA but a mapping. Its modern implementation was defined by Kruskal.⁷⁹ Each data point in high-dimensional space is mapped to a point in a space with reduced dimensionality (for display purposes it will be two- or three-dimensional). The mapping function minimizes the difference between the distances in high-dimensional space and the distances in the reduced space.

Three-dimensional activity landscapes use dimensionality reduction to display chemical space.⁸⁰ In contrast to SAS maps, each data point is a molecule rather than a molecule pair. Principal components of a descriptor set or MDS of a fingerprint-based similarity matrix are used to obtain a two-dimensional representation of chemical space. Activity is displayed on the third axis. A surface is interpolated between data points, colored based on the activity values, to create a plot reminiscent of geographical landscapes (Figure 8A). This activity landscape gives a quick overview of the SAR character of the data set. Continuous sets do not have large differences in their potency values and therefore resemble plains. Discontinuous sets, however, have a much more mountain-like character with highly potent molecules generating peaks and weakly potent molecules producing valleys. An extension to multi-target analysis has been published where the chemical and biological space are visualized together using radial coordinates.⁸¹

Coordinate-free visualizations were popularized by the Network-like Similarity Graph (NSG).⁸² The underlying data structure is a fingerprint-based similarity

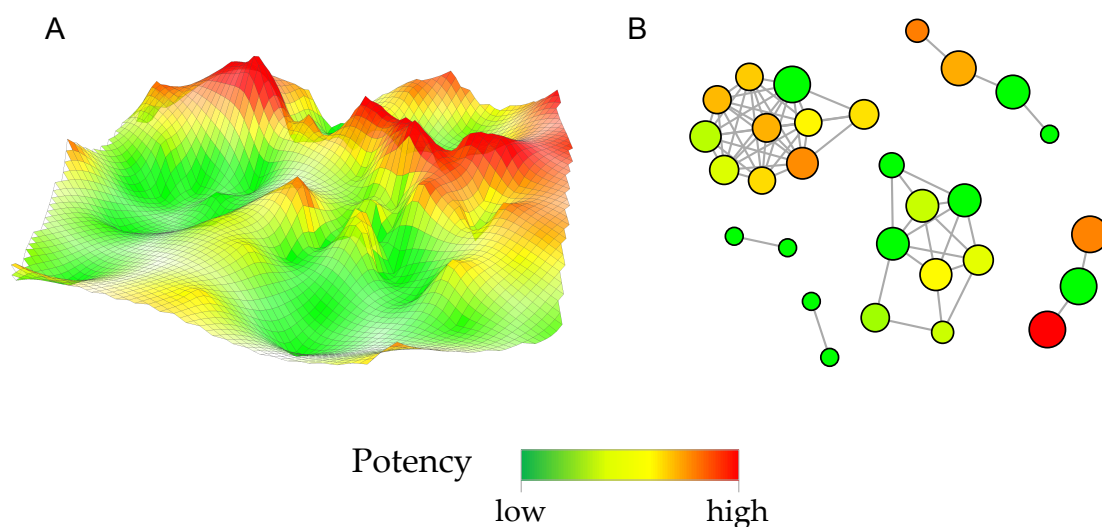


Figure 8: SAR visualization. A three-dimensional activity landscape (A) and a network-like similarity graph (B) are shown. In both visualizations, potency is represented by the same color scheme, going from red (high potency) through yellow to green (weak potency). In (B), the size of the nodes correlates with their compound-based SAR discontinuity measure.

matrix that is transformed into an adjacency matrix using a pre-specified threshold value. All pairwise relations between compounds are contained in the similarity matrix but only those that are of interest will be displayed. The nodes in the graph are colored according to activity values. Their size correlates with their SAR discontinuity, i.e., if many neighboring molecules have large differences in activity the node will be large. The layout of the graph is generated algorithmically (usually with force-directed layouts such as the Fruchterman-Reingold algorithm⁸³) to maximize clarity. Therefore, pairwise distances between molecules are meaningless in the visualization (Figure 8B). The NSG has been adapted for selectivity⁸⁴, mechanism of action⁸⁵, or multi-target analysis⁸⁶.

MMP relationships have also been used to generate network representations such as the bipartite matching molecular series graph.⁵² Each key in the index table that has at least two values creates an uncolored node connected to nodes representing the molecules that generated the value fragments. Molecule nodes are colored based on activity. If molecules participate in MMP relationships based on different key fragments, they will be connected to several key nodes. Additionally, scaffold relations have been used as the basis of a visualization. In the layered skeleton-scaffold organization graph, BMSs and CSKs are systematically generated for all compounds.⁸⁷ Each CSK is displayed as a black square containing one pie

chart per BMS they represent. Each pie chart is divided into equal sections for each compound the BMS represents and each section is colored based on potency. CSKs are laid out in concentric circles based on their ring number and connected to other CSKs with which they have substructure relationships.

1.4.3 Composite measures in compound optimization

Measures that combine several properties have been described to filter compounds with unfavorable properties. One well known example is Lipinski's rule of five that classifies compounds as orally available or not.⁸⁸ A research group at Pfizer analyzed the values of four molecular descriptors for compounds that reached phase II clinical trials. These molecules were considered to have good oral bioavailability. They found a set of threshold values that contained around 90% of the compounds under study. According to these rules, a molecule is expected to have good oral bioavailability for values of:

- molecular weight lower than 500 Da
- lipophilicity (as cLogP) lower than 5
- number of hydrogen bond donors (OH and NH groups) less than or equal 5
- number of hydrogen bond acceptors (O and N atoms) less than or equal 10

It must be emphasized that although the simplicity of these rules have made them ubiquitous they are not infallible. There are successful oral drugs, such as atorvastatin, that fail one or more of these rules.⁸⁹ Natural products, source of several drugs, are also frequent violators of these rules.⁹⁰

Another popular measure applied in compound optimization is ligand efficiency.⁹¹ The concept originated from an analysis comparing binding affinities to molecular size.⁹² According to this study, for small molecules the addition of one atom can lead to an improvement of up to 1.5 kcal/mol in binding affinity. However, as molecules become larger, this gain per atom is reduced. The energy contributed by each atom becomes lower and the efficiency of the ligand decreases.

Ligand efficiency has been measured in many ways.⁹³ At its core, it is a ratio of affinity against size. Affinity can be measured as change in free energy or through dissociation constants. Similarly, size can be measured as number of HAs or with molecular weight. Ligand efficiency has been used to drive compound optimization successfully.⁹⁴ There are alternative measures that compare activity not against size but against lipophilicity⁹⁵ or polar surface area⁹⁶.

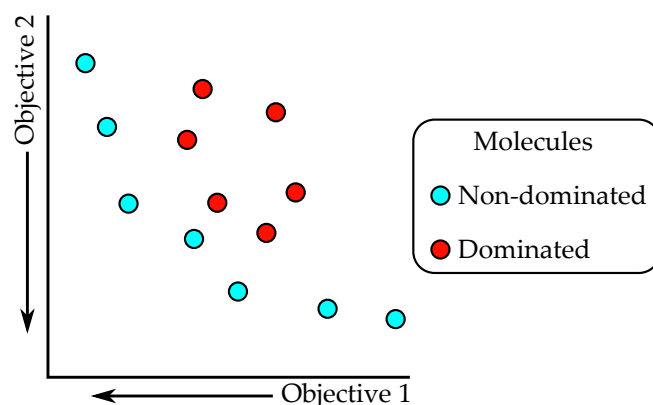


Figure 9: Pareto optimization. A schematic representation of a multi-objective optimization is shown. Two different objectives are minimized. Each data point corresponds to a molecule. If any other compound has lower values for both objective functions, it is considered dominated and colored red. If no other molecule has lower values for both objective functions, it is considered non-dominated and colored blue.

1.4.4 Multi-objective optimization

Compound optimization is straightforward when a single objective is applied to rank compounds. However, if two or more conflicting objectives are evaluated, ranking compounds becomes challenging. Rather than focusing on each objective sequentially, multi-objective optimization searches for solutions that represent the best compromise between the different objectives.⁹⁷ Consider a set of compounds ranked based on different objectives. For each objective, an individual ranking is obtained. Those molecules for which no other compound is better in all rankings are considered non-dominated and represent unique optimal trade-offs between the different objectives (Figure 9). The concept originated in economics and was developed by Vilfredo Pareto.⁹⁸ Consequently, these solutions are often called Pareto solutions and their combination the Pareto front. However, the modern mathematical formulation of multi-objective optimization was pioneered by Kuhn and Tucker.⁹⁹

Multi-objective optimization has been applied to docking, *de novo* compound design, and library generation.¹⁰⁰ Heuristic methods are often needed to drive the optimization because of the large size of chemical space. Some of the more common ones are evolutionary algorithms that simulate natural selection.¹⁰¹ Other algorithms that use biological concepts are swarm intelligence algorithms like

particle swarm optimization, which mimics the behavior of birds in a flock.¹⁰² Particle swarm optimization has been applied to obtain subsets of compounds from a data set with desired properties.¹⁰³

1.5 Outline of the thesis

This thesis focuses first on the development of novel applications of MMPA. In chapter 2, the ionization state of publicly available bioactive molecules is explored. The frequency of ionization state changes among MMP partners is analyzed. Additionally, the relation between ionization state and activity is rationalized. In chapter 3, ligand efficiency is examined. The difference in ligand efficiency is compared on the basis of different molecular representations of activity cliffs, including MMP-cliffs.

The following chapters focus on extensions of MMPs. In chapter 4, second generation MMPs created on the basis of retrosynthetic rules are described. These new MMPs, named RECAP-MMPs, have transformations that are easier to apply to chemical synthesis than standard MMPs. In chapter 5, MMS are used to obtain SAR information for confirmed hit compounds. SAR information is gained from potency-ordered MMS to which these hits are mapped through MMP fragmentation. In chapter 6, MMPs are used to develop kernel functions for SVR. These SVR models are applied to predict the potency change between MMP partners.

In the final chapters, the focus changes to representations of chemical space and their utility to multi-objective compound optimization. In chapter 7, MMPA is combined with visualization of high-dimensional space. Principal component plots are used to rationalize property changes from an MMP-driven compound optimization procedure. In chapter 8, star coordinate and parallel coordinate plots are introduced to the medicinal chemistry community. They are applied to differentiate between distinct drug-like subspaces obtained from an optimization task. In chapter 9, a novel visualization to explore high-dimensional spaces using coordinate-free representations is presented. It extends the chemical space network concept and offers an overview of important similarities in property space to quickly focus on specific compound subsets of interest. The final chapter summarizes the main points of this work and serves as a conclusion of the thesis.

2 Target-based analysis of ionization states of bioactive compounds

Introduction

The ionization state of a compound is important for its activity and *in vivo* properties. Many drugs, nearly four out of five, contain chemical groups that are partly ionized under physiological pH, i.e., the pH commonly encountered in humans.¹⁵ Because of its importance, there have been several studies of ionization state of drugs and bioactive compounds.^{104–106} However, the effect of structure modifications on the ionization state has not been previously evaluated. In this study, the ionization state of bioactive compounds is analyzed in detail. The relationship of activity and ionization state is evaluated for individual targets and superfamilies. Finally, the effect of small structure modifications is studied through MMP relationships.

My main contribution to this work was the analysis of the ionization state distribution among ligands active against specific targets and superfamilies. This study was published as:

S. Kayastha, A. de la Vega de León, D. Dimova, J. Bajorath. Target-based analysis of ionization states of bioactive compounds. *MedChemComm* **2015**, 6, 1030–1035.

Materials and methods

Bioactive compounds were obtained from the ChEMBL database¹⁰⁷ (version 19). Compounds were extracted only if equilibrium constant (K_i) values were available with the highest confidence level for human proteins. In case several activity measures were present for a single compound and they differed by more than one order of magnitude, this compound was excluded from the analysis. If all values were within one order of magnitude, the geometric mean was taken as the final activity measure. Compounds were considered highly potent if their potency value was at least 100 nM and weakly potent if it was at most 1 μ M. A total of 80 776 compounds were obtained and they were organized in 719 different target sets. Each target was assigned to a superfamily based on the ChEMBL target classification.

The dissociation constant (K_a) is the equilibrium constant between the ionized and neutral form of a chemical group. The dissociation constant for the most acidic chemical group (A_pK_a) and the most basic chemical group (B_pK_a) were obtained from ChEMBL as pK_a values, i.e., the negative decadic logarithm of K_a . Compounds were classified on the basis of A_pK_a and B_pK_a values as four ionization state classes (IS-classes): neutral, acidic, basic, and zwitterionic. The classification was based on a previously published methodology.¹⁰⁶ The Henderson-Hasselbalch equation¹⁰⁸ was employed, along with a physiological pH of 7.4, to calculate how ionized the most acidic and basic chemical groups were. If both an acidic and a basic group in the same molecule were more than 50% ionized, the compound was classified as zwitterionic. If only an acidic group or a basic group, but not both, were more than 50% ionized, the compound was classified as acidic or basic, respectively. If no chemical group was more than 50% ionized, the compound was classified as neutral. Finally, if no A_pK_a and B_pK_a values were present in the database, the molecule was not classified (NA). Target sets were excluded from the target distribution analysis if they contained less than 10 compounds or if more than 20% of their compounds were not classified, leaving 351 target sets.

MMPs were obtained for each target set (see section 1.3.4) using an in-house Java program based on the OpenEye toolkit¹¹⁰. MMPs were not calculated for compounds that could not be classified. Those target sets that generated less than 50 MMPs were excluded from the MMP analysis. Finally, 290 different target sets were left, representing 66 871 compounds and 338 419 MMPs. Compounds were assigned to three different categories based on the ionization state of MMP partners (Figure 10). If all partners had the same IS-class as the molecule, it belonged to

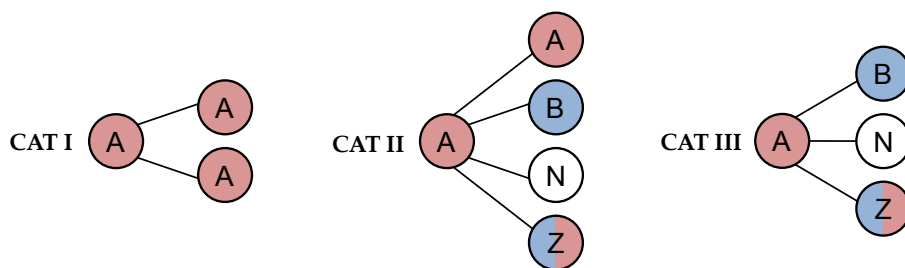


Figure 10: Chemical neighborhoods. The figure shows the IS-class composition of chemical neighborhoods formed by compounds assigned to category (CAT) I–III and their MMP partners. Color denotes IS-class (red, acidic; blue, basic; white, neutral; dual colored, zwitterionic). The figure has been adapted from [109].

category I. If some but not all partners belonged to different IS-classes, the molecule was added to category II. Finally, if all partners belonged to different IS-classes than the compound, it was considered category III.

Results and discussion

We first analyzed the frequency of different IS-classes among bioactive compounds (Figure 11A). Similar frequency of basic (39.2%) and neutral compounds (38.6%) were present in the target sets. The frequency of acidic (10.3%) and zwitterionic (3.5%) compounds was much lower. Comparable proportions were found when focusing only on highly potent compounds. Over all target sets, IS-class distribution of highly potent compounds did not differ from weakly potent compounds.

Next, we focused on the distribution of IS-classes in different target sets. For individual target sets, one IS-class was usually prevalent. In 90% of the target sets, at least 50% of the compounds belonged to the same IS-class. Moreover, for 40% of the target sets, at least 80% were assigned to the same class. Table 1 lists target sets with over 200 compounds where the largest prevalence of IS-classes was observed. For many targets, basic compounds represented the dominant IS-class. Nevertheless, large prevalence of acidic compounds (among prostaglandin D2 receptor 2 ligands) and neutral compounds (among vanilloid receptor ligands) was also observed. These values indicate that many targets may preferentially

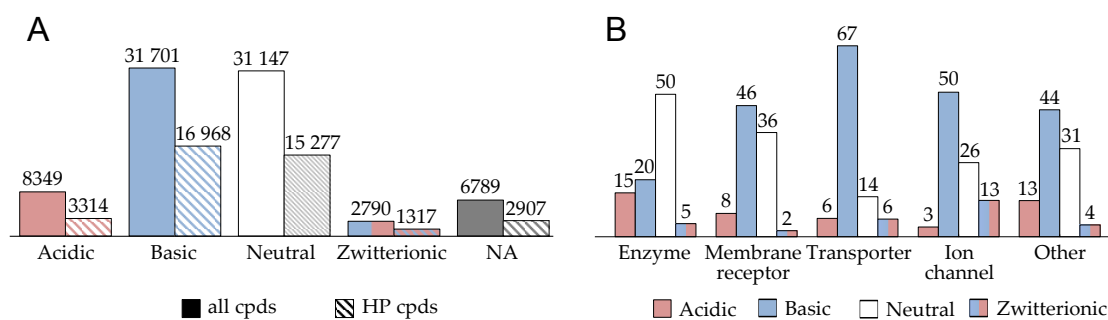


Figure 11: IS-class distribution. (A) Distribution of IS-classes among bioactive compounds (solid bars) and a subset of highly potent (HP) compounds (striped bars). Numbers over the bars represent the number of compounds for each IS-class. (B) IS-class distribution over superfamilies. The percentage of compounds belonging to each IS-class is displayed over each bar. Compounds that could not be classified are not shown. The figure has been adapted from [109].

bind compounds from a specific IS-class. Furthermore, for 57 target sets a notable difference in the distribution of IS-classes between weakly and highly potent compounds was found. For example, for the neurokinin 2 receptor target set, more than 70% of highly potent compounds were basic while only 12% of weakly potent molecules were. On the other hand, almost 80% of weakly potent neurokinin 2 ligands were neutral but only 24% of highly potent compounds were. Taken together, these results provide further evidence of the importance of ionization state for compound activity values.

Target sets were further grouped into four different superfamilies: enzymes, membrane receptors, transporters, and ion channels. Targets that did not belong to any of the previous four superfamilies were grouped together as other. Large differences in IS-class distribution were present between the superfamilies (Figure 11B). For enzymes, neutral compounds represented the majority IS-class (50%) while basic and acidic compounds had similar frequency, 20% and 15% respectively. Basic compounds were the most frequent IS-class in all other superfamilies. However, the difference in frequency to neutral compounds was small in membrane receptor (46% to 36%) and other (44% to 31%) but large in transporter (67% to 14%) and ion channel (50% to 26%) superfamilies. Acidic and zwitterionic compounds were not frequent in the superfamilies, rarely exceeding a frequency of 10%. Similar to individual target sets, superfamilies displayed marked preferences for specific IS-classes.

Table 1: Target sets that display large ionization state class prevalence^a

TID	Target name	# Cpds	IS-class
5071	Prostaglandin D2 receptor 2	468	99% acidic
4794	Vanilloid receptor	253	97% neutral
259	Melanocortin receptor 4	1217	92% basic
264	Histamine H3 receptor	2023	92% basic
1898	Serotonin 1b (5-HT1b) receptor	364	92% basic
335	Protein-tyrosine phosphatase 1B	243	91% acid
344	Melanin-concentrating hormone receptor 1	846	90% basic
4644	Melanocortin receptor 3	350	90% basic
4608	Melanocortin receptor 5	268	88% basic
1983	Serotonin 1d (5-HT1d) receptor	359	87% basic
1800	Corticotropin releasing factor receptor 1	473	84% neutral
222	Norepinephrine transporter	1010	84% basic
232	Alpha-1b adrenergic receptor	290	84% basic
228	Serotonin transporter	1337	83% basic
2492	Neuronal acetylcholine receptor protein alpha-7 subunit	253	83% basic
238	Dopamine transporter	867	81% basic
3798	Calcitonin gene-related peptide type 1 receptor	349	81% neutral
1916	Alpha-2c adrenergic receptor	295	80% basic
2954	Cathepsin S	375	80% neutral
210	Beta-2 adrenergic receptor	241	80% basic

^aThe top 20 target sets with largest prevalence of a single ionization state class (IS-class) are reported. The table lists the ChEMBL target identifier (TID), name, number of compounds (# Cpds), and IS-class.

The second part of the analysis focused on the effect of small structure changes on the ionization state and the description of chemical neighborhoods on the basis of IS-classes. Most MMPs were ionization state conservative because only in 13.6% the two compounds of the pair had different IS-classes. Even though most chemical transformations did not alter the ionization state of a molecule, almost a third of the compounds had heterogeneous chemical neighborhoods. 28.7% of the compounds were assigned to category II and 2.5% to category III. 68.8% of

all molecules explored had neighborhoods with conserved ionization states. The conservation of IS-classes for MMP partners is a favorable characteristic, because binding to a particular target often requires a specific IS-class.

Conclusions

We have systematically analyzed the ionization state of publicly available bioactive compounds on the basis of high-confidence activity data. The focus of this study was not on drug compounds and this set it apart from many previous analyses of ionization state. Bioactive compounds were predominantly neutral or basic under physiological pH. The overall distribution of IS-classes in highly potent and weakly potent compounds was very similar. However, for many target sets a strong preference for a specific IS-class was detected. There were also many target sets where different IS-class distributions were found for highly and weakly potent compounds. Small structural changes encoded in MMP transformations only rarely altered the ionization state of a molecule.

Ionization state has been further established as an important property for drug development efforts. A specific IS-class is often found in most compounds that bind to a particular target and structural changes do not often change the IS-class. Another important property used in compound optimization is ligand efficiency, a measure that relates potency and size. In the next chapter, a ligand efficiency analysis is carried out on the basis of different activity cliff representations including MMP-cliffs.

3 Formation of activity cliffs is accompanied by systematic increases in ligand efficiency from lowly to highly potent compounds

Introduction

Ligand efficiency has proven to be an effective measure to drive compound optimization. Additionally, activity cliffs represent important sources of SAR information for compound optimization. Despite the fact that both topics have seen large interest in the chemoinformatics and medicinal chemistry community, their connection has never been explored before. It is unknown if the large potency increase found in activity cliffs is correlated with a proportional increase in the size of the molecule that would leave ligand efficiency unchanged. In this study, we present an analysis of the ligand efficiency change between compounds forming activity cliffs. Ligand efficiency change is compared for activity cliffs based on fingerprint-based similarity indices and MMPs.

Reprinted with permission from “A. de la Vega de León, J. Bajorath. Formation of activity cliffs is accompanied by systematic increases in ligand efficiency from lowly to highly potent compounds. *The AAPS Journal* **2014**, 16(2), 335–341”. Copyright 2014 Springer

Research Article

Formation of Activity Cliffs Is Accompanied by Systematic Increases in Ligand Efficiency from Lowly to Highly Potent Compounds

Antonio de la Vega de León¹ and Jürgen Bajorath^{1,2}

Received 19 November 2013; accepted 9 January 2014; published online 30 January, 2014

Abstract. Activity cliffs (ACs) are defined as pairs of structurally similar compounds sharing the same biological activity but having a large difference in potency. Therefore, ACs are often studied to rationalize structure-activity relationships (SARs) and aid in lead optimization. Hence, the AC concept plays an important role in compound development. For compound optimization, ligand efficiency (LE) represents another key concept. LE accounts for the relation between compound potency and mass. A major goal of lead optimization is to increase potency and also LE. Despite their high relevance for drug development, the AC and LE concepts have thus far not been considered in combination. It is currently unknown how compounds forming ACs might be related in terms of LE. To explore this question, ACs were systematically identified on the basis of high-confidence activity data and LE values for cliff partners were determined. Surprisingly, a significant increase in LE was generally detected for highly potent cliff partners compared to their lowly potent counterparts, regardless of the compound classes and their targets. Hence, ACs reveal chemical modifications that determine SARs and improve LE. These findings further increase the attractiveness of AC information for compound optimization and development.

KEY WORDS: activity cliffs; drug development; ligand efficiency; matched molecular pairs; structure-activity relationships.

INTRODUCTION

The activity cliff (AC) concept plays a key role in structure-activity relationship (SAR) analysis (1–3). ACs are generally defined as pairs or groups of structurally similar or analogous active compounds having a large difference in potency (1–3). As such, ACs are prime indicators of SAR discontinuity (1,2) because small chemical changes lead to large biological effects. Therefore, SAR determinants can often be deduced from ACs (2,3). Although ACs have traditionally been considered on a case-by-case basis by focusing on one compound series at a time, they have recently been systematically investigated across compounds active against current pharmaceutical targets (2,3), thereby considerably increasing the knowledge base for SAR exploration and compound development (3).

Lead optimization generally aims to improve target-specific potency and other compound properties relevant for drug development (4). In order to increase potency, various R-groups are typically added to candidate compounds and their hydrophobic character is frequently increased (4). Thus, potency improvements often come at a price of increasing molecular mass and hydrophobicity, which in turn result in less favorable absorption, distribution, metabolism, and excretion

characteristics. Therefore, potency is often not considered as an individual property but related to molecular mass, thus leading to the ligand efficiency (LE) assessment (5,6). The LE concept has its origins in attempts to account for compound potency on a per-atom basis (7). Accordingly, LE is usually calculated by dividing compound potency (e.g., pK_i values) by the number of non-hydrogen atoms in a compound (5) or by its molecular weight (6). As such, LE is a simple and intuitive measure of compound optimization progress, despite some intrinsic limitations (8). Ideally, LE values should increase during compound optimization but not significantly decrease. In practice, LE values often remain more or less constant during successful optimization efforts (9–11).

Despite their intuitive nature and high relevance for compound optimization, the AC and LE concepts have thus far not been considered in combination. Rather, AC analysis has generally been potency-centric (2). Therefore, we have systematically analyzed ACs from an LE perspective and compared LE values for highly and lowly potent cliff partners across many different compound activity classes. The results of our analysis are presented herein.

MATERIALS AND METHODS

Datasets

Compounds against human targets were extracted from ChEMBL version 15 (12) by applying the following criteria. Only compounds with numerically exact K_i values reported

¹Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, 53113, Bonn, Germany.

²To whom correspondence should be addressed. (e-mail: bajorath@bit.uni-bonn.de)

for direct target interactions at the highest level of confidence (ChEMBL confidence score 9) were considered. If more than one K_i value for the same target was reported for a compound, it was discarded if these values differed by more than one order of magnitude. If the values fell within the same order of magnitude, their average was calculated as the final potency annotation. We obtained 610 different target sets comprising a total of 41,127 compounds. Individual target sets contained up to 2,307 compounds.

Molecular Representations

Molecular fingerprint- and graph-based compound representations (*e.g.*, matched molecular pairs; see below) were calculated to identify ACs (2,3). As fingerprints, MACCS structural keys (13), a dictionary containing 166 different molecular fragments, and the extended connectivity fingerprint with bond diameter 4 (ECFP4) (14), a topological fingerprint capturing layered atom environments, were calculated using the molecular operating environment (MOE) (15). These two fingerprints of different design are currently most frequently used in AC analysis (2). The molecular weight (MW) and logP value, a measure of lipophilicity, of each compound was also calculated with MOE.

Matched Molecular Pairs

Matched molecular pairs (MMPs) are defined as pairs of compounds that differ only by a structural change at a single site (16), *i.e.*, the exchange of a substructure, termed a chemical transformation (17). MMPs were systematically calculated for compounds in all target sets using an in-house implementation of the algorithm by Hussain and Rea (17) based on the OEChem toolkit (18). For AC assessment, transformation size-restricted MMPs were selected (19). The difference in size between the exchanged substructures was limited to at most 8 non-hydrogen atoms and the maximal size of an exchanged fragment was limited to 13 non-hydrogen atoms. In addition, the number of non-hydrogen atoms comprising the common parts (core structure) of two compounds had to be at least twice the size of each of two distinguishing substructures. These size restrictions generally limit transformations to chemically meaningful replacements (19). If several transformations met the size restrictions for a given compound pair, the smallest transformation was selected.

Activity Cliffs

For AC assessment, similarity and potency difference criteria must be specified. In order to limit the analysis to ACs of significant magnitude, a difference in potency (equilibrium constants) of at least two orders of magnitude was consistently applied (2,3). Alternative similarity criteria were considered. For MACCS and ECFP4 fingerprint representations, Tanimoto coefficient (20) values of at least 0.85 and 0.56, respectively, were required to qualify two compounds as cliff partners (3). ACs formed on the basis of MACCS and ECFP4 representations were designated fingerprint-cliffs. In addition, the formation of transformation size-restricted

MMPs was applied as a substructure-based similarity criterion for AC formation (3). If compounds in a transformation size-restricted MMP displayed a potency difference of at least two orders of magnitude, they formed a so-called MMP-cliff (19). For each AC, the compound with high potency and compound with low potency forming the cliff were designated the “highly potent cliff partner” and the “lowly potent cliff partner,” respectively.

Ligand Efficiency

LE was calculated using the Binding Efficiency Index (BEI) (6) defined as follows:

$$\text{BEI} = \text{p}K_i/\text{MW} \text{ [log unit/kDa]}.$$

Because BEI values were only calculated and compared for structurally similar/analogous compounds, corrections for potential size dependence were not required (8).

Statistical Analysis

Statistical analysis of data distributions was carried out using the R package stats (21).

Table I. Target Sets with Largest Numbers of ACs

Target name	No. ACs
Coagulation factor X	3,972
Melanocortin receptor 4	2,890
Mu opioid receptor	2,645
Cannabinoid CB2 receptor	2,380
Adenosine A2a receptor	2,290
Adenosine A3 receptor	2,096
Thrombin	1,810
Kappa opioid receptor	1,704
Histamine H3 receptor	1,643
Purinergic receptor P2Y12	1,601
Dopamine D2 receptor	1,524
Melanin-concentrating hormone receptor 1	1,500
Bradykinin B1 receptor	1,210
Histamine H4 receptor	1,126
Serotonin 6 (5-HT6) receptor	949
Calcitonin gene-related peptide type 1 receptor	918
Corticotropin releasing factor receptor 1	888
G protein-coupled receptor 44	853
Muscarinic acetylcholine receptor M3	832
Gonadotropin-releasing hormone receptor	755
Serotonin 1a (5-HT1a) receptor	720
Adenosine A2b receptor	695
Cannabinoid CB1 receptor	666
Vasopressin V1a receptor	590
Furin	519
Carbonic anhydrase I	518
Neuropeptide Y receptor type 5	513
Dopamine transporter	495
Dopamine D3 receptor	494
Delta opioid receptor	478

Targets yielding the largest number of ACs (fingerprint plus MMP-cliffs) are reported

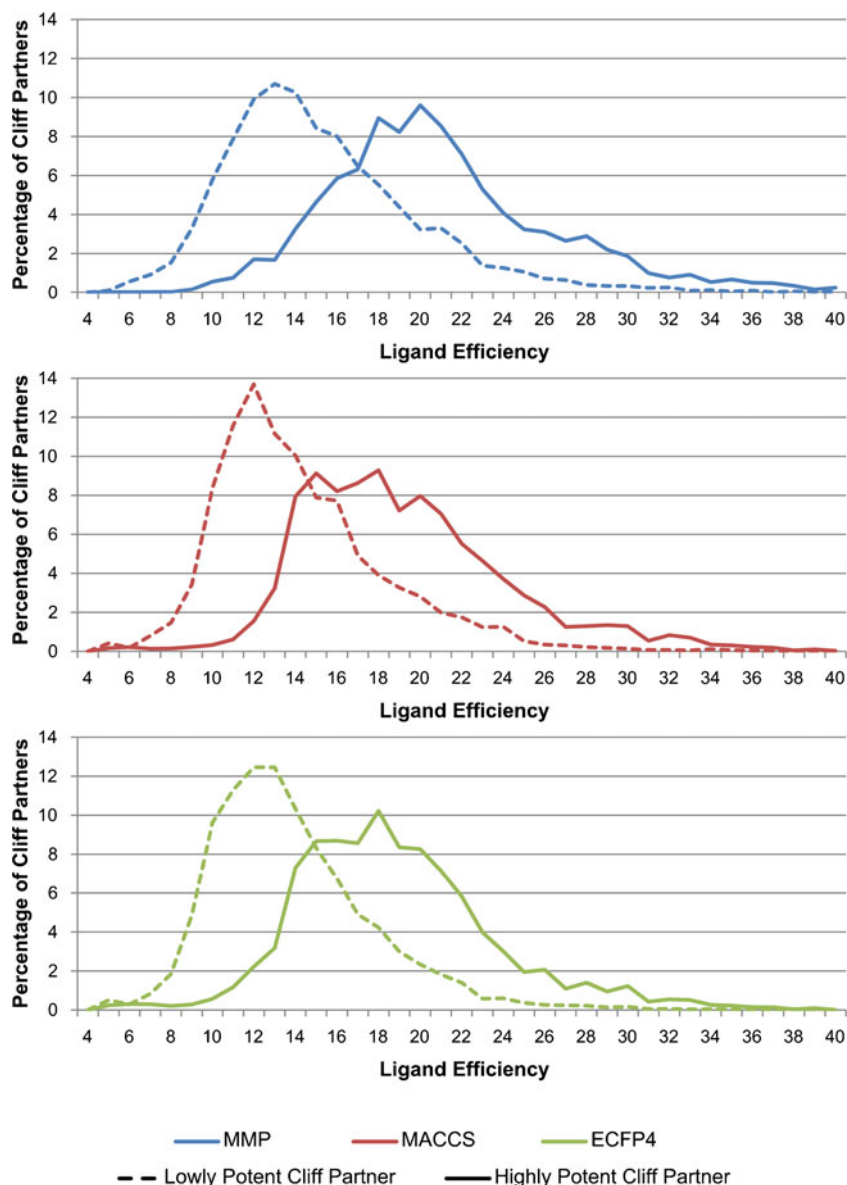


Fig. 1. Distribution of LE values. For all ACs obtained on the basis of MMPs (blue), MACCS (red), and ECFP4 (green), the distribution of LE values for lowly potent (dashed lines) and highly potent cliff partners (solid lines) is reported

RESULTS AND DISCUSSION

AC Statistics

For each of our 610 target sets, ACs were systematically calculated using alternative molecular representations. From all 41,127 compounds, 22,109 and 17,312 MACCS- and ECFP4-based fingerprint-cliffs were obtained, respectively. In addition, 18,208 MMP-cliffs were identified. Thus, a very large pool of ACs was available for our analysis, originating from compounds active against the spectrum of current pharmaceutical targets. Table I lists the 30 targets yielding most ACs. The frequency of occurrence and potency range distribution of ACs has previously been determined (22). ACs spanning a potency difference of at least two orders of magnitude on the basis of equilibrium constants over all available potency ranges

provide a statistically preferred and chemically reliable pool of ACs for further exploration (22). We adhere to this AC assessment herein.

LE Analysis

For each AC-forming compound, its LE value was calculated, and for each AC, the LE values of highly and lowly potent cliff partners were compared. Figure 1 reports the distribution of LE values for highly and lowly potent cliff partners identified on the basis of different molecular representations. In each case, LE values of highly potent cliff compounds were on average significantly larger than the values of lowly potent cliff partners. For different molecular representations, the profiles of the LE distributions were rather similar. Importantly, for 99.1, 96.9, and 97.4% of the MMP-, MACCS-, and ECFP4-based ACs, respectively, an

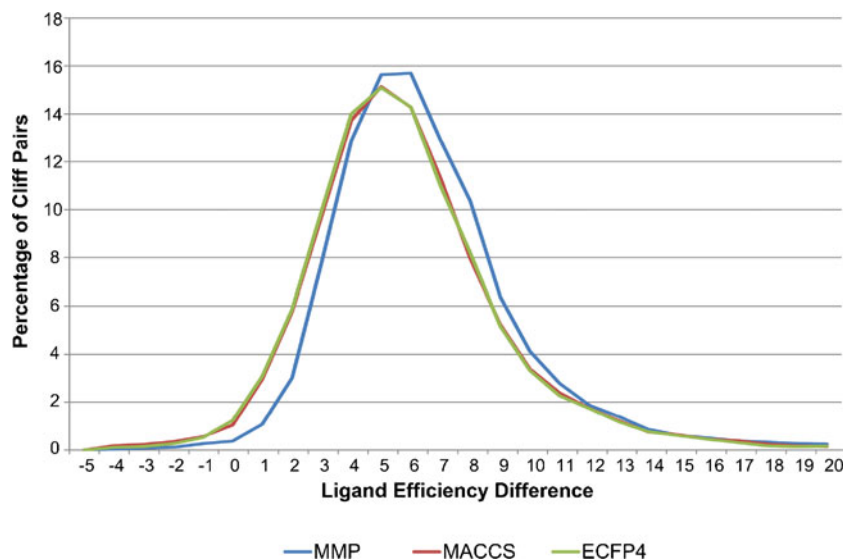


Fig. 2. LE difference distribution. The difference in LE between highly and lowly potent partners are compared for ACs obtained on the basis of MMPs (*blue*), MACCS (*red*), and ECFP4 (*green*). Negative values represent a decrease in LE as potency increases and positive values a corresponding increase in LE

increase in LE was detected for the highly potent compared to the lowly potent cliff compound; a surprising finding.

LE Differences

Figure 2 compares the distributions of LE differences between cliff-forming compounds. For fingerprint-based ACs, the distributions were extremely similar, with average LE difference values of 5.38 for both MACCS and ECFP4. However, for MMP-cliffs, the distribution was shifted towards larger LE differences, yielding an average value of 6.25. On the basis of a two sample unpaired *t* test (Table II), the difference between fingerprint- and MMP-based ACs was statistically highly significant. Thus, for the structurally more conservative MMP-based AC representations, larger differences in LE values between lowly and highly potent cliff partners were detected than for fingerprint-based AC representations that relied on the calculation of (whole-molecule) Tanimoto similarity. Hence, from an LE perspective, MMP-cliffs were preferred for AC representations.

LE vs. MW, Potency, and logP Differences

We also analyzed the relationship between LE and MW differences of AC partners. For 54.9, 58.2, and 57.3% of all MMP-, MACCS-, and ECFP4-based ACs, respectively, the highly potent cliff partners had larger MW than the lowly potent compounds. However, for 98.4, 95.4, and 95.5% of these ACs, the highly potent cliff partners also had larger LE values than their lowly potent counterparts. Figure 3 shows the comparison of LE and MW differences for MMP-cliffs and Fig. 4 the comparison of LE and potency differences. No statistically significant correlation between LE and MW or potency differences was detected. Furthermore, the relationship between LE and logP differences was also explored. LogP values of highly and

lowly potent cliff partners were calculated as a measure of lipophilicity. For MMP-, MACCS-, and ECFP4-based ACs, the average change in logP values between compounds forming an AC was 0.16, 0.19, and 0.20, respectively. Figure 5 shows a comparison of LE and logP differences for MMP-cliffs. No significant correlation between LE and logP differences was observed. Taken together, these findings indicated that the observed LE increases for ACs were largely independent of MW or lipophilicity variations between cliff partners. Hence, large potency differences between cliff partners mostly determined LE increases.

Exemplary ACs

In Fig. 6, four MMP-cliffs are shown in which the highly potent cliff partner had larger MW and LE values than the lowly potent compound. These MMP-cliffs involve compounds of different size and chemical complexity (as well as different activity). In the first two examples (from the top), the MW increase was small and the LE increase was large, as often observed for different ACs. In the two remaining examples, MW increases are nearly maximal for MMP-cliffs (given the transformation size restrictions). In these extreme cases, MW increases are large and LE increases are small. In

Table II. *T*-test for LE Difference Distributions

LE diff. distribution	T statistic	<i>p</i> value
MMP vs. MACCS	20.58	1.54E-93
MMP vs. ECFP4	19.59	5.93E-85
MACCS vs. ECFP4	-0.012	0.99

In order to compare LE difference distributions for AC sets according to Fig. 2, a two-sample unpaired *t* test was performed. Values of the *T* statistic and *p* values are reported

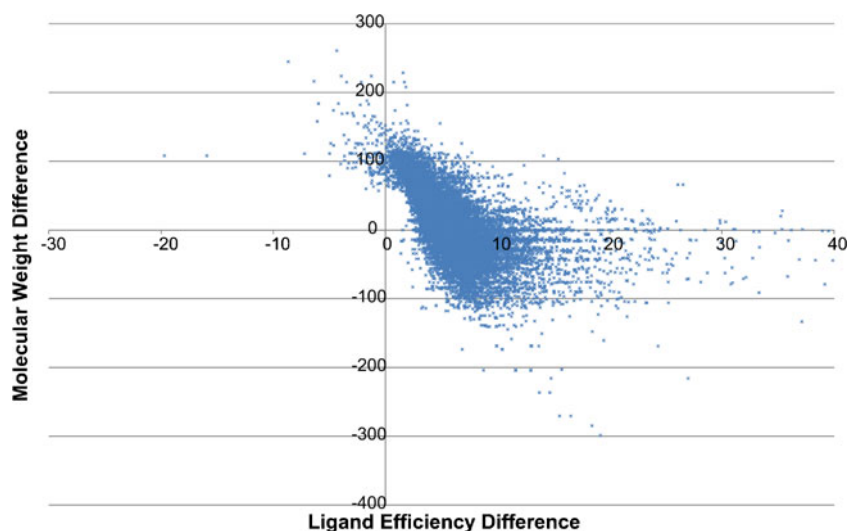


Fig. 3. LE vs. molecular weight difference. Each data point represents an MMP-cliff. Its position in the graph is determined by the LE and MW difference between highly and lowly potent cliff partners

the majority of cases, larger LE increases were observed, as reflected by the LE value and difference distributions reported herein.

CONCLUSIONS

The AC and LE concepts are focal points of SAR analysis and compound development. ACs are explored to identify SAR determinants and design analogs of active compounds, and increasing LE is utilized as a guiding principle during lead optimization. However, despite these conceptual relationships, ACs have, thus far, not been analyzed from an LE perspective. To these ends, we have carried out a large-scale analysis of ACs and calculated LE

values for cliff-forming compounds. From a total of more than 41,000 unique compounds belonging to 610 different target sets, ACs were systematically extracted on the basis of high-confidence activity data and alternative molecular representations. The resulting AC populations were subjected to LE analysis. For each AC, LE values of the highly and lowly potent cliff partners were compared. On the basis of this analysis, very strong trends were observed. Regardless of chosen molecular representations and target activities, in more than 96% of all ACs, highly potent cliff partners had consistently higher LE values than their lowly potent counterparts. Thus, the formation of ACs was accompanied by a systematic increase in LE in the direction of increasing compound potency. Increases in LE were not accompanied by

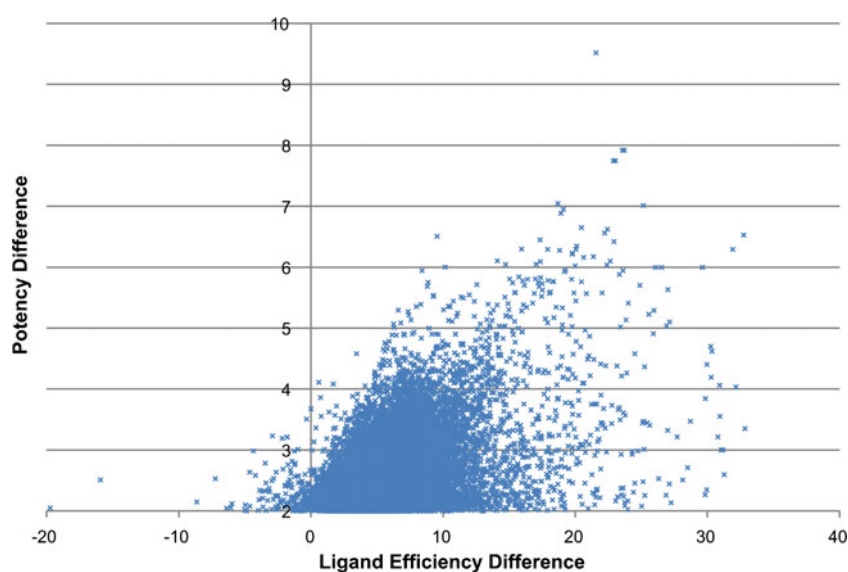


Fig. 4. LE vs. potency difference. Each data point represents an MMP-cliff. Its position in the graph is determined by the LE and potency difference between highly and lowly potent cliff partners

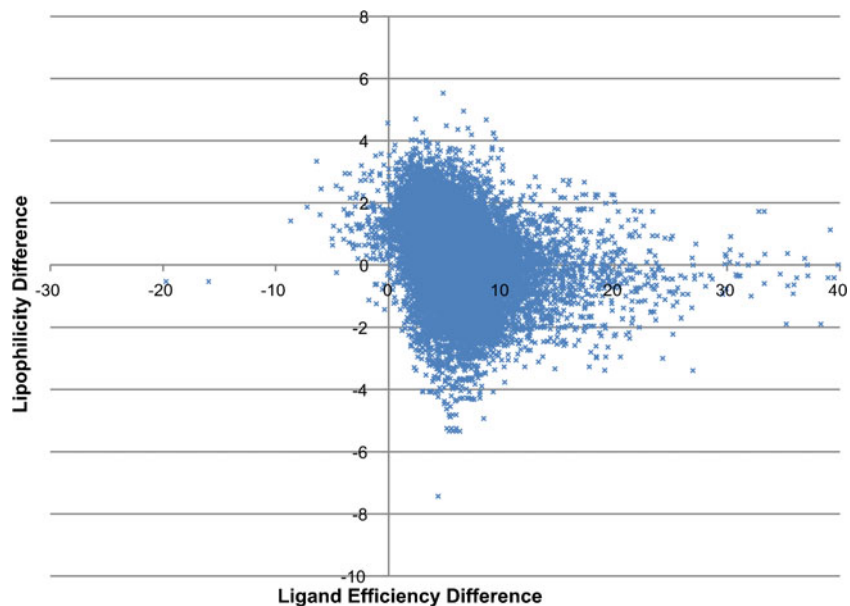


Fig. 5. LE vs. logP difference. Each data point represents an MMP-cliff. Its position in the graph is determined by the LE and logP difference between highly potent and lowly potent cliff partners

general increases in logP as a measure of lipophilicity. LE differences were larger for MMP- than for fingerprint-based ACs, and LE increases in ACs were independent of MW

variations between cliff-forming compounds. The systematic differences between LE values of highly and lowly potent AC compounds revealed by our analysis further increase the value

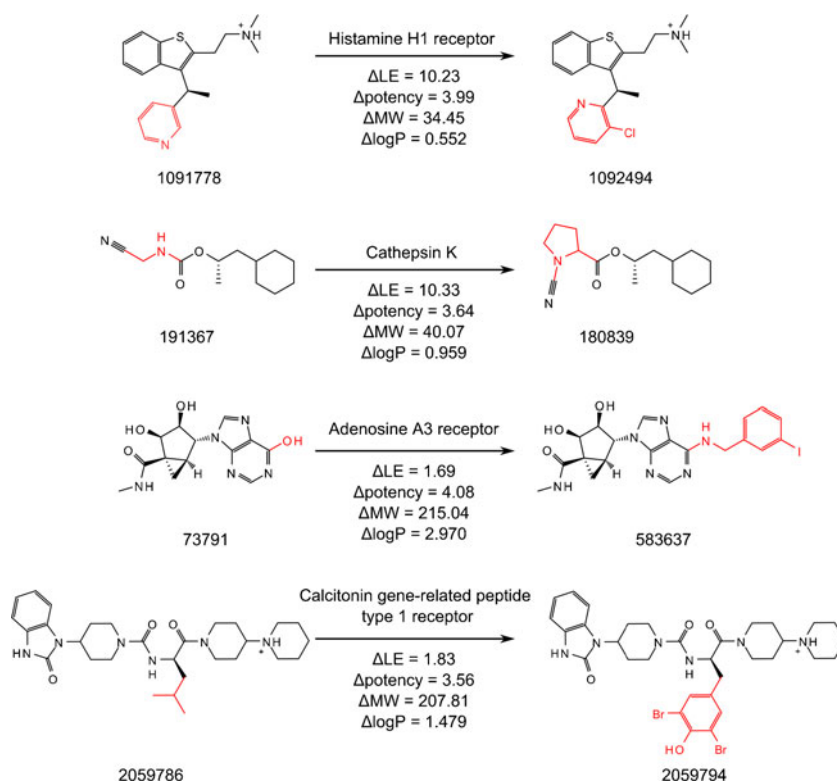


Fig. 6. Exemplary ACs. Four MMP-cliffs are shown. In each pair, the left compound represents the lowly potent and the right compound the highly potent cliff partner, (arrows point from the lowly to the highly potent compound). Substructures constituting the MMP transformation are highlighted in red. For compound pairs, ChEMBL IDs (below the compounds) are provided. In addition, potency (pK_i), MW, lipophilicity (logP), and LE differences are reported

of AC information for compound development. ACs not only uncover SAR determinants but critical chemical changes encoded by ACs also lead to LE improvements. Especially for MMP-cliffs, this dual role renders the underlying chemical transformations highly attractive for compound design.

ACKNOWLEDGMENTS

The authors thank Dilyana Dimova for help with compound datasets.

REFERENCES

1. Maggiora GM. On outliers and activity cliffs—why QSAR often disappoints. *J Chem Inf Model.* 2006;46(4):1535. doi:10.1021/ci060117s.
2. Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. *J Med Chem.* 2012;55(7):2932–42. doi:10.1021/jm201706b.
3. Stumpfe D, Hu Y, Dimova D, Bajorath J. Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *J Med Chem.* 2013. doi:10.1021/jm401120g.
4. Wermuth CG. *The practice of medicinal chemistry.* 3rd ed. London: Academic; 2008.
5. Hopkins AL, Groom CR, Alex A. Ligand efficiency: a useful metric for lead selection. *Drug Discov Today.* 2004;9(10):430–1. doi:10.1016/S1359-6446(04)03069-7.
6. Abad-Zapatero C, Metz JT. Ligand efficiency indices as guideposts for drug discovery. *Drug Discov Today.* 2005;10(7):464–9. doi:10.1016/S1359-6446(05)03386-6.
7. Kuntz ID, Chen K, Sharp KA, Kollman PA. The maximal affinity of ligands. *Proc Natl Acad Sci USA.* 1999;96(18):9997–10002. doi:10.1073/pnas.96.18.9997.
8. Reynolds CH, Bembenek SD, Tounge BA. The role of molecular size in ligand efficiency. *Bioorg Med Chem Lett.* 2007;17(15):4258–61. doi:10.1016/j.bmcl.2007.05.038.
9. Hajduk PJ. Fragment-based drug design: how big is too big? *J Med Chem.* 2006;49(24):6972–6. doi:10.1021/jm060511h.
10. Perola E. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *J Med Chem.* 2010;53(7):2986–97. doi:10.1021/jm100118x.
11. Tanaka D, Tsuda Y, Shiyama T, Nishimura T, Chiyo N, Tominaga Y, *et al.* A practical use of ligand efficiency indices out of the fragment-based approach: ligand efficiency-guided lead identification of soluble epoxide hydrolase inhibitors. *J Med Chem.* 2011;54(3):851–7. doi:10.1021/jm101273e.
12. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2012;40(D1):D1100–7. doi:10.1093/nar/gkr777.
13. MACCS structural keys; Accelrys, Inc., 5005 Wateridge Vista Drive, San Diego, CA 92121, USA
14. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model.* 2010;50(5):742–54. doi:10.1021/ci100050t.
15. Molecular operating environment (MOE), 2011.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2011
16. Kenny PW, Sadowski J. Structure modification in chemical databases. In: Oprea TI, editor. *Chemoinformatics in drug discovery.* Weinheim: Wiley-VCH; 2005. p. 271–85.
17. Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J Chem Inf Model.* 2010;50(3):339–48. doi:10.1021/ci900450m.
18. OEChemTK, v2012.Jun.1; OpenEye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, NM 87508, USA
19. 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):1138–45. doi:10.1021/ci3001138.
20. Willet P, Barnard JM, Downs GM. Chemical similarity searching. *J Chem Inf Comput Sci.* 1998;38(6):983–96. doi:10.1021/ci9800211.
21. R: A language and environment for statistical computing; R Core Team; Foundation for Statistical Computing: Vienna, Austria, 2013
22. Stumpfe D, Bajorath J. Frequency of occurrence and potency range distribution of activity cliffs in bioactive compounds. *J Chem Inf Model.* 2012;52(9):2348–53. doi:10.1021/ci300288f.

Conclusions

We have analyzed activity cliffs for more than 600 target sets extracted from ChEMBL. Regardless of their molecular representation, more than 95% of activity cliffs showed an increase in ligand efficiency from the weakly to the highly potent cliff partner. However, when the ligand efficiency distribution was compared, MMP-cliffs had a larger average ligand efficiency increase than fingerprint-based activity cliffs. The increase in ligand efficiency could not be fully explained by changes in molecular weight or lipophilicity. This result provided further evidence for the preference of substructure-based representations of activity cliffs over fingerprint-based representations.

In the previous studies, we have analyzed the effect of MMPs on properties important for drug design. However, for MMPs to become more relevant in drug discovery, the chemical change they encode should be applicable to the synthetic modification of compounds. Still, this is not always the case. In order to increase the utility of MMPs for medicinal chemists, a novel fragmentation algorithm on the basis of retrosynthetic rules was designed and is presented in the next chapter.

4 Matched molecular pairs derived by retrosynthetic fragmentation

Introduction

MMPs can be used to analyze structural relations in a systematic manner. Nonetheless, the chemical transformation encoded in an MMP is in many cases not synthetically feasible. A very frequent transformation is the exchange of a hydrogen atom for a methyl group. However, the exchange of a methyl group with one specific hydrogen atom of the molecule, e.g. at an unreactive C-H bond, can be very challenging.¹¹¹ This limits the applicability of MMP-derived chemical transformations in compound optimization procedures. Here, we present a modification of the MMP fragmentation algorithm. Reactions encoded in the RETrosynthetic Combinatorial Analysis Procedure (RECAP) are used to guide the fragmentation step in the MMP generation, creating RECAP-MMPs. The distribution of RECAP-MMPs in public data is explored.

Reproduced from "A. de la Vega de León, J. Bajorath. Matched molecular pairs derived by retrosynthetic fragmentation. *MedChemComm* **2014**, 5, 64-67" with permission from The Royal Society of Chemistry.

CONCISE ARTICLE

Matched molecular pairs derived by retrosynthetic fragmentation

Cite this: *Med. Chem. Commun.*, 2014, 5, 64

Antonio de la Vega de León and Jürgen Bajorath*

Matched molecular pairs (MMPs) are defined as pairs of compounds that only differ by a chemical change at a single site. MMPs have become popular in medicinal chemistry to support lead optimization, absorption, distribution, metabolism, excretion, and toxicity (ADMET) analysis, and other applications. Thus far, MMPs have been algorithmically defined and not on the basis of reaction information. This often limits the chemical interpretability and practical utility of MMPs. Therefore, we introduce synthetically accessible MMPs that are automatically generated by applying reaction rules following the retrosynthetic combinatorial analysis procedure (RECAP). A library of more than 92 000 RECAP-MMPs was generated from public domain compounds active against 435 different targets exclusively utilizing high-confidence activity data. This library is made freely available for use in medicinal chemistry.

Received 10th September 2013
Accepted 27th October 2013

DOI: 10.1039/c3md00259d

www.rsc.org/medchemcomm

Introduction

MMPs have been introduced as pairs of compounds that only differ by a chemical change at a single site,^{1,2} a so-called chemical transformation.³ They are mostly generated by fragmentation³ or maximum common substructure-based^{1,4} algorithms. In recent years, MMPs have become popular tools in medicinal chemistry for a variety of applications^{5,6} including structure–activity relationship (SAR)^{7,8} and activity profile⁹ analysis, lead optimization,^{10,11} ADMET analysis,^{11–13} or the exploration of bioisosterism.¹⁴ A major reason for the attractiveness of the MMP concept in medicinal chemistry is that chemical transformations such as R-group replacements or core structure modifications can directly be associated with defined property changes (*e.g.*, activity, solubility, or stability) within the context of actual compounds,^{5,6} hence providing a basis for chemically intuitive analysis. By contrast, a shortcoming of current MMPs is that participating compounds are usually not related by chemical reactions. Hence, chemical transformations constituting MMPs are often not chemically interpretable and accessible, which limits their practical utility in medicinal chemistry, for example, when attempting to convert compounds into MMP partners with more favorable properties. Therefore, we introduce herein a new category of MMPs that are generated on the basis of retrosynthetic fragmentation employing the well-known RECAP reaction rules.¹⁵ Accordingly, these second-generation MMPs are termed RECAP-MMPs. The chemical transformation relating compounds forming RECAP-MMPs to each other results from a specific reaction. We show that

RECAP-MMPs are a subset of original MMPs, with very few exceptions, and generate a large library of RECAP-MMPs for 435 different compound classes exclusively utilizing high-confidence activity data. This library is made available to the scientific community without restrictions.

Methods

All activity classes from ChEMBL¹⁶ (release 15) were collected that contained at least 5 compounds with available (assay-independent) K_i values. Equilibrium constants were exclusively used to ensure high confidence of activity data.¹⁷ A total of 435 target-specific datasets comprising 40 650 unique active compounds were obtained. Compounds with multiple K_i values for the same target were only considered if all values fell within one order of magnitude. If this confidence criterion was met, the average value of independent measurements was used as the final potency annotation.

From each activity class, MMPs were systematically generated using an in-house implementation of the Hussain and Rea algorithm.³ Each compound was subjected to systematic single-, double-, and triple-cut fragmentation of all exocyclic single bonds between non-hydrogen atoms. During fragmentation, connectivity information was retained. Core structures and variable substituents resulting from fragmentation were stored in an index table as key and value fragments, respectively. Each pair of compounds having the same key and different value fragments formed an MMP. The size of a transformation was limited to a maximum of 13 non-hydrogen atoms and the size difference between exchanged fragments to 8 non-hydrogen atoms. In addition, keys were required to have at least twice the size of value fragments for each transformation. Application of these criteria yielded transformation size-restricted MMPs¹⁸ in

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany. E-mail: bajorath@bit.uni-bonn.de; Fax: +49-228-2699-341; Tel: +49-228-2699-306

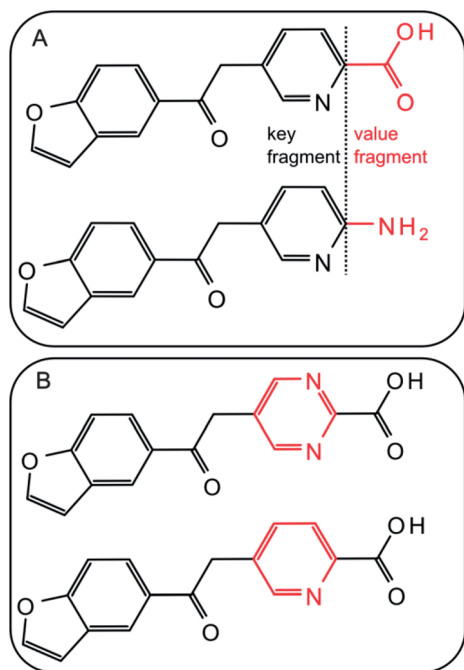


Fig. 1 MMPs. Two exemplary MMPs are shown. Exchanged fragments are highlighted in red.

which value fragments (substituents) were generally limited to relatively small substructures.¹⁸ Fig. 1 shows exemplary MMPs. In the following, MMPs generated by systematic fragmentation are referred to as “standard MMPs”.

For the generation of RECAP-MMPs, a RECAP rule-based fragmentation scheme was applied.^{15,19} Accordingly, bonds were only cut on the basis of retrosynthetic rules. In addition, a

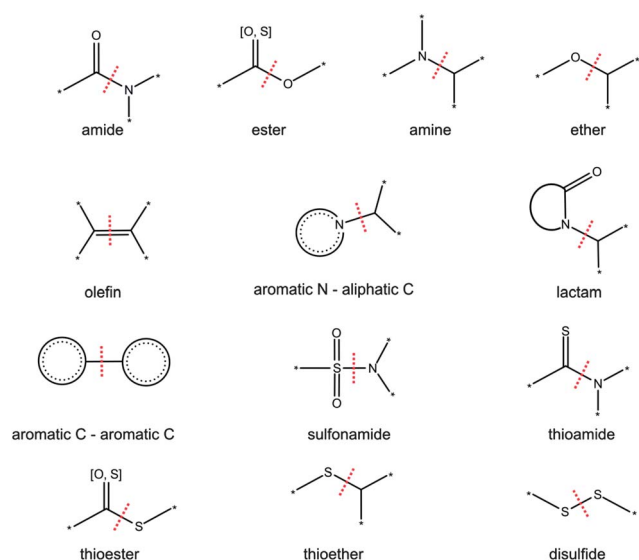


Fig. 2 RECAP rules. Thirteen retrosynthetic fragmentation rules are illustrated that were applied to generate RECAP-MMPs. The red line indicates the bond that is cut according to each reaction. In the case of amines, ethers, and thioethers the heteroatom should not be a part of any other functional group and not form exclusive bonds to multiple aromatic carbons.

transformation was only accepted if the two exchanged fragments were generated by the same reaction. Transformation size restrictions were applied as specified above. Original RECAP rules were slightly modified for single bond fragmentation. The urea and thiourea rules were not utilized because they affect multiple bonds. In addition, quaternary amines were not distinguished from non-charged amines. All applied retrosynthetic rules are reported in Fig. 2. RECAP-MMPs were systematically generated using in-house Java code and the Open Eye Toolkit.²⁰ For non-commercial applications source code is available upon request. Statistical analyses were carried out using R.²¹

Results and discussion

Standard versus RECAP-MMPs

As reported in Table 1, we obtained 435 K_1 -based datasets from ChEMBL with 40 650 compounds. From these compounds, we systematically generated standard MMPs and RECAP-MMPs. A total of 223 671 unique standard and 92 743 unique RECAP-MMPs were obtained. Many MMPs originated from multiple datasets. For 86 datasets, no RECAP-MMP was obtained, due to small compound numbers (on average, these 86 datasets contained only 10.6 compounds). The application of a confined set of retrosynthetic rules yielded fewer MMPs than systematic fragmentation, as expected. Surprisingly, however, nearly half as many RECAP-MMPs were obtained. Moreover, we found that essentially all RECAP-MMPs were reproduced by systematic fragmentation. Only 11 instances of RECAP-MMPs were detected that were not obtained by systematic fragmentation. An example is shown in Fig. 3. In this pair of compounds, qualifying exocyclic single bonds were absent. Hence, systematic fragmentation did not yield an MMP. Because RECAP-MMPs were a subset of standard MMPs, with only very few exceptions, 42% of all standard MMPs were conserved when reaction-based fragmentation was applied, a larger proportion than anticipated.

Chemical transformations

However, despite the high degree of MMP conservation, we generally observed that standard and RECAP-based transformations differed for a qualifying compound pair. Thus, although the same MMP was obtained on the basis of systematic or retrosynthetic fragmentation, the corresponding transformations were distinct. Examples are provided in Fig. 4. In general, RECAP-based transformations tended to be larger than

Table 1 Datasets and MMP statistics^a

Datasets	435
Compounds	40 650
Standard MMPs	223 671
RECAP-MMPs	92 734
Standard MMP cliffs	13 261
RECAP-MMP cliffs	4406
Standard MMP cliff frequency	5.9%
RECAP-MMP cliff frequency	4.8%

^a Statistics are reported for compound datasets, standard and RECAP-MMPs, and MMP cliffs.

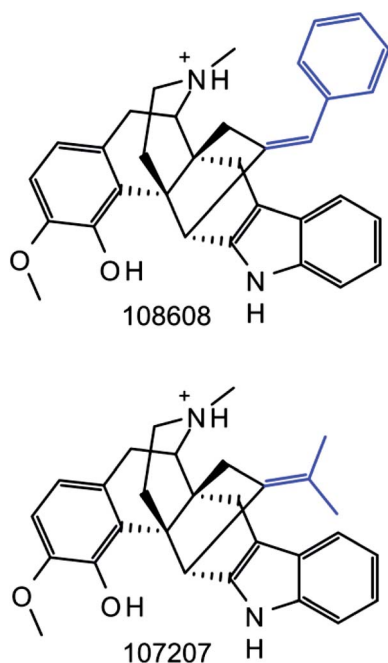


Fig. 3 Unique RECAP-MMP. Two compounds forming a RECAP-MMP are shown that was not generated by systematic fragmentation. RECAP-MMP value fragments are highlighted in blue. Compound ChEMBL IDs are given.

standard transformations, on average by 3–5 non-hydrogen atoms per MMP depending on the dataset. From RECAP transformations, reagents could often be deduced for the given reaction. By contrast, exchanges of small fragments in standard

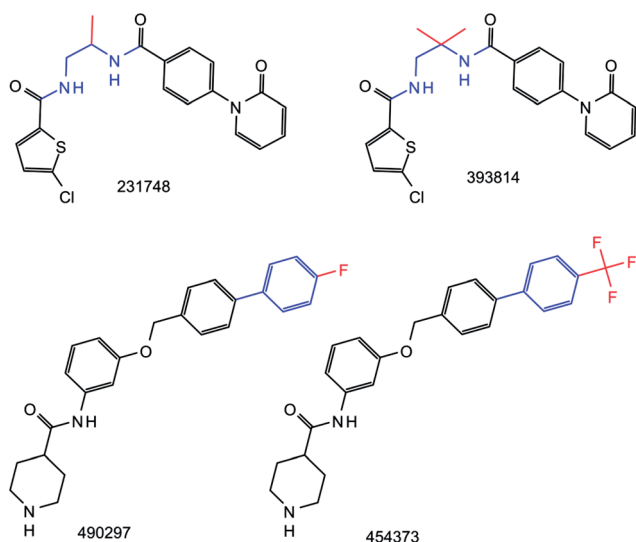


Fig. 4 Comparison of standard and RECAP-MMPs. Two pairs of compounds forming standard and RECAP-MMPs are shown. ChEMBL IDs are provided. Transformations in standard MMPs are highlighted in red and transformations in RECAP-MMPs in red and blue. The comparison illustrates that RECAP-based substructures representing a transformation were typically larger than substructures produced by systematic fragmentation. The RECAP-MMPs at the top and bottom were obtained through cuts of two amide bonds and an aromatic carbon–aromatic carbon bond, respectively.

MMPs were typically not interpretable in reaction terms. Thus, transformation information clearly distinguished RECAP-MMPs from standard MMPs.

Reaction distribution

Fig. 5 reports the fractions of RECAP-MMPs that were defined by specific retrosynthetic rules according to Fig. 2. Interestingly, no instances of RECAP-MMPs were detected that resulted from fragmentation of thioester and disulfide bonds, and thioamide bond cleavage accounted for less than 1% of all RECAP-MMPs. By contrast, amine and amide chemistry dominated the distribution of RECAP-MMPs, with 33% and 27%, respectively, followed by ethers (13%) and aromatic carbon–aromatic carbon bonds (10%), hence reflecting the current compound portfolio in medicinal chemistry.²² In addition, between 6% and 1% of RECAP-MMPs resulted from fragmentation of aromatic nitrogen–aliphatic carbon bonds, esters, lactams and olefins.

MMP cliffs

As an indicator of the SAR information content, we also determined the fraction of activity cliffs that were captured by standard and RECAP-MMPs, so-called MMP cliffs.¹⁸ Activity cliffs are generally defined as pairs of structurally similar or analogous compounds with a large difference in potency.²³ Therefore, all MMPs were determined in which the two compounds displayed a potency difference (K_i values) of at least two orders of magnitude.^{18,23} As reported in Table 1, the frequency of occurrence of standard MMP and RECAP-MMP cliffs was 5.9% and 4.8%, respectively. Thus, systematic and retrosynthetic fragmentation captured activity cliffs with similar frequency.

RECAP-MMP library

The 92 734 unique RECAP-MMPs identified in our study are made freely available as a machine-readable library organized on the basis of target sets (available at <http://www.limes.uni-bonn.de/forschung/abteilungen/Bajorath/labwebsite/downloads>). Given

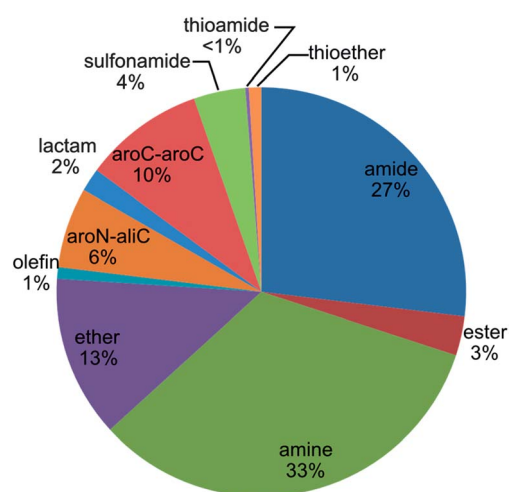


Fig. 5 Reaction frequency. The graph reports the proportions of RECAP-MMPs that were obtained on the basis of different retrosynthetic rules.

the target set organization, individual RECAP-MMPs might occur multiple times in different sets. This ensures that a complete set of RECAP-MMPs is available for each compound class. Furthermore, in the library, standard and retrosynthetic transformations are provided for each RECAP-MMP that was reproduced by systematic fragmentation to enable direct comparison of these transformations. Moreover, all RECAP-MMP cliffs are specified.

A randomly chosen sample of 50 RECAP-MMPs was traced back to compounds in original publications (via ChEMBL compound IDs) and it was examined whether the synthesis of these compounds was reported in the original publications. For more than 75% of these RECAP-MMPs, compounds were found to be synthesized by corresponding routes (in a number of original references, no compound synthesis was reported). Hence, in many cases, there was a direct link between RECAP-MMPs and synthetic routes of compounds from which these RECAP-MMPs originated.

Conclusions

Herein we have introduced second-generation MMPs defined on the basis of retrosynthetic rules and compared these RECAP-MMPs with standard MMPs. In RECAP-MMPs, chemical transformations are reaction-based and interpretable. Given the current popularity of the MMP concept, it is hoped that the library of RECAP-MMPs we provide will serve as a knowledge base to further improve the utility of matched molecular pairs in medicinal chemistry.

References

- 1 R. P. Sheridan, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 103–108.
- 2 P. W. Kenny and J. Sadowski, in *Cheminformatics in Drug Discovery*, ed. T. I. Oprea, Wiley-VCH, Weinheim, Germany, 2004, pp. 271–285.
- 3 J. Hussain and C. Rea, *J. Chem. Inf. Model.*, 2010, **50**, 339–348.
- 4 D. J. Warner, E. J. Griffen and S. A. St-Gallay, *J. Chem. Inf. Model.*, 2010, **50**, 1350–1357.
- 5 E. Griffen, A. G. Leach, G. R. Robb and D. J. Warner, *J. Med. Chem.*, 2001, **54**, 7739–7750.
- 6 A. M. Wassermann, D. Dimova, P. Iyer and J. Bajorath, *Drug Dev. Res.*, 2012, **73**, 518–527.
- 7 R. P. Sheridan, P. Hunt and J. C. Culberson, *J. Chem. Inf. Model.*, 2006, **46**, 180–192.
- 8 J. E. J. Mills, A. D. Brown, T. Ryckmans, D. C. Miller, S. E. Skerratt, C. M. Barker and M. E. Bunnage, *Med. Chem. Commun.*, 2011, **3**, 174–178.
- 9 Y. Hu and J. Bajorath, *ACS Med. Chem. Lett.*, 2011, **2**, 523–527.
- 10 P. J. Hajduk and D. R. Sauer, *J. Med. Chem.*, 2008, **51**, 553–564.
- 11 G. Papadatos, M. Alkarouri, V. J. Gillet, P. Willett, V. Kadirkamanathan, C. N. Luscombe, G. Bravi, N. J. Richmond, S. D. Pickett, J. Hussain, J. M. Pritchard, A. W. Cooper and S. J. Macdonald, *J. Chem. Inf. Model.*, 2010, **50**, 1872–1876.
- 12 A. G. Leach, H. D. Jones, D. A. Cosgrove, P. W. Kenny, L. Ruston, P. MacFaul, J. M. Wood, N. Colclough and B. Law, *J. Med. Chem.*, 2006, **46**, 6672–6682.
- 13 M. L. Lewis and L. Cuchurall-Sanchez, *J. Comput.-Aided Mol. Des.*, 2009, **23**, 97–103.
- 14 A. M. Wassermann and J. Bajorath, *Future Med. Chem.*, 2011, **3**, 425–436.
- 15 X. Q. Lewell, D. B. Judd, S. P. Watson and M. M. Hann, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 511–522.
- 16 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.
- 17 Y. Hu and J. Bajorath, *J. Chem. Inf. Model.*, 2012, **52**, 2550–2558.
- 18 X. Hu, Y. Hu, M. Vogt, D. Stumpfe and J. Bajorath, *J. Chem. Inf. Model.*, 2012, **52**, 1138–1145.
- 19 E. Lounkine and J. Bajorath, *J. Chem. Inf. Model.*, 2009, **49**, 162–168.
- 20 OpenEye Scientific Software Inc., Santa Fe, NM.
- 21 R Foundation for Statistical Computing, Vienna, Austria.
- 22 W. P. Walters, J. Green, J. R. Weiss and M. A. Murcko, *J. Med. Chem.*, 2011, **54**, 6405–6416.
- 23 D. Stumpfe and J. Bajorath, *J. Med. Chem.*, 2012, **55**, 2932–2942.

Conclusions

Novel MMPs have been developed on the basis of retrosynthetic fragmentation. In total, 13 different rules were implemented to recognize specific bonds in molecules. These rules were generated based on simple chemical reactions, such as an ester bond created from the condensation of an alcohol and a carboxylic acid. The distribution of RECAP-MMPs among compounds active against human targets was analyzed. Because of the more restrictive fragmentation, the number of RECAP-MMPs was less than half of the number of standard MMPs. Nonetheless, their SAR content, measured as activity cliff frequency, was very similar. More than half of the RECAP-MMPs found were generated because of nitrogen containing bonds such as amide bonds. The set of more than 92 000 unique RECAP-MMPs obtained was made publicly available.

Following development of RECAP-MMPs, novel applications of standard MMP relationships for drug discovery are explored. In the next chapter, a new methodology is introduced based on MMS to obtain preliminary SAR information for confirmed hit molecules. This information can be used to drive the optimization of a hit molecule.

5 Systematic identification of matching molecular series and mapping of screening hits

Introduction

MMS organize substructure relations on the basis of MMP sets. They can be rationalized as analog series and have been used to study SAR information in network representations.⁵² They have also been used to analyze SAR transfer.⁵³ In this study, MMS are systematically generated for bioactive compounds and their properties are explored. Confirmed hit compounds are mapped to MMS through MMP fragmentation in order to obtain initial SAR information. My main contribution to this study was the analysis of confirmed hit compounds and their mapping to MMS.

Reprinted with permission from "A. de la Vega de León, Y. Hu, J. Bajorath. Systematic identification of matching molecular series and mapping of screening hits. *Molecular Informatics* **2014**, 33(4), 257-263". Copyright 2014 John Wiley and Sons

Systematic Identification of Matching Molecular Series and Mapping of Screening Hits

Antonio de la Vega de León,^[a] Ye Hu,^[a] and Jürgen Bajorath^{*[a]}

Abstract: Matching molecular series (MMS) have originally been introduced as an extension of the matched molecular pair (MMP) concept to facilitate the design of substructure-based structure-activity relationship (SAR) networks. An MMP is defined as a pair of compounds that only differ by a structural change at a single site. In addition, an MMS is defined as an MMP-based series of compounds that have a conserved structural core and are distinguished by modifications at a single site. Systematic generation of MMS from specifically active compounds generalizes the search for series of structural analogs. Potency-ordered MMS pro-

vide series associated with SAR information. We have systematically extracted MMS from publicly available compounds with well-defined activity measurements and generated a large database with approx. 40 000 single- and 13 600 multi-target series, which provide a rich source of SAR information. As an application, we introduce MMP-based mapping of screening hits to MMS to search for initial SAR information and determine all SAR environments available for such hits. The MMS database is made freely available to the scientific community.

Keywords: Matching molecular series (MMS) · Structure-activity relationship (SAR) networks · Bioinformatics · Drug design · Computational chemistry

Matched molecular pairs (MMPs) are defined as pairs of compounds that only differ by the exchange of a substructure at a single site.^[1] The MMP concept is widely applied in medicinal chemistry^[2] to associate molecular property changes with defined structural modifications,^[2,3] study absorption, distribution, metabolism, and excretion (ADME) properties,^[3] or systematically analyze structure-activity relationship (SAR) information.^[4,5] MMPs can be algorithmically generated in an efficient manner,^[6,7] which enables large-scale analysis of compound structures and associated data. The MMP concept has been extended by introducing matching molecular series (MMS).^[8] An MMS is defined as a series of compounds forming pairwise MMP relationships. Hence, an MMS consists of compounds sharing the same structural core, a “key fragment” following MMP terminology,^[7] and varying substitutions (“values”) at a single site (i.e., exchanges of substructures). The MMS concept was originally introduced to facilitate the design of structure-activity relationship (SAR) network/graph representations in which similarity relationships between compounds were accounted for by MMS memberships.^[8] However, algorithmic generation of MMS can also be applied to generalize the search for series of structurally related compounds or analogs, as reported herein. In addition, ordering MMS compounds according to increasing potency often reveals SAR information.^[9] We have systematically searched public domain bioactive compounds with well-defined activity measurements for MMS, analyzed the identified MMS, and generated a comprehensive MMS database. As an exampla-

ry application, we introduce MMP-based mapping of screening hits to MMS to search for initial SAR information.

Compound data sets were assembled from ChEMBL^[10] release 17. Compound data available in ChEMBL are mostly extracted from medicinal chemistry literature. In this study, two types of potency measurements were separately considered, including assay-dependent IC_{50} values and assay-independent equilibrium constants (K_i values). From ChEMBL records, it can usually not be determined if K_i values were measured or calculated from IC_{50} values (which is frequently done using the Cheng–Prusoff estimation). Nonetheless, since IC_{50} and K_i values should not be directly compared, they are separately analyzed. In addition, only explicitly defined activity values for direct interactions with a specific human target at the highest level of confidence (with a ChEMBL confidence score of 9)^[10] were considered. All approximate potency annotations such as “>”, “<” or “~” were discarded. If one compound had more than one activity value for a given target, these values were required to fall within the same order of magnitude. Then, the geometric mean was calculated as the final potency annotation. On the basis of these selection criteria, a total of 661 K_i -

[a] A. de la Vega de León,[#] Y. Hu,[#] J. Bajorath
Department of Life Science Informatics, Bonn-Aachen
International Center for Information Technology, Rheinische
Friedrich-Wilhelms-Universität Bonn
Dahlmannstr. 2, D-53113 Bonn (Germany)
tel: +49-228-2699-306; fax: +49-228-2699-341
*e-mail: bajorath@bit.uni-bonn.de

[#] The contributions of these authors should be considered equal.

based compound data sets were obtained that contained more than 45 000 compounds with more than 77 000 potency annotations. In addition, 1203 IC_{50} -based data sets were assembled that contained more than 95 000 compounds with more than 135 000 potency annotations. The compound data sets are summarized in Table 1. These data sets were systematically searched for MMS.

Table 1. Compound data sets and MMS. For the K_i - and IC_{50} -based data sets from ChEMBL (release 17), the numbers of targets, compounds, and corresponding potency measurements are reported. In addition, the total number of target-based MMS, unique MMS, and targets for which MMS were obtained are provided. Furthermore, the number (and ratio) of MMS that were associated with single- or multi-target activities are given.

Number of	ChEMBL	
	K_i	IC_{50}
Targets	661	1203
Compounds	45 353	95 685
Potency measurements	77 421	135 291
Target-based MMS	30 452	45 607
Unique MMS	19 427	35 627
Targets with MMS	406	790
Single-target MMS	12 755 (65.7%)	28 080 (79.6%)
Multi-target MMS	6 672 (34.3%)	7 187 (20.4%)

From the PubChem BioAssay database (accessed August 20th, 2012),^[11] all confirmatory assays that corresponded to targets of our ChEMBL data sets were identified. In total, 241 confirmatory assays were obtained for 88 different targets. From these 241 assays, all confirmed active compounds with explicitly defined IC_{50} measurements were taken and searched against the ChEMBL database. A total of 3123 screening hits from the PubChem assays were not detected in ChEMBL. For these hits, a total of 5182 IC_{50} measurements were available (Table 2). The screening hits were then mapped to MMS, as described below.

The selected ChEMBL compounds were systematically fragmented using an in-house implementation of the algorithm by Hussain and Rea^[7] utilizing the OEChem toolkit.^[12]

All exocyclic single bonds and all possible combinations of two or three bonds in a compound were cleaved in subsequent fragmentation trials. Accordingly, the MMP frag-

Table 2. PubChem assay data and hits. The number of confirmatory assays taken from the PubChem BioAssay database (accessed August 20th, 2012) and the number of different targets these assays covered are reported. In addition, the number of confirmed hits reported to be active in at least one assay that were not found in ChEMBL (release 17) and the total number of activity measurements associated with these hits are reported.

Number of	PubChem
Assays	241
Targets	88
Confirmed hits	3123
Activity measurements	5182

mentation scheme differed from the generation of Bemis-Murcko scaffolds^[13] that are extracted from molecules by removing all R-groups at once and retaining ring systems and linkers between rings. An index table was created using the key fragments to organize all associated value fragments. Indexing was limited to keys that consisted of at least twice the size of corresponding values and to values with no more than 13 non-hydrogen atoms. In addition, the difference in the size between the exchanged value fragments was limited to at most eight non-hydrogen atoms.^[14] These restrictions ensured that values represented structural changes of relatively small size compared to keys (core structures).^[14] The index table contained all MMPs formed by pairs of ChEMBL compounds yielding the same key and different values.

In the index table, MMS were identified that consisted of a common key and at least three different values (i.e., three structurally related compounds). Compounds forming an MMS were ordered by increasing potency.

PubChem screening hits were also subjected to systematic fragmentation, as described above, and the resulting key fragments were searched against the MMS keys. If a match was detected, the PubChem hit was assigned to the MMS as an extension.

MMS are formed on the basis of systematically detected MMP relationships between specifically active compounds, as illustrated in Figure 1. By design, MMS comprehensively account for all possible structural relationships and include classical analog series, as shown in Figure 1, and also series with site-specific modifications in core structures (depending on the fragmentation scheme). Hence, the MMS concept represents a generalized compound series format that retrospectively accounts for all detectable pairwise structural relationships in data sets and organizes compounds according to these structural relations in a consistent manner. This is different from combinatorially decorated scaffolds where a series of previously chosen scaffolds are prospectively explored with defined structural permutations or chemical modifications.^[15] If complemented with activity information, MMS can be utilized for SAR data mining and analysis, which is facilitated by potency-based ordering of compounds within series, as also illustrated in Figure 1. Given the general applicability of the MMS concept, we have set out to determine all MMS comprising at least three currently available bioactive compounds with defined activity measurements and target annotations.

From ChEMBL, 661 K_i and 1203 IC_{50} value based data sets were assembled that contained a total of more than 133 000 compounds (Table 1). Each data set consisted of compounds active against a specific target. The K_i - and IC_{50} -based sets were separately searched for MMS to avoid the identification of series comprising compounds with different types of potency measurements that cannot be directly compared.

As reported in Table 1, 30 452 and 45 607 MMS were identified in 406 K_i - and 790 IC_{50} -based data sets, respec-

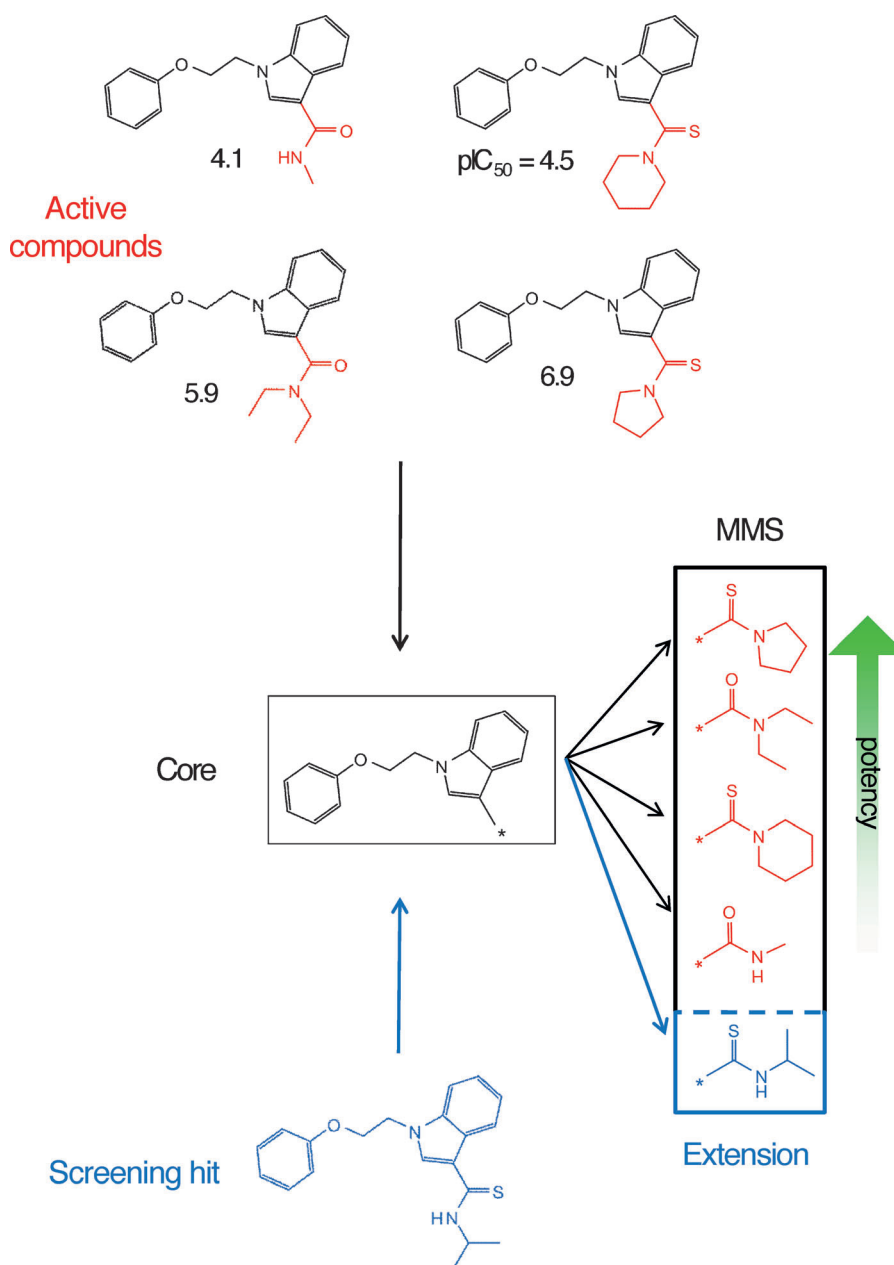


Figure 1. Exemplary MMS and its extension. Four inhibitors of protein-tyrosine phosphatase LC-PTP are shown that form pairwise MMP relationships and thus an MMS. Their common structural core (key fragment, black) is displayed and distinguishing substituents (values, red) are ordered according to increasing compound potency. The MMS is extended by mapping a screening hit (blue) that also forms MMP relationships with all compounds of this series.

tively. Thus, MMS were found in ~64% of all data sets, providing broad target coverage. Because a given MMS might be present in different data sets, we determined the total number of unique series. As reported in Table 1, 19427 and 35627 unique MMS were detected in the K_i - and IC_{50} -based sets, respectively, thus providing a large database of series for SAR exploration. The majority of these MMS was associated with single-target activities, but a significant proportion of series consisted of multi-target MMS. In the K_i -based sets, 6672 multi-target MMS were present (~34% of all

MMS) and in the IC_{50} -based sets, 7187 (~20%) multi-target MMS (Table 1). Interestingly, K_i -based sets contained a higher proportion of multi-target MMS than IC_{50} -based sets.

We then determined the composition and size distribution of MMS. Figure 2a reveals that MMS from both K_i - and IC_{50} -based sets yielded a similar distribution of core structure sizes (with the majority of key fragments consisting of 21 to 30 non-hydrogen atoms). In addition, MMS from both sets also consisted of very similar numbers of compounds

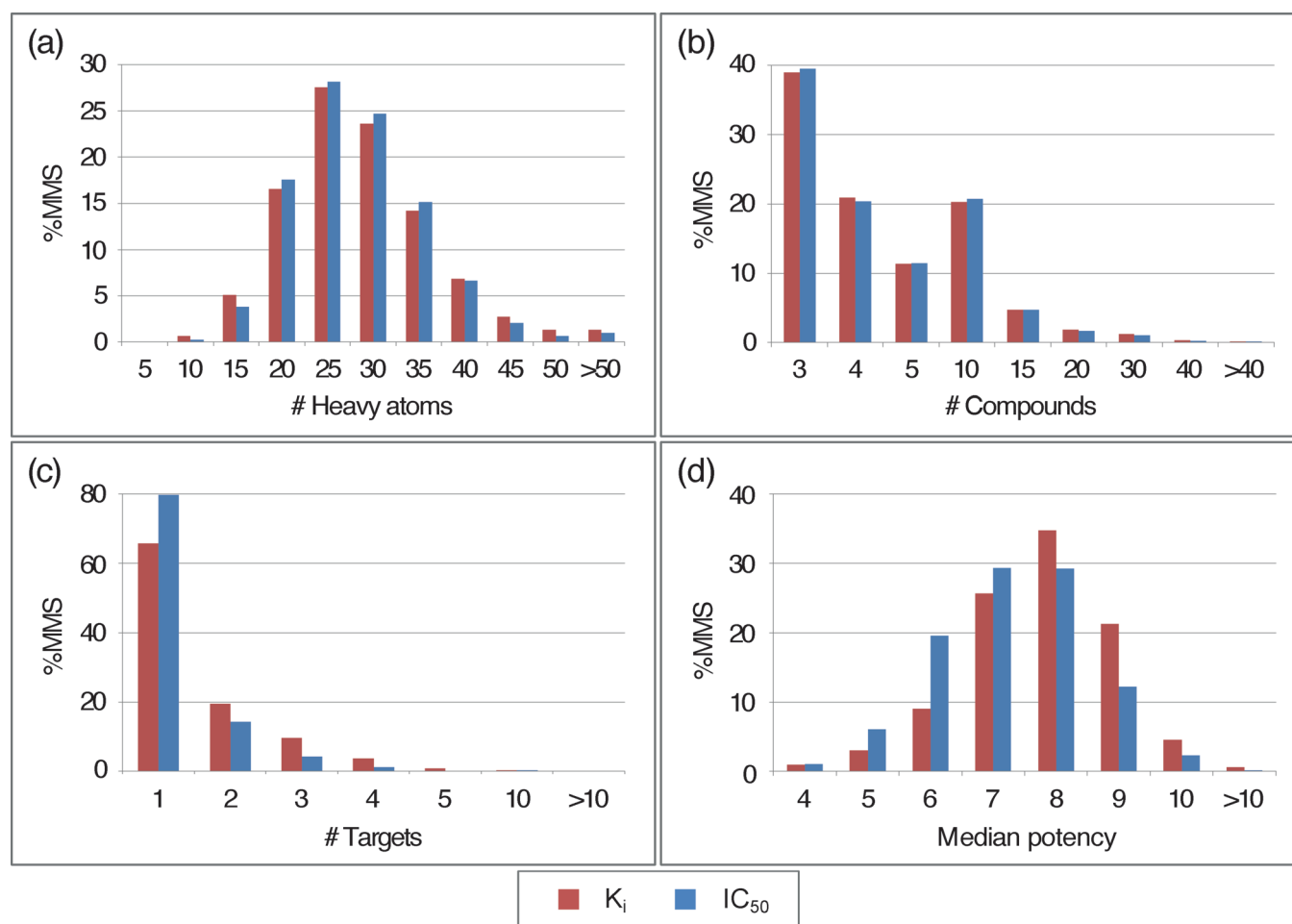


Figure 2. Characterization of MMS. Reported are the distributions of the number of (a) non-hydrogen atoms of key fragments, (b) compounds, and (c) targets over MMS as well as (d) the median potency values for MMS from the K_i - (red) and IC_{50} -based (blue) subsets.

(Figure 2b). Approx. 40% of MMS from both sets consisted of three compounds and ~50% of four to 10 compounds. Moreover, ~8% of all MMS contained 11 to 20 compounds and individual series with 40 or more compounds were also detected. Nearly 10% of all MMS comprised 10 or more compounds.

The target distribution of MMS is reported in Figure 2c. The majority of MMS was associated with single-target activities (see also Table 1). Most multi-target MMS were active against two to four targets. In Figure 2d, the distribution of median potency values of MMS is reported, revealing that most series contained compounds active in the nanomolar range, regardless of the type of potency measurements, which further emphasized the relevance of MMS for SAR analysis.

MMS can also be utilized for compound mapping, as illustrated in Figure 1 (bottom). On the basis of MMP calculations, test compounds can be searched against MMS to identify series that test compounds further extend (Figure 1). Compound mapping can be carried out for different purposes. For example, hits from screening campaigns might be searched against MMS to determine if hits

further extend MMS sharing the same activity. In this case, at least preliminary SAR information has been obtained for a given hit. Moreover, if hits are found to further extend MMS with a different activity, an additional activity hypothesis can be explored.

To illustrate the underlying idea we have systematically searched 3123 confirmed screening hits from PubChem against our MMS database that had reported activity against targets also contained in ChEMBL. A total of 40 hits were found to map to existing MMS sharing the same activity. As reported in Table 3, these 40 hits further extended 28 MMS from IC_{50} -based compound sets that were active

Table 3. Mapping of screening hits to MMS. The number of screening hits from PubChem (accessed August 20th, 2012) that extended existing MMS, their number, and targets are reported. All screening hits were found to map to IC_{50} -based MMS.

Number of	IC_{50}
Hits	40
MMS	28
Targets	15

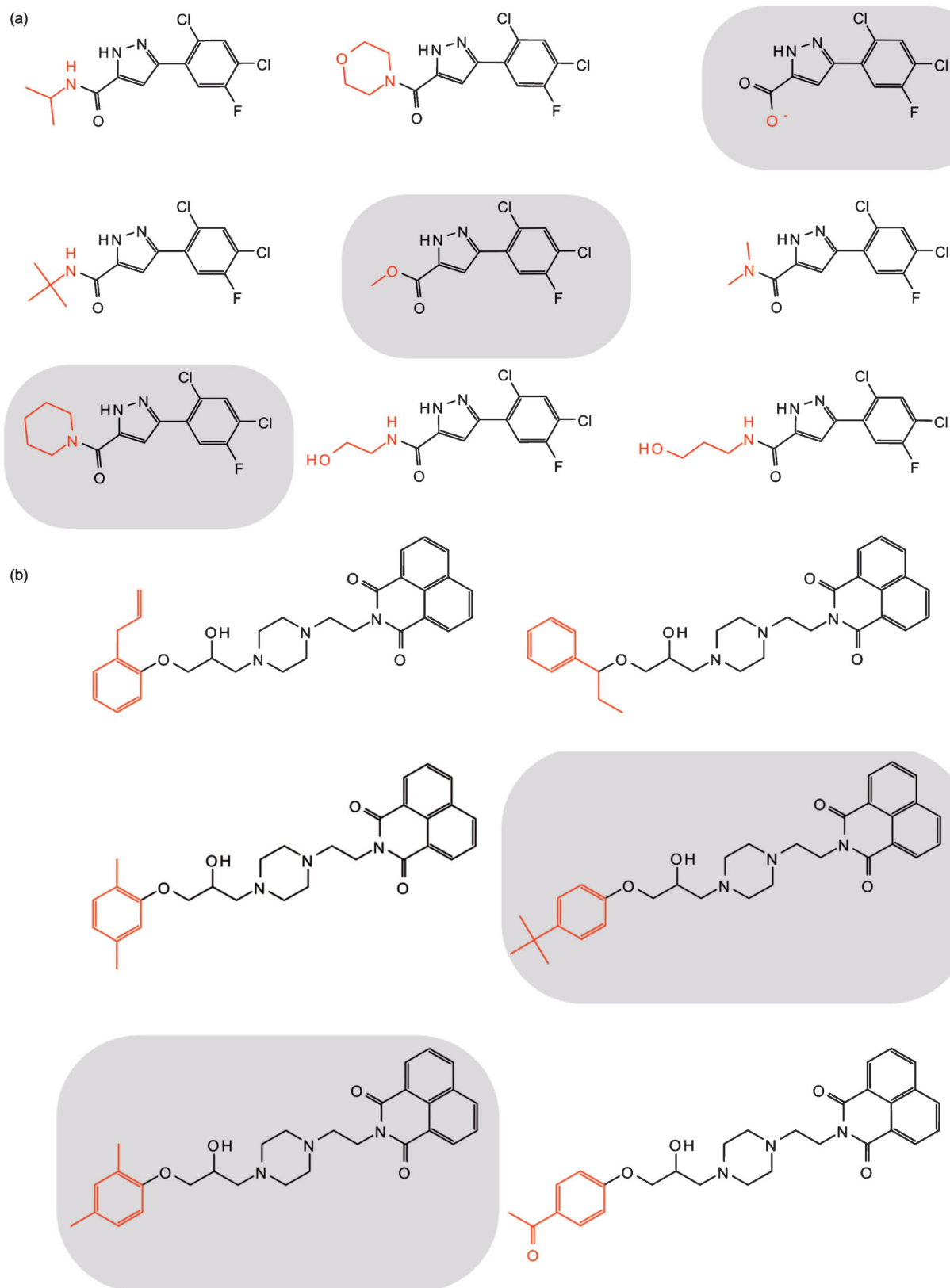


Figure 3. Extended MMS. Two exemplary MMS are shown together with mapped screening hits including inhibitors of (a) alkaline phosphatase and (b) ubiquitin-conjugating enzyme. Hits extending each MMS are displayed on a gray background. Structural differences between compounds (value fragments) are highlighted in red. From the top (left) to the bottom (right) compounds are arranged in the order of increasing potency.

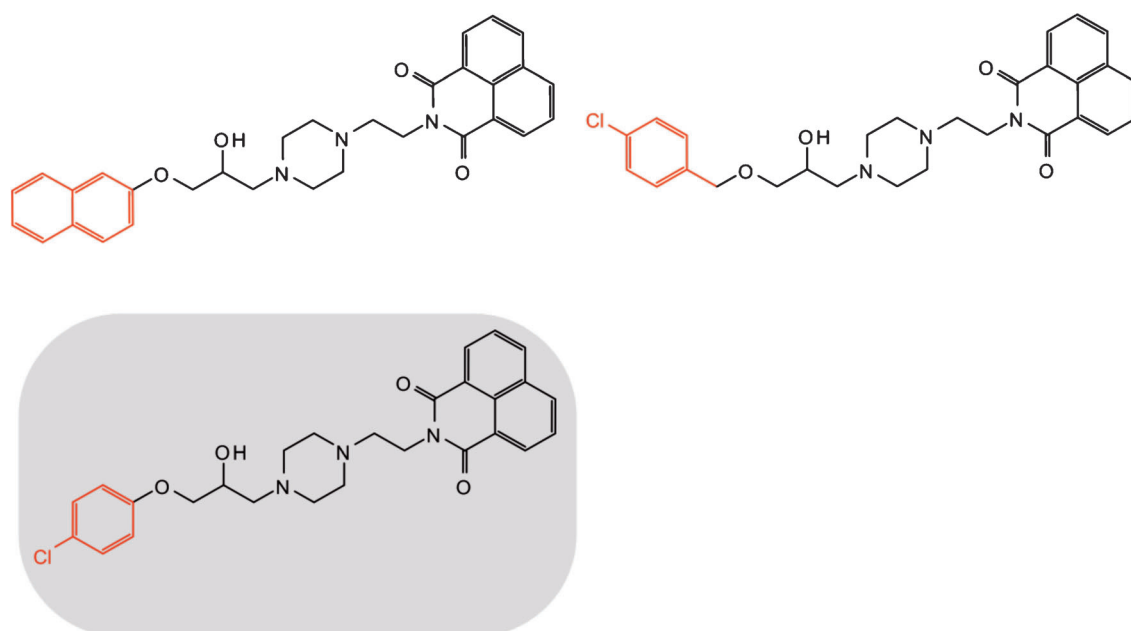


Figure 3. (Continued)

against 15 different targets. We found that 30 of these hits mapped to a single MMS, whereas the remaining 10 hits mapped to two or more target-based MMS. In addition, five of these hits further extended multi-target MMS, hence providing attractive activity hypotheses for further investigation.

In Figure 3, exemplary MMS are shown that were further extended by multiple screening hits sharing the same activity. In Figure 3a, an MMS comprising six alkaline phosphatase inhibitors is shown that was extended by three confirmed hits from a screen against this enzyme. In Figure 3b, an MMS consisting of six inhibitors of ubiquitin-conjugating enzyme is shown to which three confirmed hits mapped. Both MMS represent classical analog series and illustrate how mapping of screening hits complements available SAR information.

In conclusion, we have reported a systematic search for matching molecular series in target-based sets of bioactive compounds with well-defined activity data. Different types of potency measurements were separately considered. In total, approx. 53 000 MMS with activity against 877 different targets were identified and characterized. Potency-based ordering of compounds forming MMS renders these series attractive for SAR analysis. The series we identified included a significant proportion of multi-target MMS. We have also introduced an MMP-based strategy for compound mapping to MMS that can be utilized in different ways. As an exemplary application, it has been shown how screening hits can be mapped to MMS to search for initial SAR information and further extend existing series. Our large database of potency-ordered MMS has been made freely available^[16] as a resource for exploring single- and multi-target SARs,

compound mapping, and other medicinal chemistry applications.

Conflict of interests

No conflict of interests declared.

Acknowledgement

We are grateful to *OpenEye Scientific Software, Inc.*, for the free academic license of the OpenEye Toolkits.

References

- [1] P. W. Kenny, J. Sadowski, *Cheminform. Drug Discov.* **2004**, 271–285.
- [2] E. Griffen, A. G. Leach, G. R. Robb, D. J. Warner, *J. Med. Chem.* **2011**, *54*, 7739–7750.
- [3] A. G. Leach, H. D. Jones, D. A. Cosgrove, P. W. Kenny, L. Ruston, P. MacFaul, J. M. Wood, N. Colclough, B. Law, *J. Med. Chem.* **2006**, *49*, 6672–6682.
- [4] J. E. J. Mills, A. D. Brown, T. Ryckmans, D. C. Miller, S. E. Skerratt, C. M. Barker, M. E. Bunnage, *Med. Chem. Commun.* **2012**, *3*, 174–178.
- [5] A. M. Wassermann, P. Haebel, N. Weskamp, J. Bajorath, *J. Chem. Inf. Model.* **2012**, *52*, 1769–1776.
- [6] D. J. Warner, E. J. Griffen, S. A. St-Gallay, *J. Chem. Inf. Model.* **2010**, *50*, 1350–1357.
- [7] J. Hussain, C. Rea, *J. Chem. Inf. Model.* **2010**, *50*, 339–348.
- [8] M. Wawer, J. Bajorath, *J. Med. Chem.* **2011**, *54*, 2944–2951.
- [9] A. M. Wassermann, J. Bajorath, *J. Chem. Inf. Model.* **2011**, *51*, 1857–1866.

- [10] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.
- [11] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte, S. H. Bryant, *Nucleic Acids Res.* **2012**, *40*, D400–D412.
- [12] *OEChem*, Version 1.7.7, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, **2012**; <http://www.eyesopen.com>.
- [13] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.
- [14] X. Hu, Y. Hu, M. Vogt, D. Stumpfe, J. Bajorath, *J. Chem. Inf. Model.* **2012**, *52*, 1138–1145.
- [15] A. R. Katritzky, J. S. Kiely, N. Hebert, C. Chassaing, *J. Comb. Chem.* **2000**, *2*, 2–5.
- [16] A. de la Vega de León, Y. Hu, J. Bajorath, *Data Sets of Matching Molecular Series*; ZENODO.10.5281/zenodo.8342. <https://zenodo.org/record/8342>

Received: February 13, 2014

Accepted: February 18, 2014

Published online: March 13, 2014

Conclusions

We have systematically generated potency-ordered MMS from bioactive compounds. 53 000 unique MMS were generated from 133 000 compounds on the basis of K_i and IC_{50} activity data. The properties of MMS for K_i and IC_{50} data were very similar. MMS were found that contained large numbers of compounds or that were annotated with activity against more than one target. Confirmed hits from 241 confirmatory assays were fragmented and mapped to previously generated MMS. Several hit compounds with different activities were mapped. Preliminary SAR information was obtained for mapped screening hits that could inform compound optimization efforts. The set of potency-ordered MMS was made available to the medicinal chemistry community for compound mapping applications.

We have presented a method to obtain SAR information from potency-ordered MMS to aid compound optimization efforts. Activity comparisons between chemically related compounds provide initial SAR information. However, several hit compounds could not be mapped to any MMS and no SAR information was present for these molecules. In the next study, we developed a predictive methodology based on MMP representations and SVR that can predict the difference in activity between MMP partners. With this methodology, potency difference caused by chemical transformations not found in the data can be estimated.

6 Prediction of compound potency changes in matched molecular pairs using support vector regression

Introduction

Activity prediction is the goal of QSAR. Traditionally, activity prediction was confined to structurally related compounds sets. However, in order to predict activity from heterogeneous sets of compounds, methods from machine learning have been applied in chemoinformatics. These methods, such as SVR, are thought to be able to model complex nonlinear relationships between structure and activity. Thus far, MMPs have rarely been used in machine learning. A previous application accurately predicted the presence of MMP-cliffs.⁷¹ In this study, we expand on this effort by using SVR and MMP-based kernel functions to predict the activity change between compounds forming an MMP. Different kernel functions that represent only the chemical transformation (transformation kernels) or that combine the key fragment with the transformation (MMP kernels) are compared.

Reprinted with permission from "A. de la Vega de León, J. Bajorath. Prediction of compound potency changes in matched molecular pairs using support vector regression. *Journal of Chemical Information and Modeling* **2014**, 54(10), 2654-2663".
Copyright 2014 American Chemical Society

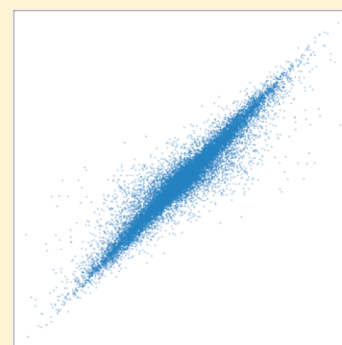
Prediction of Compound Potency Changes in Matched Molecular Pairs Using Support Vector Regression

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, Dahlmannstr. 2, D-53113 Bonn, Germany

ABSTRACT: Matched molecular pairs (MMPs) consist of pairs of compounds that are transformed into each other by a substructure exchange. If MMPs are formed by compounds sharing the same biological activity, they encode a potency change. If the potency difference between MMP compounds is very small, the substructure exchange (chemical transformation) encodes a bioisosteric replacement; if the difference is very large, the transformation encodes an activity cliff. For a given compound activity class, MMPs comprehensively capture existing structural relationships and represent a spectrum of potency changes for structurally analogous compounds. We have aimed to predict potency changes encoded by MMPs. This prediction task principally differs from conventional quantitative structure–activity relationship (QSAR) analysis. For the prediction of MMP-associated potency changes, we introduce direction-dependent MMPs and combine MMP analysis with support vector regression (SVR) modeling. Combinations of newly designed kernel functions and fingerprint descriptors are explored.

The resulting SVR models yield accurate predictions of MMP-encoded potency changes for many different data sets. Shared key structure context is found to contribute critically to prediction accuracy. SVR models reach higher performance than random forest (RF) and MMP-based averaging calculations carried out as controls. A comparison of SVR with kernel ridge regression indicates that prediction accuracy has largely been a consequence of kernel characteristics rather than SVR optimization details.



INTRODUCTION

The prediction of changes in compound potency as a consequence of chemical modifications typically falls into the domain of classical quantitative structure–activity relationship (QSAR) analysis.¹ However, QSAR modeling is usually limited to congeneric compound series¹ and cannot be systematically applied to large and structurally diverse compound data sets. In addition, conventional QSAR predictions are based upon linear regression models. To account for the nonlinearity of many SARs, machine learning approaches such as random forest² (RF) and support vector regression³ (SVR) analysis have gained popularity in recent years. In addition to potency prediction,^{4–6} these methods have also been applied, for example, to predict ADME properties^{7,8} or toxicology end points.^{9–11}

The congeneric compound series constraint of standard QSAR can be addressed by considering alternative structural representations, which can also be combined with nonlinear prediction methods. For example, an attractive opportunity to further extend potency predictions to large and heterogeneous compound data sets is provided by the matched molecular pair (MMP) formalism.¹² An MMP is defined as a pair of compounds that only differ by a structural change at a single site.¹² This modification is accounted for by the exchange of a pair of substructures termed a chemical transformation. MMPs can be systematically generated for a given compound data set, which reveals all possible pairs of structural analogs and comprehensively captures structural relationships present within the set. Then, it can be attempted to predict changes

in potency or other chemical properties associated with chemical transformations at the level of compound pairs, rather than series. For example, property value changes in multiple MMPs have been used to predict value changes associated with equivalent transformations.^{13,14} However, such MMP-based extrapolations cannot be generalized and have often limited statistical significance.¹⁴

Accordingly, first attempts have been made to combine MMP analysis with machine learning, for example, by predicting changes in potency and ADME properties using RF calculations¹⁵ or by predicting activity cliffs (i.e., pairs of structurally analogous compounds having a large difference in potency)¹⁶ using support vector machines (SVMs).^{17,18} SVM-based activity cliff prediction has distinguished pairs of compounds forming activity cliffs from others,¹⁸ without predicting numerical potency differences. In addition to using SVM models for classification and ranking, prediction of numerical potency (difference) values encoded by MMPs can be attempted via SVR.

Herein, we combine MMP analysis with SVR and systematically predict potency difference values for MMPs using kernel functions of different design. Combinations of different kernel functions and fingerprint descriptors used as molecular fragment representations are explored, and preferred combinations are identified. In calculations on a variety of compound data sets, preferred kernel-fingerprint combinations yield high

Received: July 4, 2014

Published: September 5, 2014

SVR accuracy in predicting numerical potency differences (and reach higher accuracy than RF calculations). The MMP-based SVR methodology introduced herein is generally applicable for numerical potency predictions.

MATERIALS AND METHODS

Compound Data Sets. From ChEMBL (version 17),¹⁹ 17 compound activity classes including a variety of targets were selected. For each class, all compounds having defined K_i values (with activity relation “=”, assay confidence score 9, the highest possible score, and target relationship “D” indicating “direct” relationships) were collected. Compounds having multiple potency values that differed by more than 1 order of magnitude (considering the highest and lowest values) were discarded. For compounds with multiple activity values falling within 1 order of magnitude range, the arithmetic mean was calculated as the final potency annotation. The compound data sets contained between 1200 and 2500 compounds, as reported in Table 1. From these compound data sets, between 5700 and 32 000 direction-dependent MMPs were obtained, as also reported in Table 1.

Table 1. Compound Data Sets^a

	ID	name	abbreviation	Cpds	MMPs
A	205	carbonic anhydrase II	CA2	1566	8248
B	214	serotonin 1a receptor	5-HT1A	1276	9352
C	217	dopamine D2 receptor	DRD2	1916	17 630
D	218	cannabinoid CB1 receptor	CB1	1673	16 004
E	226	adenosine A1 receptor	ADORA1	2107	24 534
F	228	serotonin transporter	SHTT	1317	9352
G	233	mu opioid receptor	MOR1	1447	15 712
H	234	dopamine D3 receptor	DRD3	1332	9532
I	237	kappa opioid receptor	KOR-1	1302	17 654
J	251	adenosine A2a receptor	ADORA2A	2538	32 086
K	253	cannabinoid CB2 receptor	CB2	1903	18 548
L	256	adenosine A3 receptor	ADORA3	2037	22 316
M	259	melanocortin receptor 4	MC4R	1209	28 274
N	261	carbonic anhydrase I	CA1	1528	7,804
O	264	histamine H3 receptor	H3R	1,849	19 256
P	3371	serotonin 6 receptor	5-HT6	1291	9648
Q	3594	carbonic anhydrase IX	CA9	1220	5756

^aFor each data set, the number of compounds and direction-dependent MMPs is given. In addition, the ChEMBL identifier (ID), the target name, and abbreviation are provided. Data sets are labeled A–Q.

Direction-Dependent Matched Molecular Pairs.

Matched molecular pairs were systematically calculated for data set compounds using an in-house Java implementation of the algorithm developed by Hussain and Rea²⁰ based upon the OEChem Toolkit²¹ from OpenEye. Single-, dual-, and triple-cut fragmentation of exocyclic bonds was carried out generating conserved key and variable value fragments stored in an index table.²⁰ In addition, transformation size restrictions²² were applied to ensure a meaningful distinction between core structures (key) and substituents (values). Accordingly, a key fragment was required to consist of at least twice the number of non-hydrogen atoms as a value fragment; a value fragment was permitted to contain a maximum of at most 13 non-hydrogen atoms, and the size difference between values of a given key was

limited to at most eight non-hydrogen atoms.²² If two compounds formed several MMPs, the one having the largest key fragment was selected.

For each MMP, the potency difference between the participating compounds was recorded in a direction-dependent manner, as illustrated in Figure 1A. Thus, for each original MMP, two direction-dependent MMPs were generated encoding a potency-decreasing and a potency-increasing transformation (i.e., value fragment 1 → 2 vs 2 → 1).

Molecular Representation. For each direction-dependent MMP, five fingerprint representations were calculated, as also illustrated in Figure 1A, including a fingerprint of the key fragment (KeyFP), fingerprint of the value fragments (V1FP and V2FP, respectively), a fingerprint containing bits shared by V1FP and V2FP (CommV1V2FP), and two value fragment difference fingerprints (V1FP-V2FP and V2FP-V1FP, respectively). For keyed fingerprint descriptors such as MACCS²³ (in which each bit position is assigned to a specific structural fragment or pattern), a difference fingerprint of size $2n$ was calculated from value fingerprints of size n by merging the value fingerprints with uniquely set bit positions. For hashed fingerprints such as extended connectivity fingerprints (ECFPs),²⁴ which represent feature sets, a difference fingerprint was generated by combining unique features of the first value and unique features of the second value with inverted sign. These types of difference fingerprints accounted for direction-dependent transformations. To implement these five fingerprint designs, different fingerprint descriptors were used including ECFPs of bond diameter 2, 4, and 6 as well as MACCS structural keys (166 bit version). Fingerprint calculations were carried out using in-house Python scripts based upon the OEChem Toolkit.²¹

Support Vector Regression. SVR is a nonlinear prediction method and variant of SVM classification that maps data points into higher-dimensional chemical (descriptor) reference spaces with the aid of kernel functions. A linear regression is performed in the high dimensional space. More formally, the predicted value y of the input vector x is calculated as

$$y = K(w, x) + b$$

$K(w, x)$ is the kernel function that maps x into the high-dimensional space; w (the normal weight vector) and b (the bias) determine a hyperplane in this space to separate positive and negative training instances and are estimated by minimizing the so-called structural risk:

$$R_{\text{SVR}}(C) = C \frac{1}{n} \sum_n^{i=1} L_\epsilon(d_i, y_i) + \frac{1}{2} \|w\|^2$$

$$L_\epsilon(d_i, y_i) = \begin{cases} |d - y| - \epsilon & |d - y| \geq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

R_{SVR} represents the risk function and depends on a loss function L_ϵ and a regularization parameter depending on w . L_ϵ measures the difference between the predicted value y and observed value d and is only applied when the difference is larger than ϵ . C is a factor determining the trade-off between the loss function and the regularization term.

Kernel Functions. For SVR, six alternative kernels were designed and investigated. All kernels were based on the Tanimoto similarity function²⁵ but utilized different types of fingerprint representations. The design of these kernels was

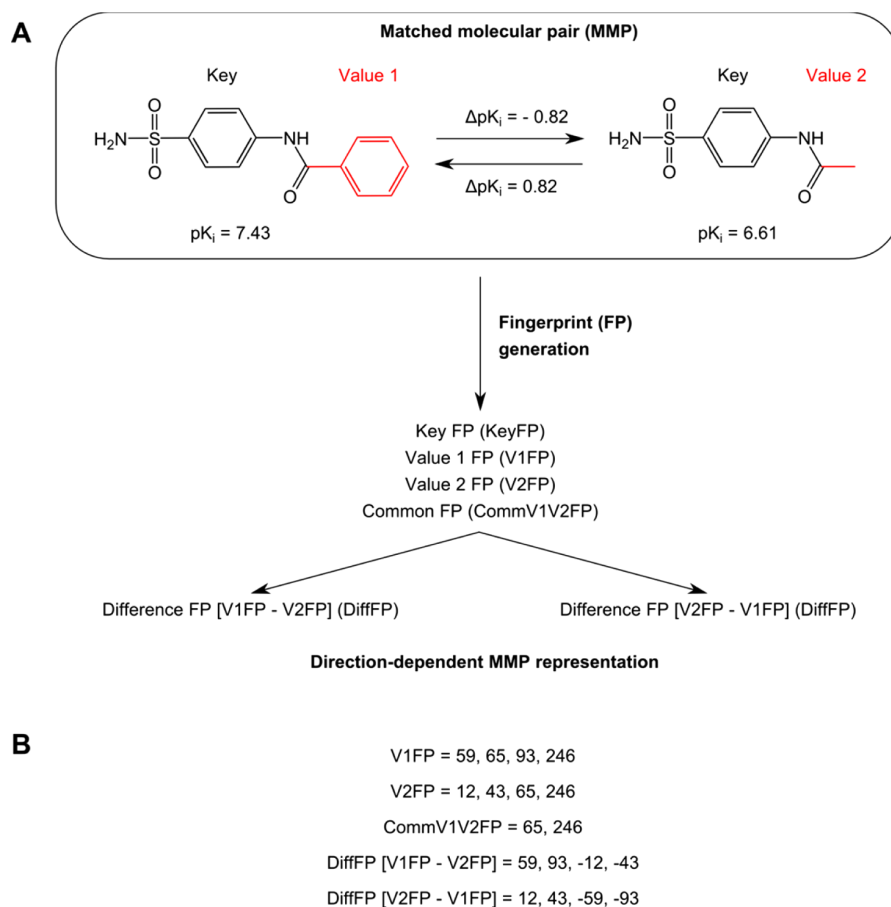


Figure 1. Molecular representation. (A) At the top, an exemplary MMP is shown. The shared core structure (key fragment) is colored black, and the distinguishing value fragments (value 1 and value 2) are colored red. For potency prediction, the potency difference between MMP compounds is monitored in a direction-dependent manner. Accordingly, two direction-dependent MMPs are obtained encoding a potency-decreasing (value 1 → 2) and a reverse potency-increasing (value 2 → 1) transformation. For each MMP, a fingerprint of the key fragment (KeyFP) and a fingerprint of its value fragments (V1FP and V2FP, respectively) are calculated. From V1FP and V2FP, a common fingerprint (CommV1V2FP) is generated consisting of shared bit positions. In addition, two difference fingerprints (DiffFP) are calculated from V1FP and V2FP to yield a direction-dependent transformation representation. For direction-dependent MMPs, KeyFP and CommV1V2FP are conserved. (B) Schematic representation of feature set fingerprints of two value fragments and of CommV1V2FP and DiffFP derived from these feature sets. The derivation of CommV1V2FP and DiffFP is described in detail in the text.

inspired by MMP-based kernel functions successfully used for the prediction of activity cliffs.¹⁸

Transformation kernels only utilized value fragment-based fingerprints and represented a chemical transformation in three different ways, based upon

- (i) only the difference fingerprint (DiffFP, Figure 1B)—this kernel was termed 1VD
- (ii) both CommV1V2FP and DiffFP (2VCD)
- (iii) the two value fragments fingerprints (2V12)

In addition, *MMP kernels* were constructed by adding the key fingerprint representation to i–iii producing kernels 2VKD, 3VKCD, and 3VK12, respectively.

These six kernels were implemented using the Tanimoto coefficient (Tc) formula and are defined by the following equations:

$$K_{1VD}(i, j) = \text{Tc}(\text{DiffFP}_i, \text{DiffFP}_j)$$

$$K_{2V12}(i, j) = \text{Tc}(\text{V1FP}_i, \text{V1FP}_j) \cdot \text{Tc}(\text{V2FP}_i, \text{V2FP}_j)$$

$$K_{2VCD}(i, j) = \text{Tc}(\text{DiffFP}_i, \text{DiffFP}_j) \cdot \text{Tc}(\text{CommV1V2FP}_i, \text{CommV1V2FP}_j)$$

$$K_{2VKD}(i, j) = \text{Tc}(\text{KeyFP}_i, \text{KeyFP}_j) \cdot K_{1VD}(i, j)$$

$$K_{3VK12}(i, j) = \text{Tc}(\text{KeyFP}_i, \text{KeyFP}_j) \cdot K_{2V12}(i, j)$$

$$K_{3VKCD}(i, j) = \text{Tc}(\text{KeyFP}_i, \text{KeyFP}_j) \cdot K_{2VCD}(i, j)$$

where i and j are two direction-dependent MMPs and $\text{Tc}(\text{FP}_i, \text{FP}_j)$ represents the Tc for comparison of the two fingerprints. MMP kernels were expected to project potency-annotated MMPs into feature spaces in which linear modeling algorithms could be successfully applied. SVMlight²⁶ was used to perform all SVR calculations. Except for the kernel functions, default parameter settings were used.

Control Calculations. For comparison with SVR, two conceptually different approaches were used including MMP-based averaging analysis (MMPAV)^{13,14} and RF predictions.^{2,15} To predict the potency change of an MMP in the test set for

Table 2. SVR Predictions^a

		A	B	C	D	E	F	G	H	I
1VD	MACCS	0.21	0.26	0.32	0.34	0.38	0.34	0.30	0.42	0.45
	ECFP2	0.32	0.35	0.43	0.43	0.43	0.41	0.36	0.51	0.53
	ECFP4	0.34	0.39	0.46	0.47	0.46	0.42	0.38	0.53	0.55
	ECFP6	0.34	0.40	0.46	0.47	0.46	0.42	0.39	0.53	0.55
2V12	MACCS	0.20	0.28	0.33	0.38	0.40	0.33	0.32	0.44	0.47
	ECFP2	0.27	0.35	0.43	0.41	0.43	0.38	0.36	0.50	0.51
	ECFP4	0.26	0.36	0.41	0.41	0.43	0.36	0.35	0.49	0.51
	ECFP6	0.25	0.35	0.40	0.40	0.41	0.36	0.34	0.48	0.50
2VCD	MACCS	0.19	0.28	0.33	0.36	0.39	0.33	0.31	0.43	0.46
	ECFP2	0.29	0.36	0.43	0.42	0.43	0.39	0.35	0.50	0.51
	ECFP4	0.30	0.38	0.44	0.44	0.45	0.40	0.36	0.51	0.53
	ECFP6	0.29	0.38	0.43	0.45	0.45	0.39	0.36	0.51	0.53
2VKD	MACCS	0.36	0.43	0.57	0.58	0.62	0.53	0.60	0.60	0.71
	ECFP2	0.49	0.58	0.73	0.67	0.74	0.64	0.71	0.73	0.79
	ECFP4	0.50	0.63	0.77	0.72	0.78	0.67	0.75	0.75	0.83
	ECFP6	0.49	0.64	0.78	0.73	0.78	0.68	0.75	0.76	0.83
3VK12	MACCS	0.28	0.38	0.52	0.52	0.58	0.47	0.54	0.56	0.66
	ECFP2	0.35	0.51	0.66	0.58	0.66	0.54	0.63	0.66	0.71
	ECFP4	0.33	0.53	0.66	0.58	0.66	0.53	0.64	0.65	0.72
	ECFP6	0.31	0.52	0.64	0.58	0.65	0.53	0.63	0.63	0.71
3VKCD	MACCS	0.28	0.39	0.52	0.51	0.58	0.47	0.54	0.55	0.65
	ECFP2	0.41	0.53	0.69	0.61	0.69	0.59	0.65	0.68	0.74
	ECFP4	0.41	0.57	0.71	0.65	0.71	0.60	0.69	0.69	0.77
	ECFP6	0.40	0.57	0.71	0.66	0.72	0.61	0.69	0.69	0.78
		J	K	L	M	N	O	P	Q	
1VD	MACCS	0.38	0.37	0.48	0.60	0.24	0.40	0.43	0.18	
	ECFP2	0.45	0.46	0.52	0.72	0.28	0.49	0.50	0.21	
	ECFP4	0.47	0.49	0.55	0.74	0.29	0.52	0.53	0.22	
	ECFP6	0.47	0.50	0.55	0.74	0.29	0.52	0.54	0.22	
2V12	MACCS	0.40	0.40	0.50	0.64	0.24	0.42	0.44	0.15	
	ECFP2	0.45	0.44	0.52	0.73	0.25	0.47	0.47	0.17	
	ECFP4	0.44	0.44	0.53	0.72	0.24	0.46	0.48	0.17	
	ECFP6	0.43	0.44	0.52	0.71	0.24	0.45	0.47	0.16	
2VCD	MACCS	0.38	0.38	0.48	0.62	0.23	0.41	0.42	0.15	
	ECFP2	0.44	0.44	0.52	0.72	0.26	0.47	0.48	0.19	
	ECFP4	0.46	0.46	0.54	0.73	0.26	0.48	0.51	0.20	
	ECFP6	0.46	0.47	0.54	0.73	0.26	0.49	0.51	0.19	
2VKD	MACCS	0.63	0.63	0.70	0.79	0.44	0.67	0.61	0.37	
	ECFP2	0.77	0.73	0.79	0.89	0.51	0.76	0.73	0.48	
	ECFP4	0.80	0.77	0.83	0.91	0.55	0.81	0.77	0.53	
	ECFP6	0.81	0.79	0.83	0.91	0.57	0.81	0.77	0.54	
3VK12	MACCS	0.58	0.57	0.66	0.76	0.35	0.61	0.57	0.29	
	ECFP2	0.69	0.64	0.72	0.85	0.38	0.67	0.64	0.36	
	ECFP4	0.69	0.64	0.72	0.85	0.38	0.67	0.64	0.38	
	ECFP6	0.67	0.63	0.71	0.84	0.39	0.65	0.63	0.38	
3VKCD	MACCS	0.57	0.56	0.65	0.75	0.37	0.60	0.55	0.32	
	ECFP2	0.72	0.67	0.75	0.86	0.44	0.70	0.67	0.42	
	ECFP4	0.75	0.70	0.78	0.87	0.47	0.73	0.70	0.46	
	ECFP6	0.75	0.71	0.78	0.87	0.48	0.73	0.71	0.47	

^aFor each data set, average R^2 values after 10-fold cross-validation are reported for all fingerprint-kernel combinations. Overall best results are shown in bold.

MMPAV, the training set was searched for MMPs containing the same chemical transformation. If no such MMP was found, prediction was not possible. If qualifying training set MMPs were detected, the average potency difference of these MMPs was calculated to predict the potency change for the test MMP. RF modeling utilizes ensembles of decision trees for consensus predictions. RF calculations were performed using the R²⁷ package randomForest.²⁸ An MMP was represented as the

difference of 51 2D numerical descriptors calculated with the Molecular Operating Environment (MOE)²⁹ and the potency value of the first compound of the MMP. The numerical descriptor set, which was not used for MMP-based SVM modeling, was previously designed for machine learning applications.¹⁵ For RF calculations, the number of trees was set to 400; for all other parameters, default settings were used.

Table 3. SVR Results for the ECFP6-2VKD Combination (10-Fold Cross-Validation)^a

	R^2	SD(R^2)	MAE	SD(MAE)	RMSE	SD(RMSE)	r	SD(r)
A	0.49	0.06	0.48	0.04	0.84	0.10	0.73	0.04
B	0.64	0.06	0.30	0.02	0.48	0.06	0.81	0.04
C	0.78	0.01	0.23	0.01	0.37	0.01	0.89	0.01
D	0.73	0.02	0.34	0.01	0.50	0.02	0.86	0.01
E	0.78	0.02	0.24	0.01	0.37	0.03	0.89	0.01
F	0.68	0.03	0.31	0.01	0.47	0.03	0.84	0.02
G	0.75	0.03	0.29	0.02	0.46	0.03	0.87	0.02
H	0.76	0.02	0.30	0.02	0.47	0.03	0.88	0.01
I	0.83	0.02	0.27	0.01	0.41	0.02	0.91	0.01
J	0.81	0.02	0.24	0.01	0.38	0.02	0.91	0.01
K	0.79	0.03	0.33	0.02	0.50	0.04	0.89	0.02
L	0.83	0.02	0.29	0.01	0.46	0.02	0.92	0.01
M	0.91	0.01	0.18	0.00	0.28	0.01	0.96	0.00
N	0.57	0.04	0.47	0.03	0.73	0.05	0.77	0.03
O	0.81	0.02	0.22	0.01	0.35	0.02	0.91	0.01
P	0.77	0.04	0.27	0.02	0.40	0.02	0.89	0.02
Q	0.54	0.06	0.35	0.03	0.56	0.06	0.75	0.04

^aFor each data set, results of SVR predictions for the preferred ECFP6-2VKD combination are reported using different performance measures including the average and standard deviation (SD) of the coefficient of determination (R^2), mean absolute error (MAE), root-mean-square error (RMSE), and correlation coefficient (r) over 10 independent trials.

The RF protocol followed a previous report of predictions of MMP-encoded property changes.¹⁵

In order to evaluate if prediction accuracy resulted from contributions of newly designed kernels or was essentially determined by SVR, the performance of SVR calculations using the overall preferred kernel/fingerprint combination was compared to kernel ridge regression (KRR),³⁰ which represents an alternative kernel-based regression method. Therefore, the MMP kernels were implemented in R using the package kernlab,³¹ and KRR calculations were carried out using the R package CVST.³² Except for the kernel function, default parameter settings were used.

Learning and Scoring. Training calculations for SVR, RF, and MMPAV were based on 10-fold cross-validation. Compound pairs forming direction-dependent MMPs were randomly divided into 10 nonoverlapping groups. For each compound pair, the two direction-dependent MMPs were assigned to the same group. Regression was performed 10 times. Each time a different group was chosen as the test set, and the remaining nine groups were combined as the training set.

As further control calculations, SVR was also performed using 4-fold cross-validation on the basis of four data set subsets (instead of 10) and, in addition, without cross-validation on larger test sets (i.e., randomly selecting half of each data set as the training and the other half as the test set). These calculations were carried out in order to evaluate the influence of training set composition and size on prediction accuracy. For comparing KRR and SVR, no cross-validation was applied due to the high computational expense of KRR calculations.

For each prediction, the coefficient of determination (R^2), mean absolute error (MAE), root-mean-square error (RMSE), and Pearson's correlation coefficient (r) were calculated comparing the predicted and the observed potency difference values for the test set. Following cross-validation, the average and standard deviation (SD) of the scores for each independent trial were calculated and used as the final prediction result. The different performance measures applied herein are defined by the following equations:

$$R^2 = 1 - \frac{\sum (\text{obs}_i - \text{pred}_i)^2}{\sum (\text{obs}_i - \overline{\text{obs}})^2}$$

$$\text{MAE} = \frac{\sum |\text{obs}_i - \text{pred}_i|}{N}$$

$$\text{RMSE} = \sqrt{\frac{\sum (\text{obs}_i - \text{pred}_i)^2}{N}}$$

$$r = \frac{\sum (\text{obs}_i - \overline{\text{obs}}) \cdot (\text{pred}_i - \overline{\text{pred}})}{\sqrt{\sum (\text{obs}_i - \overline{\text{obs}})^2 \cdot \sum (\text{pred}_i - \overline{\text{pred}})^2}}$$

RESULTS AND DISCUSSION

Kernel Characteristics. Newly designed kernel functions accounted for transformation and MMP information in

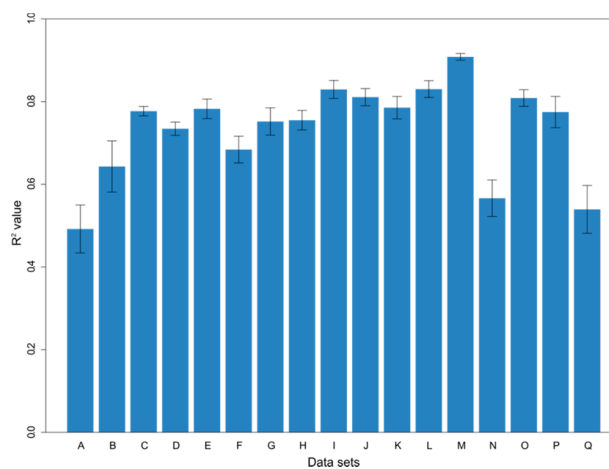


Figure 2. Best SVR predictions. For each data set, the R^2 value for the ECFP6-2VKD fingerprint-kernel combination is reported. R^2 standard deviations for 10 independent trials are given at the top of each bar.

Table 4. SVR Results for the ECFP6-2VKD Combination (4-Fold Cross-Validation)^a

	R^2	SD(R^2)	MAE	SD(MAE)	RMSE	SD(RMSE)	r	SD(r)
A	0.45	0.03	0.50	0.03	0.88	0.09	0.71	0.02
B	0.61	0.02	0.32	0.01	0.51	0.02	0.79	0.01
C	0.75	0.01	0.24	0.00	0.39	0.01	0.88	0.01
D	0.72	0.01	0.35	0.01	0.52	0.01	0.86	0.01
E	0.76	0.01	0.25	0.01	0.38	0.01	0.88	0.01
F	0.66	0.02	0.32	0.01	0.48	0.01	0.83	0.01
G	0.74	0.02	0.30	0.00	0.47	0.01	0.87	0.01
H	0.74	0.01	0.31	0.01	0.49	0.01	0.87	0.01
I	0.82	0.00	0.28	0.01	0.43	0.01	0.91	0.00
J	0.79	0.01	0.26	0.01	0.40	0.01	0.90	0.00
K	0.78	0.02	0.34	0.00	0.51	0.02	0.89	0.01
L	0.81	0.01	0.31	0.00	0.48	0.01	0.91	0.01
M	0.90	0.01	0.19	0.00	0.29	0.01	0.95	0.00
N	0.54	0.02	0.49	0.01	0.75	0.03	0.75	0.01
O	0.78	0.01	0.23	0.00	0.37	0.01	0.90	0.00
P	0.75	0.01	0.29	0.01	0.43	0.01	0.88	0.01
Q	0.50	0.02	0.36	0.02	0.58	0.02	0.72	0.02

^aFor each data set, results of SVR predictions for the preferred ECFP6-2VKD combination are reported using different performance measures including the average and standard deviation (SD) of the coefficient of determination (R^2), mean absolute error (MAE), root-mean-square error (RMSE), and correlation coefficient (r) over four independent trials.

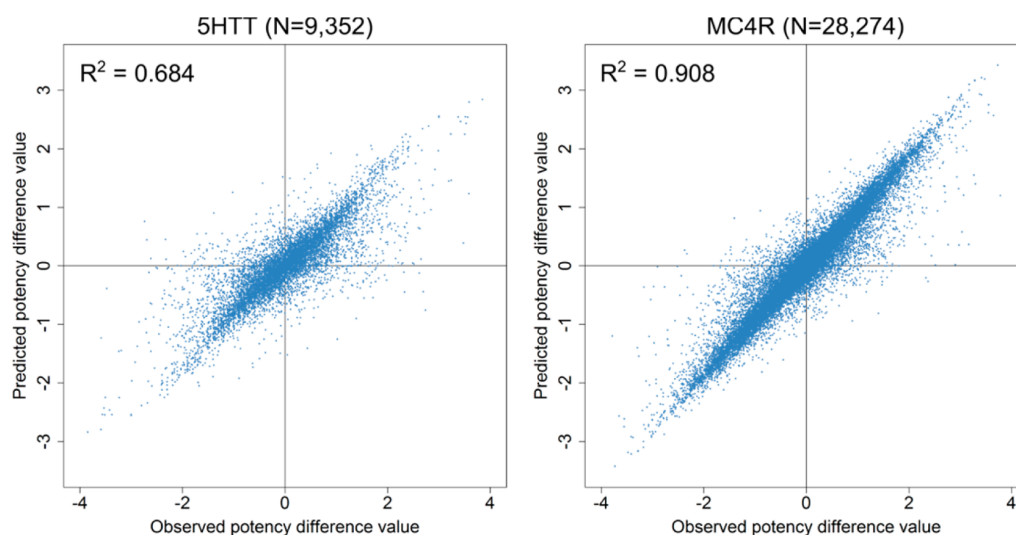


Figure 3. Comparison of predicted and observed potency difference values. For two exemplary data sets, 5HTT and MC4R, predicted and observed potency value differences are compared in a scatterplot. R^2 values and the number of MMPs (N) plotted are given. Each data point represents a direction-dependent MMP.

different ways. For example, CommV1V2FP and DiffFPs generated for value fragments and the kernels built using these representations took combined transformation information and/or structural differences between exchanged substructures into account. By contrast, the incorporation of KeyFPs added core structure information to transformation kernels and hence represented the structural context in which transformations occurred. The design of MMP pair product kernels is based upon pre-existing pairwise Tanimoto kernel products.

Systematic SVR Predictions. For each data set, 24 SVR predictions were carried out resulting from combinations of four fingerprint descriptors (ECFP2, ECFP4, ECFP6, and MACCS) and six kernel functions (1VD, 2V12, 2VCD, 2VKD, 3VK12, and 3VKCD), as described in detail in the Methods section. The R^2 results of 10-fold cross-validated calculations

are reported in Table 2. R^2 values significantly varied for different combinations and data sets and ranged from 0.15 to 0.91, hence reflecting significant differences in prediction accuracy. However, regardless of the magnitude of R^2 values, most trials produced stable results with low R^2 standard deviations of, on average, only 0.037.

Kernel and Descriptor Performance. For all data sets, MMP kernels (taking core structure and transformation information into account) were found to perform better than transformation kernels. The average R^2 value for all calculations with MMP kernels was 0.63 compared to 0.42 for transformation kernels. Thus, structural context information was of critical importance for accurate potency difference value predictions associated with specific chemical transformations. Furthermore, the best representation of a transformation was the value difference fingerprint (1VD and the combined 2VKD

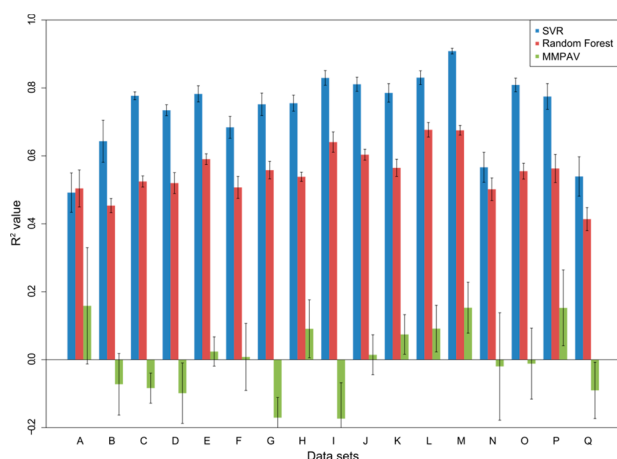


Figure 4. Control calculations. For each data set, the performance of SVR using the ECFP6–2VKD combination (blue) is compared to RF (red) and MMPAV (green) calculations. R^2 values and standard deviations over 10 independent trials are reported.

version), rather than the common plus difference fingerprints (2VCD and 3VKCD) or the two individual value fingerprints (2V12 and 3VK12). The average R^2 for 1VD (0.43) was slightly larger than for 2VCD (0.41) and 2V12 (0.40). Similarly, the average R^2 for 2VKD (0.68) was larger than for 3VKCD (0.62) and 3VK12 (0.58). Hence, kernel 2VKD, which combined the value difference fingerprint and the key fingerprint, performed overall best.

ECFP fingerprints consistently yielded higher performance than MACCS structural keys. Furthermore, increasing the diameter of the ECFPs (and hence their topological resolution) improved SVR prediction performance in most cases, albeit with generally low differences between ECFP4 and ECFP6.

Preferred Combination. The combination of the 2VKD kernel and the high-resolution ECFP6 gave the maximum R^2 value among all 24 kernel-fingerprint combinations for 16 of 17 compound data sets (Table 2). In Table 3, prediction results are reported for 10-fold cross-validation for the preferred 2VKD–ECFP6 combination. R^2 values varied between 0.49

(CA2) and 0.91 (MC4R), as also shown in Figure 2. For 11 of 17 compound data sets, R^2 values of at least 0.75 were obtained (Tables 2 and 3), reflecting generally high prediction accuracy. Again, R^2 standard deviations for these calculations were very low (Table 3), ranging from 0.01 (MC4R) to 0.06 (5-HT1A). Alternative performance measures were applied. MAE and RMSE results showed trends similar to those observed for R^2 . MAE values varied between 0.18 (MC4R) and 0.48 (CA2) and RMSE values between 0.28 (MC4R) and 0.84 (CA2). The correlation between observed and predicted values was generally high, even for predictions yielding intermediate R^2 values. Only two compound data sets had correlation coefficient (r) values below 0.8 and five data sets had values above 0.9. In Table 4, prediction results are reported for 4-fold cross-validation for 2VKD–ECFP6 combination. R^2 values were very similar to those obtained with 10-fold cross-validation (Table 3) showing that test set size and composition had no major influence on prediction accuracy.

Figure 3 shows scatterplots comparing observed and predicted potency difference values for two exemplary data sets with moderate (SHTT) and high (MC4R) prediction accuracy. The observed potency differences reported in these plots are also representative for the compound data sets under study. Across all data sets, only ~4.5% of the direction-dependent MMPs encoded potency differences of 2 orders of magnitude or more. The comparison in Figure 3 revealed that predicted potency differences covered the entire range of observed differences. Moreover, the pronounced diagonal patterns resulted from the presence of many highly accurate predictions.

Control Calculations. To put SVR performance into perspective, control calculations were carried out using MMPAV and RF. Especially RF calculations were relevant for comparison with SVR, given a previous report that utilized RF analysis to predict MMP-associated changes in property values.¹⁵ Figure 4 reports the results of control calculations compared to SVR using the preferred 2VKD–ECFP6 combination. MMPAV performed poorly and even produced negative R^2 for eight compound sets. Details are provided in Table 5. Between 35% (CA9) and 70% (MC4R) of the test

Table 5. MMP-Based Averaging Analysis^a

	R^2	SD(R^2)	MAE	SD(MAE)	RMSE	SD(RMSE)	R	SD(R)
A	0.16	0.17	0.69	0.03	0.98	0.04	0.47	0.14
B	-0.07	0.09	0.60	0.03	0.81	0.06	0.26	0.09
C	-0.08	0.04	0.60	0.03	0.81	0.03	0.29	0.04
D	-0.10	0.09	0.67	0.02	0.90	0.02	0.29	0.04
E	0.02	0.04	0.53	0.02	0.73	0.03	0.38	0.03
F	0.01	0.10	0.57	0.04	0.76	0.05	0.35	0.06
G	-0.17	0.06	0.69	0.02	0.93	0.04	0.24	0.04
H	0.09	0.09	0.64	0.03	0.86	0.04	0.44	0.05
I	-0.17	0.11	0.67	0.02	0.90	0.03	0.26	0.06
J	0.01	0.06	0.61	0.02	0.83	0.04	0.38	0.04
K	0.07	0.06	0.70	0.03	0.93	0.05	0.42	0.04
L	0.09	0.07	0.69	0.03	0.95	0.04	0.45	0.04
M	0.15	0.07	0.51	0.02	0.70	0.03	0.52	0.04
N	-0.02	0.16	0.75	0.08	1.06	0.10	0.33	0.11
O	-0.01	0.10	0.51	0.03	0.71	0.04	0.38	0.07
P	0.15	0.11	0.48	0.03	0.65	0.04	0.46	0.07
Q	-0.09	0.08	0.60	0.06	0.85	0.09	0.28	0.04

^aFor each data set, results of MMPAV calculations are reported using different performance measures according to Table 3.

Table 6. Random Forest Control Calculations^a

	R^2	$SD(R^2)$	MAE	$SD(MAE)$	RMSE	$SD(RMSE)$	r	$SD(r)$
A	0.50	0.05	0.57	0.03	0.83	0.08	0.71	0.04
B	0.45	0.02	0.43	0.02	0.60	0.04	0.68	0.02
C	0.52	0.02	0.39	0.01	0.53	0.01	0.73	0.01
D	0.52	0.03	0.52	0.02	0.67	0.02	0.72	0.02
E	0.59	0.02	0.37	0.01	0.51	0.02	0.77	0.01
F	0.51	0.03	0.44	0.02	0.58	0.02	0.72	0.02
G	0.56	0.03	0.45	0.01	0.61	0.02	0.75	0.02
H	0.54	0.01	0.49	0.02	0.65	0.02	0.74	0.01
I	0.64	0.03	0.44	0.02	0.60	0.02	0.81	0.02
J	0.60	0.02	0.40	0.01	0.55	0.02	0.78	0.01
K	0.56	0.03	0.54	0.02	0.71	0.03	0.76	0.02
L	0.68	0.02	0.47	0.01	0.63	0.01	0.83	0.01
M	0.68	0.01	0.39	0.01	0.52	0.01	0.83	0.01
N	0.50	0.03	0.57	0.02	0.78	0.03	0.71	0.02
O	0.55	0.02	0.39	0.01	0.53	0.02	0.75	0.02
P	0.56	0.04	0.42	0.01	0.56	0.02	0.75	0.03
Q	0.41	0.03	0.44	0.03	0.63	0.05	0.64	0.03

^aFor each data set, results of RF calculations are reported using different performance measures according to Table 3.

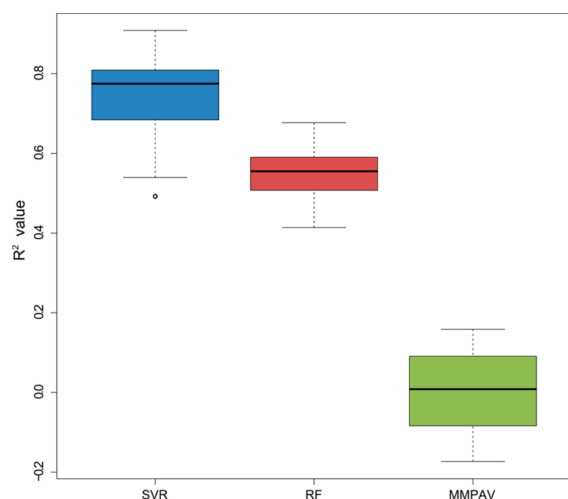


Figure 5. R^2 comparison. The R^2 value distributions resulting from the preferred SVR combination (blue), RF (red), and MMPAV (green) calculations are compared in a boxplot representation.

MMPs could not be predicted using MMPAV because a qualifying transformation was not available in the learning set. It should be noted that negative R^2 indicated that better predictions would be obtained by using the mean of the entire potency change distribution of a data set (essentially corresponding to random predictions). This very low performance was not unexpected for a simple averaging method. However, the results clearly indicated nonlinearity of many MMP-encoded SARs.

RF calculations yielded much better prediction performance than MMPAV. For most data sets, R^2 values between 0.5 and 0.6 were observed for RF predictions, with a maximum value of 0.68 (MC4R and ADORA3). Details are provided in Table 6. However, as reported in Figure 4, RF predictions did not reach the prediction accuracy of SVR for 16 of 17 compound data sets. Figure 5 compares R^2 values for the different calculations in boxplots and shows that the interquartile range of RF was lower than SVR and that the median R^2 value of SVR was ~ 0.2 units higher.

Table 7. Comparison of SVR and KRR for the ECFP6-2VKD Combination^a

	R^2		MAE		RMSE		r	
	SVR	KRR	SVR	KRR	SVR	KRR	SVR	KRR
A	0.35	0.52	0.56	0.51	0.96	0.82	0.62	0.73
B	0.53	0.59	0.36	0.33	0.56	0.52	0.74	0.77
C	0.68	0.72	0.27	0.24	0.43	0.40	0.84	0.86
D	0.64	0.69	0.41	0.38	0.59	0.55	0.82	0.84
E	0.70	0.74	0.29	0.26	0.44	0.41	0.85	0.87
F	0.60	0.63	0.36	0.34	0.52	0.50	0.79	0.80
G	0.68	0.72	0.35	0.32	0.53	0.49	0.83	0.86
H	0.67	0.72	0.37	0.33	0.56	0.52	0.83	0.86
I	0.77	0.80	0.32	0.29	0.48	0.44	0.88	0.90
J	0.73		0.30		0.45		0.87	
K	0.70	0.75	0.41	0.36	0.58	0.54	0.85	0.87
L	0.75		0.36		0.55		0.87	
M	0.86		0.22		0.33		0.93	
N	0.48	0.57	0.54	0.49	0.80	0.73	0.71	0.76
O	0.71	0.76	0.27	0.24	0.43	0.39	0.86	0.88
P	0.69	0.74	0.34	0.31	0.47	0.43	0.84	0.87
Q	0.44	0.49	0.40	0.38	0.61	0.59	0.67	0.70

^aFor each data set, results of SVR and KRR predictions for the preferred ECFP6-2VKD combination are reported using different performance measures including the coefficient of determination (R^2), mean absolute error (MAE), root-mean-square error (RMSE), and correlation coefficient (r). For data sets J, L, and M, KRR calculations could not be completed.

Kernel ridge regression was compared to SVR using the preferred 2VKD-ECFP6 combination without cross-validation (i.e., on larger test sets). KRR is computationally much more demanding than SVR both in terms of memory requirements and in CPU time (KRR calculations could not be completed for three large data sets). Results are reported in Table 7. Compared to SVR, KRR calculations yielded a small increase in R^2 values for most data sets. These findings indicated that the newly designed kernels were largely responsible for the observed prediction accuracy, rather than SVR optimization details.

CONCLUSIONS

In this work, we have explored support vector regression using newly designed kernel functions for the prediction of numerical potency differences between compounds forming MMPs. Application of the MMP formalism for potency prediction further expands the applicability domain of QSAR-type approaches. This is the case because (i) many different structural relationships are captured at the level of compound pairs and (ii) MMPs encode well-defined chemical transformations in different structural environments. For potency difference prediction, the MMP approach was further refined by introducing direction-dependent MMPs. By combining MMP-based transformation analysis and machine learning approaches such as SVR, nonlinear SARs can be captured in structurally heterogeneous data sets. In our calculations, overall high SVR prediction accuracy was achieved for a preferred combination of a kernel taking transformation and core structure information into account and a high-resolution topological fingerprint descriptor. Transformation information was best captured by a fingerprint representation accounting for structural differences between the exchanged substructures. Given that potency difference values were predicted using SVR with reasonable to high accuracy for structurally analogous compounds from many different data sets, the methodology introduced herein should merit further consideration for compound potency predictions to complement and potentially extend existing QSAR approaches.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to OpenEye Scientific Software, Inc., for the free academic license of the OpenEye Toolkits.

REFERENCES

- (1) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.
- (2) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.
- (3) Drucker, H.; Burges, C. Support Vector Regression Machines. *Adv. Neural Inform. Process. Systems* **1997**, *9*, 155–161.
- (4) Yuan, Y.; Zhang, R.; Hu, R.; Ruan, X. Prediction of CCR5 Receptor Binding Affinity of Substituted 1-(3,3-diphenylpropyl)-piperidinyl Amides and Ureas Based on the Heuristic Method, Support Vector Machine and Projection Pursuit Regression. *Eur. J. Med. Chem.* **2009**, *44*, 25–34.
- (5) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. QSAR Models for the Prediction of Binding Affinities to Human Serum Albumin Using the Heuristic Method and a Support Vector Machine. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1693–1700.
- (6) Sun, M.; Chen, J.; Wei, H.; Yin, S.; Yang, Y.; Ji, M. Quantitative Structure-activity Relationship and Classification Analysis of Diaryl Ureas Against Vascular Endothelial Growth Factor Receptor-2 Kinase Using Linear and Non-linear Models. *Chem. Biol. Drug Des.* **2009**, *73*, 644–654.
- (7) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855–1859.
- (8) Gombar, V. K.; Hall, S. D. Quantitative Structure-activity Relationship Models of Clinical Pharmacokinetics: Clearance and Volume of Distribution. *J. Chem. Inf. Model.* **2013**, *53*, 948–957.
- (9) Fatemi, M. H.; Gharaghani, S. A Novel QSAR Model for Prediction of Apoptosis-inducing Activity of 4-aryl-4H-chromenes Based on Support Vector Machine. *Bioorg. Med. Chem.* **2007**, *15*, 7746–7754.
- (10) Leong, M. K. A Novel Approach Using Pharmacophore Ensemble/Support Vector Machine (PhE/SVM) for Prediction of hERG Liability. *Chem. Res. Toxicol.* **2007**, *20*, 217–226.
- (11) Song, M.; Clark, M. Development and Evaluation of an in Silico Model for hERG Binding. *J. Chem. Inf. Model.* **2005**, *46*, 392–400.
- (12) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Cheminformatics in Drug Discovery*; Oprea, T. L., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271–285.
- (13) Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180–192.
- (14) de la Vega de León, A.; Bajorath, J. Compound Optimization Through Data Set-dependent Chemical Transformations. *J. Chem. Inf. Model.* **2013**, *53*, 1263–1271.
- (15) Beck, J. M.; Springer, C. Quantitative Structure-activity Relationship Models of Chemical Transformations from Matched Pairs Analyses. *J. Chem. Inf. Model.* **2014**, *54*, 1226–1234.
- (16) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18–28.
- (17) Cortes, C.; Vapnik, V. Support-vector Networks. *Machine Learning* **1995**, *20*, 273–297.
- (18) Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of Activity Cliffs Using Support Vector Machines. *J. Chem. Inf. Model.* **2012**, *52*, 2354–2365.
- (19) 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*, D1100–1107.
- (20) 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.
- (21) OEChem, v. Feb2014; OpenEye Scientific Software Inc: Santa Fe, NM, 2014.
- (22) 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.
- (23) MACCS Structural Keys; Symyx Software: San Ramon, CA, 2005.
- (24) Rogers, D.; Hahn, M. Extended-connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (25) Willett, P.; Barnard, J.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (26) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods – Support Vector Learning*; Schölkopf, B.; Burges, C. J. C.; Smola, A. J., Eds.; MIT-Press: Cambridge, MA, 1999; pp 169–184.
- (27) R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2008.
- (28) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
- (29) Molecular Operating Environment (MOE); Chemical Computing Group Inc.: Montreal, Canada, 2011.
- (30) Christianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
- (31) Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab - An S4 Package for Kernel Methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20.

(32) CVST R package. <http://cran.r-project.org/web/packages/CVST/index.html>.

Conclusions

An SVR application to predict the potency difference between MMP partners has been presented. Six different kernel functions (three transformation kernels and three MMP kernels) were compared. MMP kernels performed generally better than transformation kernels across 17 data sets. SVR was compared to random forest (RF) and kernel ridge regression (KRR). SVR outperformed RF in all but one data set. KRR (an alternative kernel-based regression method) yielded high prediction accuracy, slightly higher than SVR. However, its implementation was much slower and for three data sets KRR calculations could not be completed. Because good performance was obtained for KRR, SVR optimizations were probably not the origin of the prediction accuracy. Rather, the kernel functions seemed responsible for the good overall concordance between observed and predicted values.

SVR has proven to be a powerful methodology for property change prediction. However, it is a complex methodology with a marked “black box” character. For the next study, we used a more intuitive prediction method based on MMPA and coupled it to a high-dimensional visualization to rationalize the results of a multi-objective compound optimization campaign.

7

Compound optimization through data set-dependent chemical transformations

Introduction

During multi-objective compound optimization, the structure of a molecule is modified to alter its properties. These modifications are expected to improve the ranking of the compounds for different objective functions. This process can be rationalized as searching for specific regions of chemical space where compounds with favorable property values preferentially congregate. In this study, we combine visualization of high-dimensional property space, multi-objective optimization, and MMPA of data set-dependent chemical modifications. Principal component plots allow the identification of favorable regions of property space. Chemical transformations that encode consistent property changes are used to modify compounds in stages, moving them in the scatter plot from unfavorable to favorable regions of property space. This study provides a proof-of-concept of MMP-based compound optimization rationalized on the basis of visualizations of high-dimensional property space.

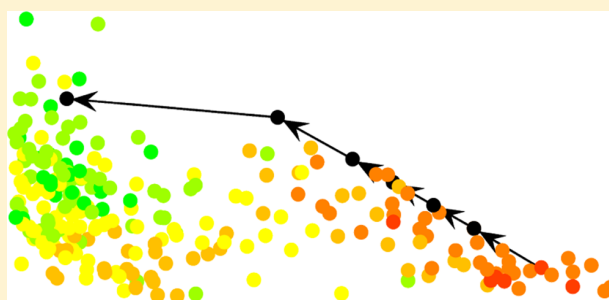
Reprinted with permission from "A. de la Vega de León, J. Bajorath. Compound optimization through data set-dependent chemical transformations. *Journal of Chemical Information and Modeling* **2013**, 53(6), 1263-1271". Copyright 2013 American Chemical Society

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 classes⁶ 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 Toolkit¹⁴ 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

Published: May 8, 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: $\text{LogP} \leq 3$, $\text{MW} \leq 400$, $\text{TPSA} \leq 120$,

$\text{rotN} \leq 7$; intermediate: $\text{LogP} 3-5$, $\text{MW} 400-500$, $\text{TPSA} 120-140$, $\text{rotN} 8-10$; unfavorable: $\text{LogP} > 5$, $\text{MW} > 500$, $\text{TPSA} > 140$, $\text{rotN} > 10$.

Chemical Transformations. Transformation size-restricted MMPs were calculated as described previously⁸ 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., $A \rightarrow B$ and $B \rightarrow A$. This was done because transformations were

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

^aFor each data set, the total number of compounds, MMPs, corresponding transformations, and the number of frequent and preferred transformations are reported.

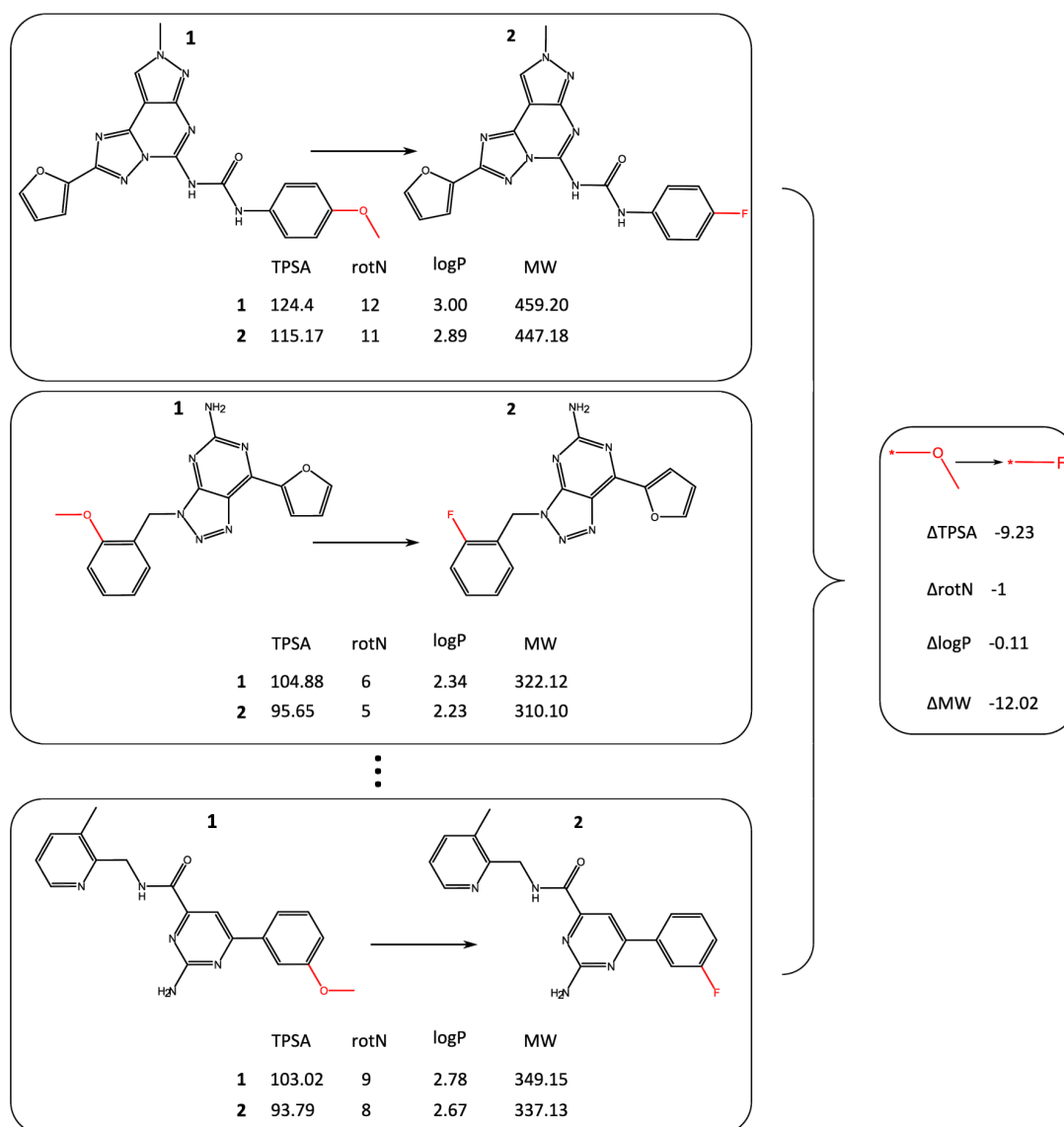


Figure 1. Transformation evaluation. To assess property changes as a consequence of a frequently occurring transformation, MMPs representing the transformation are analyzed. The descriptor value differences between compounds 2 and 1 forming each MMP are calculated and averaged over the MMPs.

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.

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.

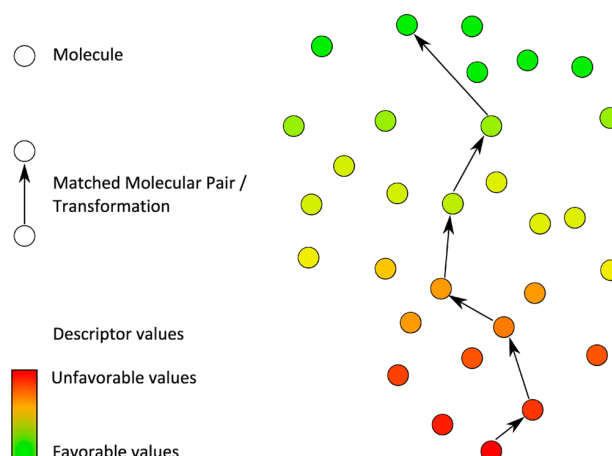


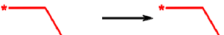

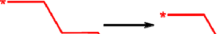
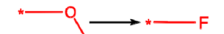


Figure 2. Compound pathway visualization. A compound pathway is delineated using arrows in the PCA projection of a hypothetical data set. Molecules are represented as nodes that are color-coded using a spectrum ranging from unfavorable to favorable descriptor values. Compounds connected by an arrow form 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

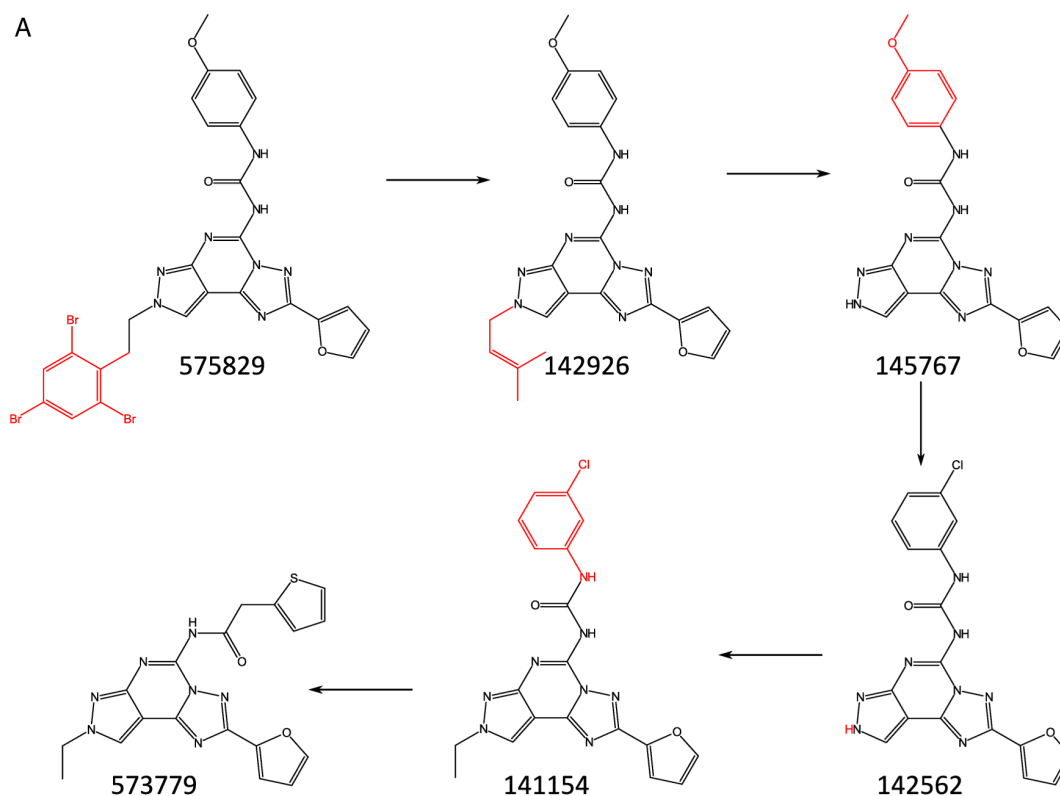
^aPreferred transformations are listed that consistently occurred in all four compound data sets together with their average descriptor value changes.

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 key¹⁸ Tanimoto similarity¹⁹ >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

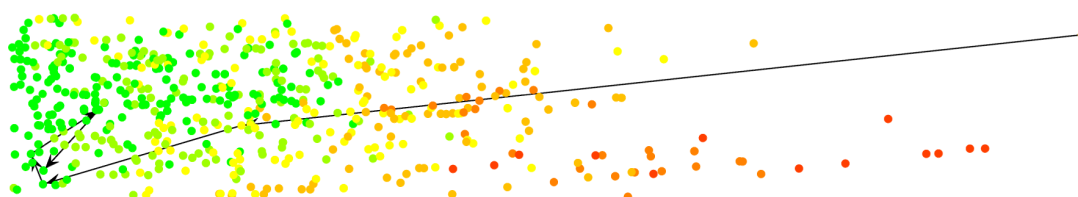
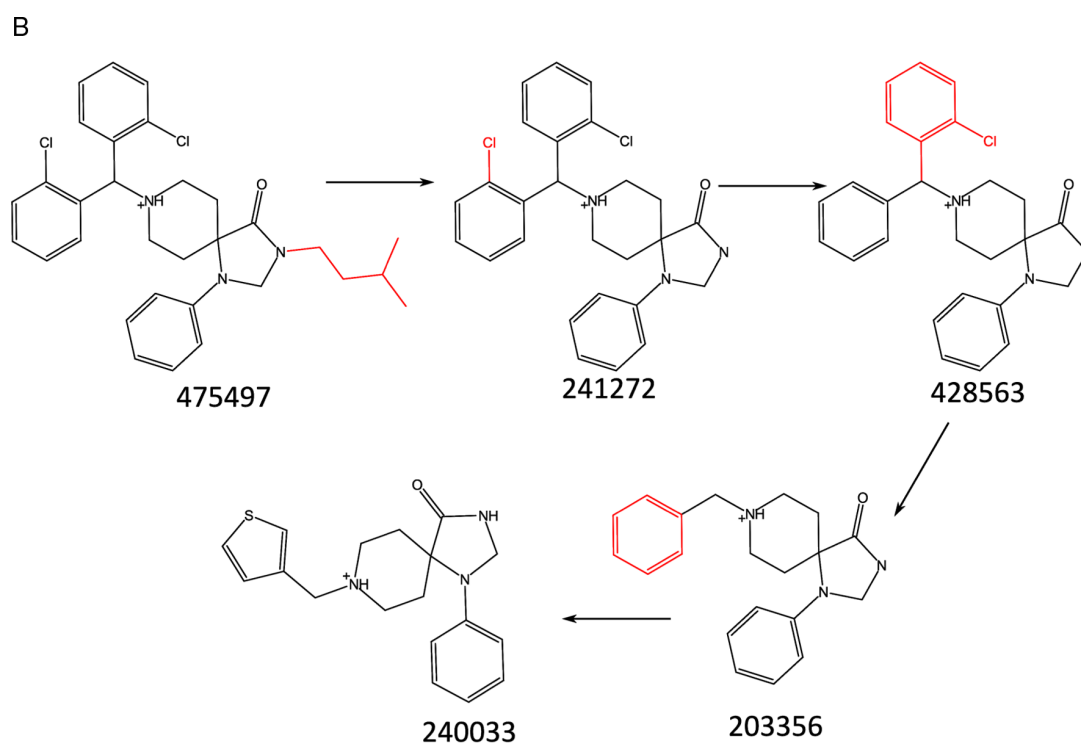


Figure 3. continued



Molecule	pK _i	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

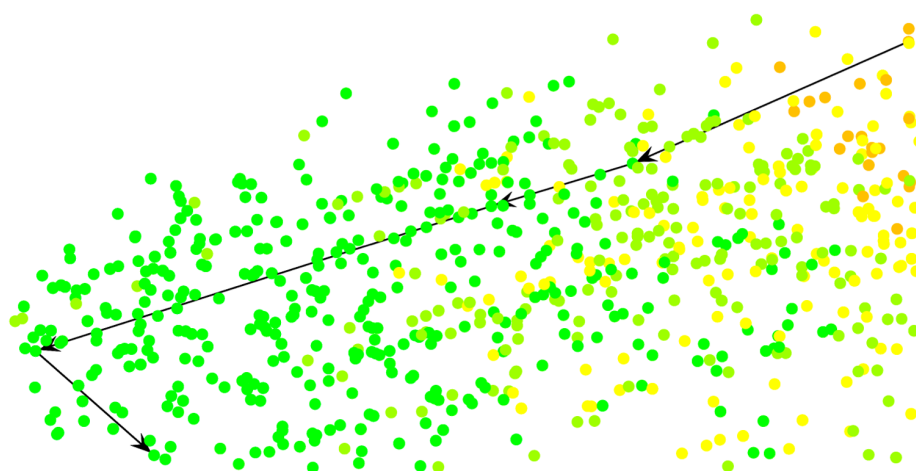


Figure 3. Compounds pathways. In (A) and (B), representative pathways from the A2AR and MOR data sets are shown, respectively. At the top of each representation, the compounds forming the pathway are shown and labeled with their ChEMBL ID. Substructures exchanged along the path are colored red. Below 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

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

^aFor 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.

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

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

1

596135

4

592896

7

596133

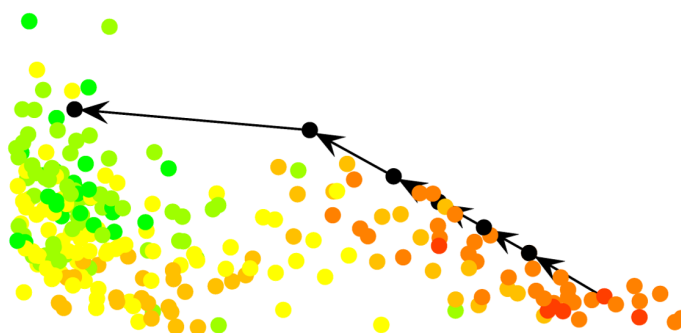


Figure 4. continued

B

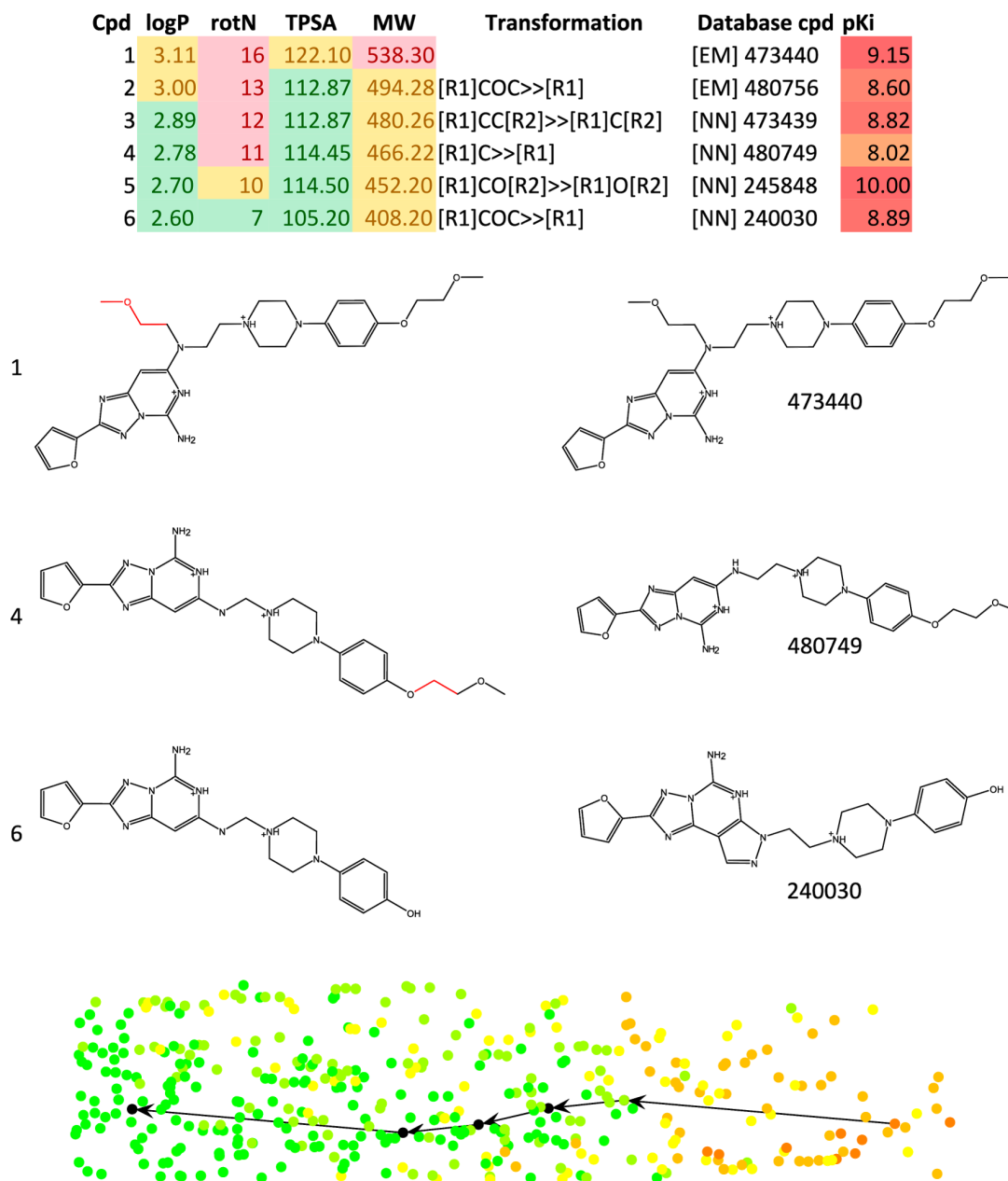


Figure 4. Compound optimization. In (A) and (B), exemplary compound optimization paths are shown originating from compound 596135 of data set CB2 and from compound 473440 (A2AR), respectively. Representation elements are according to Figure 3. Structures of designed compounds (left) and database matches (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 *Cheminformatics 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) Dosseter, 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.; Meinel, 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.; Rodriguez, 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.

Conclusions

Here, we have presented a novel methodology that combines multi-objective optimization based on MMPA and visualizations of high-dimensional space. Four sets of compounds active against different G protein coupled receptors were obtained. Four previously described descriptors and eight threshold values¹¹² were used to guide the multi-objective optimization procedure. Principal components were used to display the four-dimensional space and follow the modification of compounds from unfavorable to favorable property values. Chemical transformations were obtained from MMPs for each data set and their effect on the four properties was analyzed. Those transformations that had a consistent effect were used to modify compounds with unfavorable property values. Half of the selected compounds could be optimized in this manner. In some cases, compounds similar to the final, optimized molecule could be found in public databases with defined activity against the data set target. Therefore, data set-dependent chemical transformations were in many cases activity conservative.

In this study, visualization of high-dimensional space was used to rationalize a computational compound optimization procedure. Principal component plots of property space provided an overview of compound distribution. In the next study, we introduced star coordinate and parallel coordinate plots to study chemical space. These displays are used to explore different drug-like subspaces obtained as equivalent solutions of a multi-objective optimization procedure.

8 Visualization of multi-property landscapes for compound selection and optimization

Introduction

Chemical space is frequently studied using a set of molecular descriptors to generate high-dimensional property spaces where each descriptor represents a dimension. These spaces are vast and finding drug-like subspaces, where active and bioavailable compounds are located, is of prime importance to find promising candidate molecules for optimization. In this study, we present two high-dimensional visualizations, the star coordinate and the parallel coordinate plot, to the medicinal chemistry community. Star coordinates are used to represent distinct drug-like subspaces obtained from an optimization task allowing to differentiate between numerically equivalent solutions. Parallel coordinates are used to study the distribution of descriptor values and compare drugs and bioactive compounds. My main contribution to this study was the generation and analysis of star coordinate and parallel coordinate plots.

Reprinted with permission from “A. de la Vega de León, S. Kayastha, D. Dimova, T. Schultz, J. Bajorath. Visualization of multi-property landscapes for compound selection and optimization. *Journal of Computer-Aided Molecular Design* **2015**, 29(8), 695-705”. Copyright 2015 Springer

Visualization of multi-property landscapes for compound selection and optimization

Antonio de la Vega de León¹ · Shilva Kayastha¹ · Dilyana Dimova¹ · Thomas Schultz² · Jürgen Bajorath¹

Received: 16 June 2015 / Accepted: 27 July 2015
© Springer International Publishing Switzerland 2015

Abstract Compound optimization generally requires considering multiple properties in concert and reaching a balance between them. Computationally, this process can be supported by multi-objective optimization methods that produce numerical solutions to an optimization task. Since a variety of comparable multi-property solutions are usually obtained further prioritization is required. However, the underlying multi-dimensional property spaces are typically complex and difficult to rationalize. Herein, an approach is introduced to visualize multi-property landscapes by adapting the concepts of star and parallel coordinates from computer graphics. The visualization method is designed to complement multi-objective compound optimization. We show that visualization makes it possible to further distinguish between numerically equivalent optimization solutions and helps to select drug-like compounds from multi-dimensional property spaces. The methodology is intuitive, applicable to a wide range of chemical optimization problems, and made freely available to the scientific community.

Keywords Compound optimization · Activity landscapes · Structure–property relationships · Multi-objective optimization · Multi-property landscapes · Visualization

Introduction

The exploration of structure–activity relationships (SARs) in large and structurally heterogeneous compound data sets is strongly supported by SAR visualization methods [1]. The concept of activity landscapes (ALs) [2] provides integrated views of compound similarity and activity relationships and has been applied for SAR visualization [1, 2]. Several approaches to the design of two- (2D) and three-dimensional (3D) ALs have been introduced that typically consider activity as the sole compound property. Exemplary 2D AL designs include simple “structure–activity similarity (SAS) maps” [3] that plot structural similarity against activity similarity on the basis of pairwise comparisons of data set compounds and, in addition, various network representations. For example, the “network-like similarity graph” (NSG) [4] has been an original network-based AL design in which nodes represent compounds and edges pairwise (fingerprint) similarity relationships. Nodes in NSGs are annotated with potency and numerical SAR score information. Another more recent design has been “intuitive networks for structure–activity relationship analysis” (inSARa) [5] in which reduced graphs of active compounds are used to determine their maximum common substructures (MCSs). These MCSs are then represented as nodes that are connected by edges indicating hierarchical MCS relationships. Original compounds are then assigned to corresponding MCSs and represented as a second node category, i.e., compound nodes colored by potency. MCS-based visualization methods have

Antonio de la Vega de León and Shilva Kayastha have contributed equally to this work.

✉ Jürgen Bajorath
bajorath@bit.uni-bonn.de

¹ Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, 53113 Bonn, Germany

² Institute of Computer Science II, Computer Graphics, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Ebert-Allee 144, 53113 Bonn, Germany

also been introduced to organize individual compound series and elucidate SAR patterns [6–8]. In addition to network representations, tree-like structures have been designed to graphically organize compound series and study SAR trends in chemical neighborhoods [8, 9]. Several network- [7] or tree-like [8, 9] visualizations can be rationalized as local 2D ALs because they predominantly or exclusively focus on individual compound series (rather than structurally heterogeneous data sets).

Going beyond global or local 2D representations, the design of 3D ALs can be approached in different ways. Given a 2D representation of structural similarity relationships, an intuitive way of generating a 3D AL is adding a biological response surface as a third dimension. This typically requires extrapolation of a hypersurface from sparsely distributed compound activity values, which has been accomplished by adapting the kriging method from geostatistics [10]. An alternative approach to 3D AL design is subjecting a high-dimensional chemical descriptor space to dimension reduction to obtain a 3D view, as exemplified by the ligand induced structure–activity relationship display (LiSARD) [11]. Compound positions in this space can then be annotated with activity information.

Chemical space visualization is not confined to AL views. Rather, different visualization techniques have also been introduced to generalize chemical space display including, for example, similarity-based compound networks [12] and molecular layout algorithms [13] for smaller data sets, projections from high-dimensional descriptors spaces based on principal component analysis for large (or very large) data sets [14, 15], and generative topographic mapping (GTM) [16]. GTM was designed to project from high-dimensional feature spaces onto latent 2D space representations in which points (nodes) correspond to normal probability distributions derived from the original data space that determine the mapping of compounds to the latent space. As such, GTM does not represent an AL view as conventionally defined.

Returning to the AL concept, we emphasize two of its cardinal features: firstly, it is activity-centric (i.e., activity is considered as a single structure-related property); secondly, it is descriptive in nature (i.e., ALs are used to analyze SARs but not predict active compounds). Both of these features limit the applicability of AL representations for compound optimization, which typically is a multi-objective task. During iterative optimization, multiple biologically relevant compound properties are considered in combination with activity, focusing on the key question, which compound(s) to make next [17]. In the practice of medicinal chemistry, this process is predominantly driven by chemical experience and intuition, although it can also be supported by computational means. In computational chemistry, multi-property optimization is typically attempted using evolutionary

algorithms [18–20] or property-weighted objective functions [20], often in combination with Pareto ranking [19, 20] of numerical solutions. These multi-objective methods usually produce reasonable numerical solutions of optimization tasks but are not expected to find the globally best solution. Multi-objective optimization typically produces a variety of comparable solutions and it is often difficult to further differentiate between them and rationalize characteristic features in multi-dimensional property space.

Herein, we introduce an approach to visualize multi-property landscapes, further extending the AL concept, and graphically analyze solutions of property-weighted objective functions. The methodology makes it possible to further differentiate between numerically equivalent optimization solutions and prioritize them for specific tasks by viewing them in a multi-dimensional data set context.

Materials and methods

Compound data selection

In order to model compound optimization processes, data sets were assembled that consisted of two types of compounds active against the same target: bioactive compounds from medicinal chemistry sources and approved drugs. Bioactive compounds were extracted from ChEMBL [21] (version 20). Only compounds with reported direct interactions (i.e., target relationship type “D”) against human targets at the highest assay confidence level (i.e., confidence score 9) and precisely defined equilibrium constants (K_i values) were considered. Compounds with multiple K_i measurements for the same target were retained if all reported values fell within the same order of magnitude. In this case, the arithmetic mean was calculated as the final potency annotation. Approved small molecule drugs with specific target annotations were assembled from DrugBank [22] (version 4.1). To ensure that potency information was available for all drugs and bioactive compounds considered in the analysis, only drugs were retained for which high-confidence activity measurements were available in ChEMBL. All qualifying compounds and drugs with activity against the same target were organized into target-based compound sets. Each target set was required to contain at least 100 bioactive compounds and at least 10 approved drugs. Table 1 summarizes the composition of six target sets satisfying the above criteria assembled for our analysis.

Multi-dimensional property space

A multi-dimensional property space was generated using 14 descriptors accounting for different molecular properties relevant for chemical optimization, as summarized in

Table 1 Data sets combining bioactive compounds and approved drugs

Target ID	Target name	Bioactive CPDs	Drugs
231	Histamine H1 receptor	572	25
1867	Alpha-2a adrenergic receptor	453	23
210	Beta-2 adrenergic receptor	355	19
2035	Muscarinic acetylcholine receptor M5	282	14
4302	P-glycoprotein 1	242	49
4605	Small intestine oligopeptide transporter	181	14

For the six target-based data sets, the ChEMBL target ID, number of bioactive compounds (CPDs), and approved drugs are reported

Table 2. Properties represented by 13 calculated descriptors included, among others, hydrophobic and aromatic character, molecular complexity, hydrogen bonding potential, charge, and surface properties. In addition, compound potency (pK_i ; negative decadic logarithm of the equilibrium constant) was used as a descriptor. Experimental pK_i values for data set compounds were taken from ChEMBL (version 20). The descriptor `a_ringR` (fraction of ring atoms in a molecule) was calculated with the aid of the OpenEye toolkit [23] and the remaining 12 descriptors were calculated using the Molecular Operating Environment (MOE) [24]. This 14-dimensional feature space was designed as a reference space for exemplary multi-property optimization. The feature set selected for our proof-of-concept investigation can of course be replaced by any other number of calculated descriptors and/or experimentally determined properties, depending on the specific optimization tasks.

Property space projection and optimization

Compound subsets with preferred feature value combinations were selected from multi-dimensional feature space. Therefore, compound distributions in 14-dimensional feature space were projected onto a one-dimensional space. A projection of the data was obtained by multiplying an $n \times p$ data matrix, X , with n sample points in p dimensions, with a $p \times d$ projection matrix, A (here with $p = 14$ and $d = 1$). Accordingly, the projection of compound i was given by the formula: $val_i = \sum_{j=1}^p w_j v_j$, where v_j (from X) was the value for descriptor j and w_j (from A) the weight given to descriptor j [25]. The value of this projection was used as the *multi-objective function* (MOF) value for numerical optimization of a compound subset selection.

Values of the 13 numerical descriptors were scaled relative to the observed pK_i range to ensure that no descriptors numerically dominated the value distributions.

Table 2 Descriptors

No.	Name	Definition	Property	Unit
1	<code>a_acc</code>	Number of hydrogen bond acceptors	Hydrogen bonding	Integer
2	<code>a_aroR</code>	Fraction of aromatic ring atoms	Aromaticity	Percentage
3	<code>a_don</code>	Number of hydrogen bond donor atoms	Hydrogen bonding	Integer
4	<code>a_ringR</code>	Fraction of ring atoms	Molecular complexity	Percentage
5	<code>b_rotR</code>	Fraction of rotatable bonds	Flexibility	Percentage
6	<code>chiral_u</code>	Number of chiral centers	Stereochemistry	Integer
7	<code>Fcharge</code>	Sum of formal charges	Charge	Integer
8	<code>logP(o/w)</code>	Log of octanol/water partition coefficient	Hydrophobicity	Log unit
9	<code>logS</code>	Log of aqueous solubility	Solubility	Log (mol/L)
10	<code>PEOE_VSA_FHYD</code>	Fractional hydrophobic van der Waals surface area	Surface property	Percentage
11	<code>PEOE_VSA_FPNEG</code>	Fractional negative polar van der Waals surface area	Surface property	Percentage
12	<code>PEOE_VSA_FPPOS</code>	Fractional positive polar van der Waals surface area	Surface property	Percentage
13	<code>Pot</code>	Potency (pK_i)	Activity	Log (M)
14	<code>Weight</code>	Molecular weight	Molecular size	Da

The set of 14 descriptors used for feature space generation is listed and defined

Optimization was guided by maximizing the MOF value. Therefore, a systematic search was performed using four different weight values for each descriptor $\{-1.0, -0.33, 0.33, 1.0\}$. All 4^{14} (~ 270 millions) possible projections were systematically explored. The weighting scheme chosen for our analysis can be easily exchanged for different properties and optimization tasks. The search procedure is not dependent on a specific methodology or strategy. Descriptor weights can be obtained using alternative approaches including, among others, regression techniques. If the number of features becomes too large for an exhaustive search, stochastic search strategies can also be applied.

Compounds were ranked based on their MOF value and the top 20 compounds were analyzed. Projections were prioritized based on the number of approved drugs within the top 20 ranking. In prioritized set of projections, MOF value corresponded to our drug-likeness model of compounds meaning that compounds with higher MOF values had properties similar to approved drugs. Thus, projections with a significant enrichment of drugs among top-ranked compounds were considered to originate from *drug-like subspaces* representing favorable multi-feature combinations. Our current analysis scheme is focused on the exploration of drug-like subspaces for the generation of which reference sets of known drugs are essential. However, compound reference sets with other characteristic properties of interest can be used for mapping and derivation of descriptor weights.

Visualization of projections

For the visualization of individual projections, the *Star Coordinate* (STC) [26] representation was adopted from computer science. STC is a multi-dimensional visualization technique that arranges coordinates in predefined positions sharing the same origin at the center. The position of a compound in the STC visualization was dependent on the position of each coordinate (descriptor) and the values of the compound for each coordinate.

More formally, the position of compound i in the STC visualization was given by the formula: $\vec{i} = \sum_{j=1}^p v_j \vec{d}_j$, where \vec{d}_j represented the position of descriptor j and v_j the value for descriptor j . The position of descriptor j was calculated as follows: its weight obtained from MOF optimization provided the y -axis value. Along the x -axis, all descriptors were ordered lexicographically and given incremental values between -1 and 1 to distribute them evenly. Figure 1a shows a schematic STC visualization for an individual compound. For a given projection, the STC visualization provides a 2D representation of the data set distribution in multi-dimensional property space. STC for

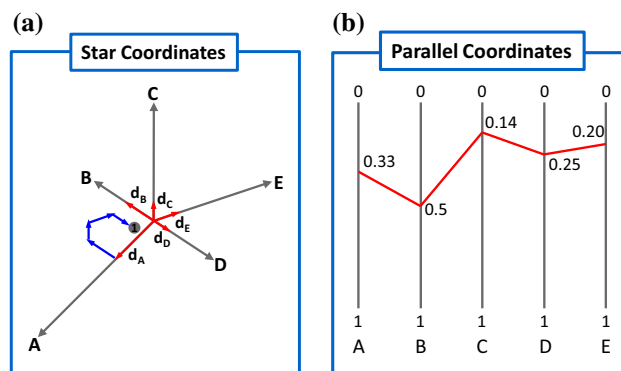


Fig. 1 Star and parallel coordinates. **a** A schematic STC representation for a single compound (gray dot) and five descriptors (A–E). Gray arrows represent descriptor vectors forming the star coordinate. Red arrows (d_A to d_E) represent weighted vectors obtained by multiplying the descriptor value of the compound with the corresponding vector. The position of the compound is determined by the sum of all weighted vectors (indicated by blue arrows for d_B to d_E). **b** An exemplary PAC plot for the same compound. Descriptors (A–E) are assigned to parallel horizontal lines. The red line traces the descriptor values of the compound

multi-property space display was implemented in-house in Java based upon the JUNG library [27].

STC visualization was complemented by the *Parallel Coordinate* (PAC) [28] representation, another multi-dimensional visualization technique from computer science that organizes features (descriptors) on parallel axes. Each axis represents all possible values for a descriptor, ranging from the minimum (top of the axis) to the maximum value (bottom). Compounds are then represented as lines that traverse all descriptor axes at positions corresponding to the value for each descriptor. Figure 1b shows an exemplary PAC representation. The molecular PAC representation was also implemented in-house in Java. STC visualizations of projections were generated to further differentiate numerically comparable optimization solutions and view subsets of top-ranked compounds in the context of global data distributions from multi-dimensional feature space.

For comparison, principal component analysis (PCA) of unweighted and weighted descriptor spaces was carried out using R [29] and the first and second principal components (PCs) were used to generate conventional PC plots. Because these plots generate a two-dimensional view of multi-dimensional data that maximize the original variance, they are often used to represent high-dimensional spaces. However, their primary goal is the generation of an uncorrelated view with maximum variance and hence the visualization might not be chemically informative.

Results and discussion

Methodological principles

The simultaneous consideration of multiple properties beyond potency is a requirement of compound optimization in medicinal chemistry. Therefore, the activity-centric AL concept, which is useful for SAR exploration, might be further extended to rationalize multi-property landscapes. Analyzing multi-dimensional property spaces generally is a complicated task, which is typically addressed using dimensionality reduction. The basic idea underlying the methodology introduced herein was to visualize compound distributions in multi-property space in which numerical optimization is carried out. Multi-property optimization carried out in the context of our analysis was guided by the use of approved drugs as internal standards. Compound rankings based upon projections with a significant enrichment of drugs at top ranked positions were thought to originate from drug-like subspaces in multi-dimensional property space. Thus, highly ranked data set compounds had property combinations comparable to drugs and were thus considered preferred candidates for selection and further optimization efforts. A known conundrum of numerical multi-objective optimization is that typically a variety of high-scoring solutions are obtained that are difficult to distinguish. Therefore, it was attempted to visualize compound distributions underlying best projections to analyze rankings within the data set context and further differentiate them. These visualizations were designed to provide a detailed view of multi-property landscapes, as discussed in the following.

Multi-property landscape display

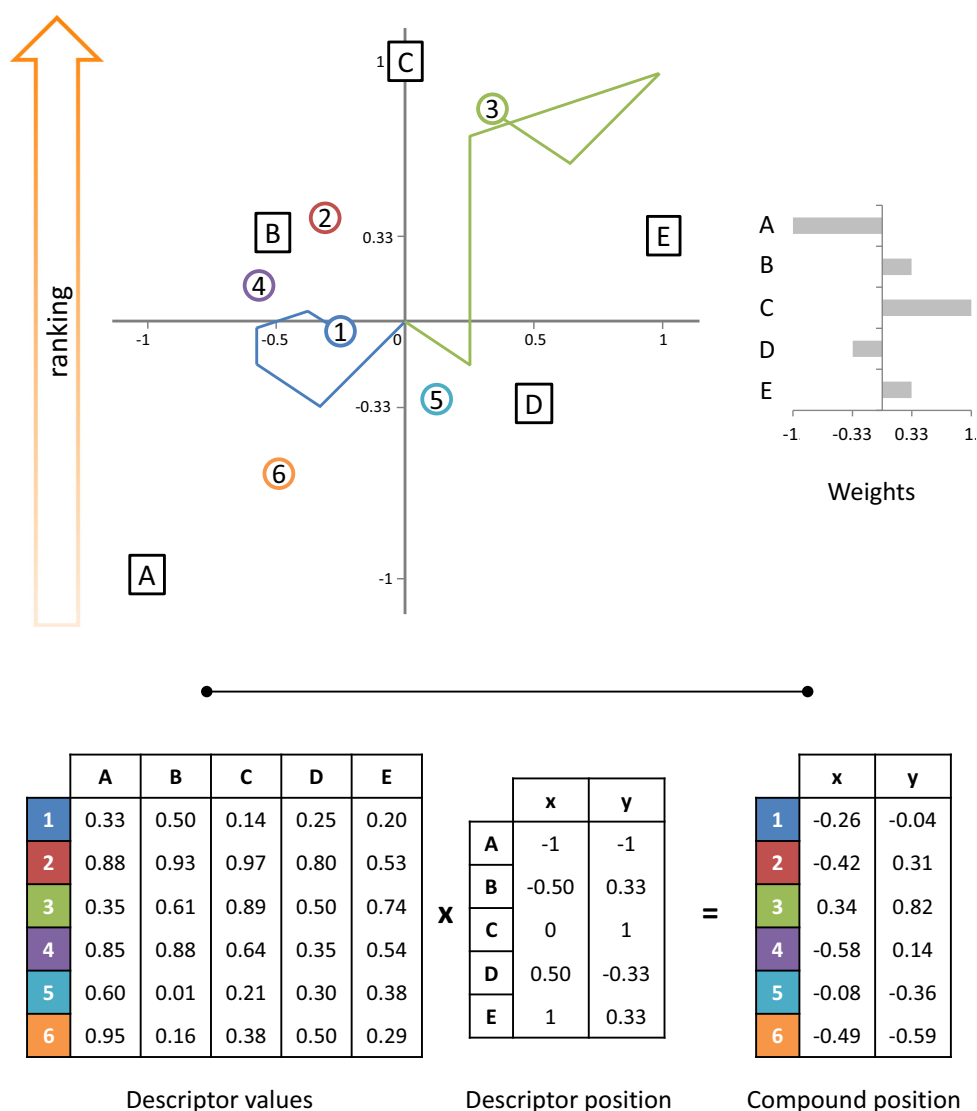
The STC representation provides the core visualization of multi-property space. Figure 2 illustrates how an STC view is obtained for a model compound set from descriptor weights and multi-dimensional coordinates. For a given projection, the STC visualization provides a 2D view of the underlying compound distributions in multi-dimensional property space. Figure 3a shows an exemplary STC visualization for an actual compound data set and a given projection. In addition, Fig. 3b shows the corresponding PAC view and Fig. 3c the top five compounds from the ranking. The five compounds have similar chemical structure. Hence, MOF value ranking likely includes a similarity-property principle component. At the top of the STC view, the drug-like subspace is delineated by the subset of highly-ranked compounds including 13 drugs, with the majority of compounds being

clearly separated from the prioritized subspace (Fig. 3a). Similarity relationships between compounds in STC views were substantially different from those in high-dimensional space (on average 21.5 % nearest neighbors overlap). Comparable average overlap values (12.8–25.2 %) were obtained for other STC views shown in Fig. 5b, d. The corresponding PAC representation reveals which descriptor contributions dominate the projection (Fig. 3b). For some descriptors, values of highly ranked compounds significantly differed (e.g., *a*_acc, logP(o/w)), whereas their values were narrowly confined in other cases (e.g., *a*_don, *a*_ringR, *b*_rotR). Moreover, largely distinct value ranges of a few descriptors were observed for highly ranked molecules compared to many other bioactive compounds (e.g., *a*_ringR, *b*_rotR), which strongly contributed to the separation. Thus, the PAC representation complements the STC visualization by identifying property settings that distinguish compounds in drug-like subspaces from others and evaluating relationships between descriptor settings. Thus, PAC representations can be used to study feature correlation patterns. For example, the line traces in Fig. 3b reveal a negative correlation between the *a*_ringR and *b*_rotR descriptors. Finally, PAC also provides a visual representation of the original high-dimensional space, as it displays all descriptor values for each compound. Therefore, the PAC representation is independent of specific projections and helpful to analyze the STC view.

Multi-property optimization

We next carried out a systematic multi-property optimization as a basis for practical applications of the newly introduced visualization approach. The set of chemically intuitive features selected for our conceptual investigation can be replaced by any other calculated or experimentally determined compound characteristics relevant for optimization tasks. For the multi-objective function containing our 14-descriptor set with four possible weights per descriptor, a systematic search of all possible projections from multi-dimensional space was carried out. Each projection yielded a MOF value for any bioactive compound and drug based on which a ranking was generated. More than 270 million weight combinations were analyzed and prioritized based on the number of drugs in the top 20 compound ranking. For all data sets, drug enrichment was only detected in a small subset of possible weight combinations, as shown in Fig. 4. Hence, delineation of subspaces populated with drugs required very specific multi-parameter settings, as one should expect. Nonetheless, for the different data sets, there were between 20 and ~500 projections that yielded maximum

Fig. 2 Star coordinate representation of multi-dimensional compound data. For a model data set comprising six compounds with five different properties, the generation of an STC view is illustrated. Descriptor positions and compound 1 correspond to Fig. 1a. Descriptor positions resulted from lexicographical ordering along the *horizontal axis* combined with weight settings for a given projection (with a descriptor weight combination shown in the *inset*). Compound positions resulted from matrix calculations shown at the *bottom* and summation of descriptor contributions (pathway calculations). For two exemplary compounds, 1 and 3, pathways are traced. Compound rank positions increase along the *vertical axis*



drug enrichment (between nine and 18 drugs for the different sets), as also shown in Fig. 4. Thus, these projections represented equivalent numerical optimization solutions. The corresponding compound rankings covered most drugs in the data sets (43–95 %; on average 70 %) but only a small fraction of bioactive compounds (3–16 %; on average 8 %) mapping to drug-like subspaces. Furthermore, many projections producing maximum drug enrichment had very similar weight combinations. However, projections with very different combinations (descriptor contributions) were also found. Therefore, solutions with maximal drug enrichment having similar or distinct weight combinations were further analyzed through visualization. The successful delineation of specific drug-like subspaces for all data sets indicated that the search procedure took compound similarity relationship implicitly into account.

Visualization of projections and comparison of compound distributions

A large number of STC representations were generated for different data set projections. Figure 5 shows exemplary comparisons. In Fig. 5a, two projections with distinct descriptor weight combinations are shown for beta-2 adrenergic receptor ligands that produced large drug enrichment (and shared 11 of 13 drugs in their top 20 rankings). Figure 5b compares the STC representations of these projections. The compound distributions differed significantly for these two projections representing numerically equivalent optimization solutions. This might be expected because distinct weight combinations characterized these projections. Although both projections displayed significant drug enrichment, projection 1 clearly separated top ranked compounds from others and also

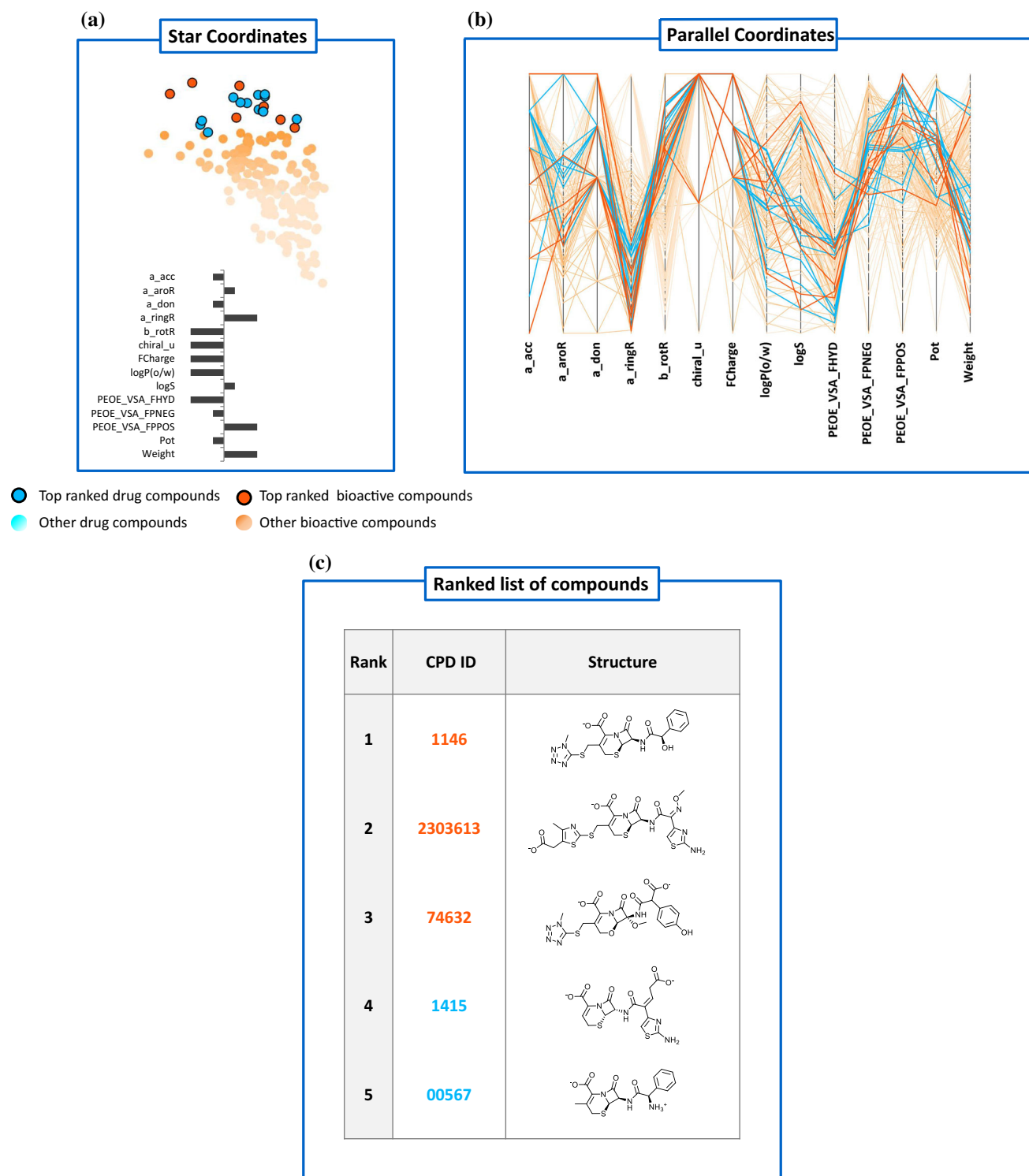


Fig. 3 Views of a multi-property landscape. Compound distributions of small intestine oligopeptide transporter ligands (ChEMBL target ID 4605) for a given projection were displayed using **a** STC and **b** PAC representations. In the STC representation, points represent individual compounds and color-coding distinguishes drugs (cyan) and bioactive compounds (orange). Top ranked molecules are depicted with a black border. In addition, shading of compounds indicates their rank, from dark colors (high rank, beginning at rank

21) to light colors (low rank). In the PAC representation, descriptors are assigned to vertical evenly spaced lines (spanning their value ranges) and compounds are depicted as lines (horizontal traces) color-coded as in (a). **c** The top five compounds from the ranking of the projection including two drugs (4 and 5). Orange and cyan compound (CPD) IDs correspond to ChEMBL and DrugBank IDs, respectively

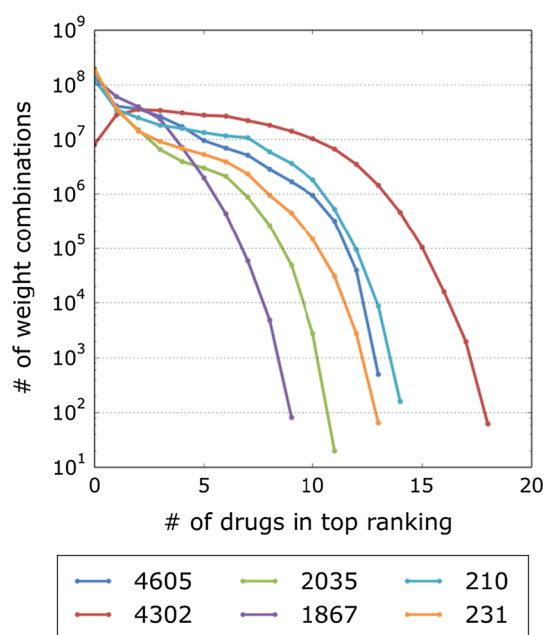


Fig. 4 Numerical comparison of projections. A projection was created for each weight value setting of the multi-objective function containing 14 descriptors and the number of drugs within the 20 top ranked compounds was determined. The graph reveals the number of weight combinations yielding largest numbers of highly-ranked drugs across the different target sets (colored by target IDs given in Table 1)

spread the compound data set across the property space, providing a clear view of compounds in increasingly large distances from the prioritized subspace. By contrast, in projection 2, the bulk of the data set was concentrated in a small region of property space and the separation of highly ranked and other compounds was only marginal. Hence, the property settings of projection 2 rendered data set compounds much more similar in multi-dimensional property space than the settings of projection 1, as clearly revealed by STC visualization. Therefore, for the selection of candidate compounds for chemical optimization efforts focusing on drug-like subspace, preference would be given to projection 1.

In Fig. 5c, two projections with similar descriptor weights are shown for alpha-2a adrenergic receptor ligands that yielded large drug enrichment (and shared seven of eight drugs among the top 20 compounds). Figure 5d shows the STC visualizations of these projections. Although the weight combinations were very similar, the compound distributions were distinct, contrary to expectations, as further discussed below.

For comparison with STC, Fig. 5e shows PC plots (using the first and second PC) of the unweighted descriptor space and weighted descriptor combinations of projections 1 and 2. In unweighted descriptor space, PCA did not yield a separation of drugs and bioactive compounds. Moreover, the PC plots of projection 1 and 2 were very difficult to

interpret and remained essentially inconclusive. By contrast, the STC representations of projection 1 and 2 in Fig. 5d reveal a clear separation of top ranked and other data set compounds, but with different characteristics. The STC view of projection 1 shows that many data set compounds including remaining drugs were located proximal to the prioritized subspace, while only a small number of lowly ranked compounds were far removed from it. However, the STC view of projection 2 in Fig. 5d reveals a significant spread of the compounds across multi-dimensional property space (similar to projection 1 in Fig. 5b) including the majority of drugs, although the weight settings of projection 1 and 2 were comparable. In the case of projection 2, the STC view also shows that the drug-like subspace was less well-defined than in other cases, with many drugs (including two highly ranked ones) located distantly from many top ranked compounds. From these STC views, individual compounds can be easily selected for further analysis. Taken together, the STC visualizations provided a well-resolved picture of compound distributions in multi-dimensional property space for otherwise very similar projections.

Concluding remarks

High-dimensional property spaces for compound optimization or data set analysis are generally difficult to represent and navigate. While the potency-centric AL concept has substantially contributed to graphical SAR exploration, especially for larger and structurally heterogeneous data sets, little efforts have thus far been made to visualize multi-dimensional property landscapes that combine activity with other optimization-relevant properties. Typically, dimension reduction techniques such as PCA are applied to evaluate feature contributions in multi-dimensional space. Different types of graphical analysis are expected to aid in the rationalization of multi-dimensional property spaces. Therefore, a visualization methodology for multi-dimensional property spaces has been developed, as reported herein. Our analysis was based upon the generation of drug-like subspaces in chemical space, which takes molecular similarity relationships implicitly into account. However, it would also be feasible to focus an analysis explicitly on selected distance relationships in chemical space (or generate subspaces for compound reference sets with other characteristic properties).

Our study introduces the STC and PAC concepts, adapted from computer graphics, to the medicinal chemistry community. STC/PAC visualization of compound data is designed to complement multi-objective optimization, provide access to multi-dimensional data distributions, and aid in compound selection. For a given

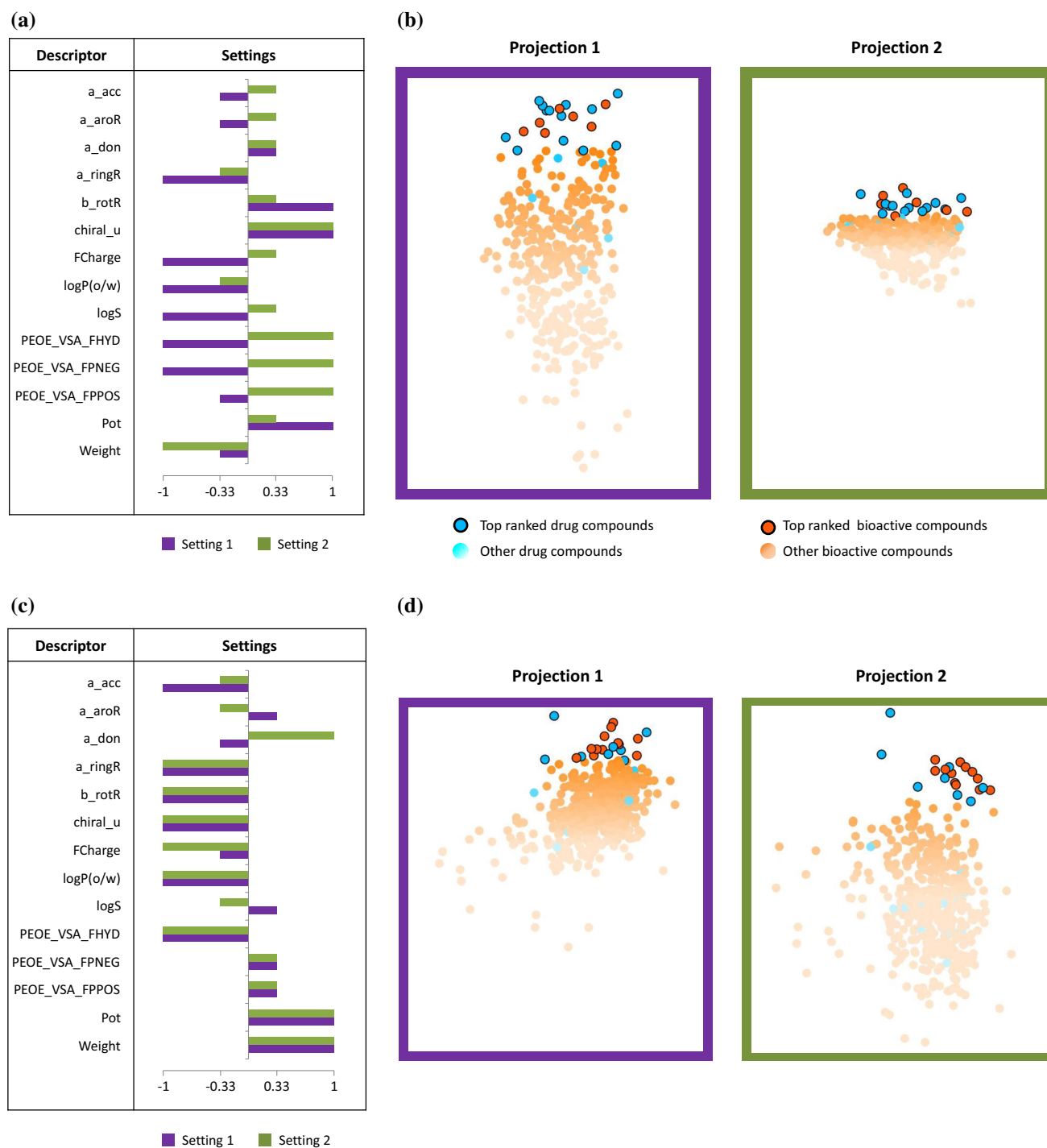


Fig. 5 Visualization of projections. Exemplary projections are visualized and compared. In **(a)** and **(b)**, two projections generated for beta-2 adrenergic receptors (ChEMBL target ID 210) are shown. The corresponding top 20 rankings contained 13 drugs each (11 of which were the same). **a** Compares the weight combinations (settings) for these projections and **b** their STC visualizations. *Points* represent individual compounds and are *color-coded* according to Fig. 3a. In **(c)** and **(d)**, two projections generated for alpha-2a adrenergic receptor ligands (ID 1867) are shown. The corresponding top 20

rankings contained eight drugs each (seven of which were the same). **c** Compares the weight combinations (settings) for these projections and **d** their STC visualizations. In **(b)** and **(d)**, STC visualizations were scaled to the same value ranges. **e** PCA-based data set projections (using the first two PCs) with unweighted descriptors (*top*, drugs colored *cyan* and bioactive compounds *gray*) and weighted descriptors from projection 1 (*middle*) and 2 (*bottom*) taken from **(c)**. PCA plots of projections are color-coded as in **(d)**

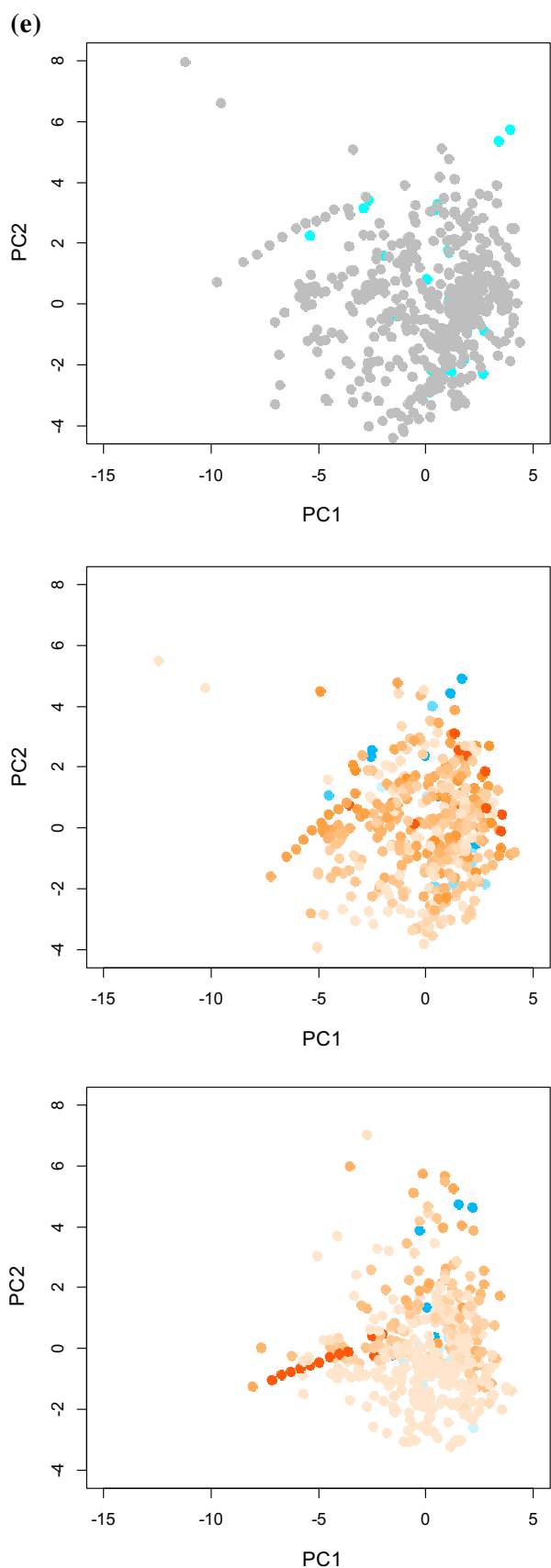


Fig. 5 continued

projection and compound ranking, the STC visualization provides a 2D representation of a compound distribution in multi-dimensional property space and views highly ranked compound subsets in the data set context. In addition, the PAC representation compares individual property contributions and identifies property settings that distinguish highly ranked compounds from others. We have demonstrated that STC visualizations help to differentiate numerically equivalent optimization solutions with similar or distinct property settings. The data sets used herein are made freely available [30].

References

1. Stumpfe D, Bajorath J (2012) Methods for SAR visualization. *RSC Adv* 2:369–378
2. Wassermann AM, Wawer M, Bajorath J (2010) Activity landscape representations for structure-activity relationship analysis. *J Med Chem* 53:8209–8223
3. Shanmugasundaram V, Maggiora GM (2001) Characterizing property and activity landscapes using an information-theoretic approach. In: Proceedings of 222nd American chemical society national meeting, division of chemical information, Chicago, IL, August 26–30, 2001; American Chemical Society: Washington, D.C., 2001; abstract no. 77
4. Wawer M, Peltason L, Weskamp N, Teckentrup A, Bajorath J (2008) Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J Med Chem* 51:6075–6084
5. Wollenhaupt S, Baumann K (2014) inSARa: Intuitive and interactive SAR interpretation by reduced graphs and hierarchical MCS-based network navigation. *J Chem Inf Model* 54:1395–1409
6. Agrafiotis DK, Shemanarev M, Connolly PJ, Farnum M, Lobanov VS (2007) SAR maps: a new SAR visualization technique for medicinal chemists. *J Med Chem* 50:5926–5937
7. Wassermann AM, Bajorath J (2012) Directed R-group combination graph: a methodology to uncover structure-activity relationship patterns in a series of analogues. *J Med Chem* 55:1215–1226
8. Peltason L, Weskamp N, Teckentrup A, Bajorath J (2009) Exploration of structure-activity relationship determinants in analogue series. *J Med Chem* 52:3212–3224
9. Wawer M, Bajorath J (2010) Similarity-potency trees: a method to search for SAR information in compound data sets and derive SAR rules. *J Chem Inf Model* 50:1395–1409
10. Peltason L, Iyer P, Bajorath J (2010) Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J Chem Inf Model* 50:1021–1033
11. Reutlinger M, Guba W, Martin RE, Alanine AI, Hoffmann T, Klenner A, Hiss JA, Schneider P, Schneider G (2011) Neighborhood-preserving visualization of adaptive structure-activity landscapes: application to drug discovery. *Angew Chem Int Ed* 50:11633–11636
12. Zwierzyna M, Vogt M, Maggiora GM, Bajorath J (2015) Design and characterization of chemical space networks for different compound data sets. *J Comput-Aided Mol Des* 29:113–125
13. Ertl P, Rohde B (2012) The molecule cloud-compact visualization of large collections of molecules. *J Cheminf* 4:12

14. Awale M, van Deursen R, Reymond J-L (2010) MQN-maplet: visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. *J Chem Inf Model* 50:1395–1409
15. Reymond J-L (2015) The chemical space project. *Acc Chem Res* 48:722–730
16. Kireeva N, Baskin II, Gaspar HA, Horvath D, Marcou G, Varnek A (2012) Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling, and dataset comparison. *Mol Inf* 3(4):301–312
17. Wermuth CG (2008) *The practice of medicinal chemistry*, 3rd edn. Academic Press-Elsevier, Burlington, London
18. Gillet VJ, Khatib W, Willett P, Fleming P, Green DVS (2002) Combinatorial library design using multiobjective genetic algorithm. *J Chem Inf Comput Sci* 42:375–385
19. Gillet VJ (2004) Applications of evolutionary computation in drug design. *Struct Bond* 110:133–152
20. Nicolaou CA, Brown N, Pattichis CS (2007) Molecular optimization using computational multi-objective methods. *Curr Opin Drug Discov Develop* 10:316–324
21. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40:D1100–D1107
22. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 42:D1091–D1097
23. OEChem TK (2012) OpenEye scientific software Inc, Santa Fe, NM, USA
24. Molecular Operating Environment (2012) Chemical computing group Inc.: Montreal, Quebec, Canada
25. Cook D, Buja A, Lee EK, Wickham H (2008) Grand tours, projection pursuit guided tours and manual controls. In: Chen C, Härdle W, Unwin A (eds) *Handbook of data visualization*. Springer, Heidelberg, pp 295–314
26. Kandogan E (2000) Star coordinates: a multi-dimensional visualization technique with uniform treatment of dimensions. In: *LBHT Proc IEEE information visualization symposium*, pp 9–12
27. Java universal network/graph framework. <http://jung.sourceforge.net/>. Accessed May 1, 2014
28. Inselberg A (1985) The plane with parallel coordinates. *Visual Comput* 1:69–91
29. R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria, 2012
30. de la Vega de León A, Kayastha S, Dimova D, Schultz T, Bajorath J (2015) ChEMBL20 data sets for multi-property landscape analysis. ZENODO. doi:10.5281/zenodo.21782

Conclusions

We have performed a systematic search of projections of the high-dimensional space on the basis of a multi-objective function. This function was a weighted linear combination of descriptor values and was used to prioritize compounds. The selection of weights was modeled as an optimization task, where the number of drug compounds with largest function values was maximized. Different weight combinations gave identical results and were considered numerically equivalent by the optimization. However, when they were used to create star coordinate plots to explore these drug-like subspaces, compound distributions were very different even among similar weight combinations. Relationships between different descriptors were analyzed with parallel coordinate plots. These plots could also differentiate descriptor value distributions between drugs and bioactive compounds.

The previous two studies have revealed applications of coordinate-based visualizations in different multi-objective optimization settings. The use of these plots can offer an overview of the data, like principal component plots, or can focus on drug-like subspaces, like star coordinate. Nonetheless, distances in these plots may not reflect true similarity relationships in high-dimensional space. In the next study, we develop a novel coordinate-free representation of high-dimensional space that better emphasizes important similarity relationships.

9 Chemical space visualization: transforming multi-dimensional chemical spaces into similarity-based molecular networks

Introduction

Chemical space can be portrayed in different ways. Coordinate-free (such as graphs) and coordinate-based displays (such as scatter plots) are both used in chemoinformatics. Molecular representations frequently dictate which visualization type is applied. Networks are traditionally built for fingerprint-based similarity and substructure relationships. Molecular descriptors are displayed through coordinate-based plots, such as those described in the last two chapters. The analysis of descriptor distance relations in network representations has not received much attention. Here, we introduce a novel coordinate-free visualization based on chemical space networks (CSNs). Distance relations in high-dimensional property space are transformed to similarity relations and used to build networks. This visualization is called transformation-CSN (TRANS-CSN) because of its ability to transform coordinate-based property spaces to coordinate-free network representations.

This study is currently in press at the Future Medicinal Chemistry journal with no formatted article. The submitted manuscript is shown as a preview.

Chemical space visualization: transforming multi-dimensional chemical spaces into similarity-based molecular networks

Antonio de la Vega de León and Jürgen Bajorath

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany.

Background: The concept of chemical space is of fundamental relevance for medicinal chemistry and chemical informatics. Multi-dimensional chemical space representations are coordinate-based. Chemical space networks (CSNs) have been introduced as a coordinate-free representation. **Results:** A computational approach is presented for the transformation of multi-dimensional chemical space into CSNs. The design of transformation CSNs (TRANS-CSNs) is based upon a similarity function that directly reflects distance relationships in original multi-dimensional space. **Conclusions:** TRANS-CSNs provide an immediate visualization of coordinate-based chemical space and do not require the use of dimensionality reduction techniques. At low network density, TRANS-CSNs are readily interpretable and make it possible to evaluate structure-activity relationship (SAR) information originating from multi-dimensional chemical space.

Introduction

The concept of chemical space is popular in medicinal and computational chemistry [1, 2]. Chemical space is generally rationalized as the union of all chemically feasible compounds [1]. With on the order of 10^{60} possible small molecules [2], chemical space is ultimately finite but so vast that it cannot be studied or represented in its entirety. Rather, only small sections of theoretically possible chemical space are typically explored, in particular, biologically relevant chemical space [1], which is populated with biologically active small molecules or compounds having the potential to be active, given their chemical properties. Chemical space is often also intuitively envisioned as a multi-dimensional space across which compounds are distributed in a star-like manner [1]. In computational chemistry and chemical informatics, coordinate-based chemical space representations essentially mimic such an imaginary multi-dimensional space [3, 4]. Since there is no generally accepted or applicable computational representation of chemical space, coordinate-based representations make use of varying numbers of molecular descriptors, i.e., more or less complex mathematical models of chemical structure and/or properties, to generate a coordinate system into which compounds are placed based on their descriptor values. Depending on the specific requirements of different applications, such as compound classification, diversity analysis, or activity prediction, the composition of descriptor spaces varies, but they have in common that they are multi- or high-dimensional. Accordingly, visualization of coordinate-based chemical space representations is not straightforward and requires the application of statistical dimensionality reduction techniques [3, 5] such as principal component analysis (PCA) [1, 3] or multi-dimensional scaling (MDS) [5] to generate two- or three-dimensional projections of chemical space.

As an alternative to coordinate-based chemical space, coordinate-free representations can also be considered,

which are generated by determining all pair-wise relationships between compounds [6]. For example, a rudimentary coordinate-free space representation is provided by a similarity value matrix of a compound collection. As more advanced representations, similarity-based molecular networks have been introduced as a paradigm for coordinate-free chemical space representation [6]. In such networks, nodes represent compounds and edges connecting pairs of nodes represent similarity relationships. However, distances between nodes in network representations do not correlate with chemically relevant distances and this is why they are regarded as coordinate-free representations. From an algorithmic viewpoint, they are indeed coordinate-free because they are generated from pair-wise similarity matrices. These similarity-based compound networks were originally generated on the basis of fingerprint-based Tanimoto similarity values using pre-defined threshold values for SAR analysis [7]. They have also been applied to explore chemical libraries [8] and the applicability domain of QSAR models [9]. Chemical space networks (CSNs) [6] represent a generalization of similarity-based compound networks. The major determinant of the topology of CSNs, their characteristic features, and SAR information content is the way in which molecular similarity is accounted for, i.e., the choice of similarity measures [6], which has been systematically investigated in CSNs of different design. Therefore, although CSNs have originally been introduced on the basis of Tanimoto similarity [10], other types of CSNs have been constructed on the basis of substructure-based similarity [11], hybrid measures combining numerical similarity measures and substructure similarity [12], or asymmetric similarity functions, leading to the presence of directed edges in CSNs [13]. As coordinate-free chemical space representations, CSNs have the advantage that they provide an immediate visualization of chemical space and enable its interactive navigation [6].

Table 1: Compound sets^a

Target ID	Target Name	Compounds
210	Beta-2 adrenergic receptor	374
231	Histamine H1 receptor	597
1867	Alpha-2a adrenergic receptor	476
2035	Muscarinic acetylcholine receptor M5	296
4302	P-glycoprotein 1	291
4605	Small intestine oligopeptide transporter	195

^aFor each data set, the ChEMBL target ID, target name, and the number of active compounds is reported.

Herein, we introduce a new computational approach for chemical space display, which establishes a methodological link between coordinate-based and coordinate-free space representations. Therefore, multi-dimensional chemical spaces based on continuous descriptor values (rather than fingerprints) are transformed into a new type of CSN using a specifically defined similarity function that converts inter-compound distances in multi-dimensional space into scaled similarity relationships. This makes it possible to directly visualize multi-dimensional chemical space representations in a new format, without the need for dimension reduction techniques. Given the uniqueness of the descriptor coordinate-to-similarity transformation, these CSNs conceptually differ from previous network representations and provide a view of transformed chemical space. The central aspect of this study is the transformation of coordinate-based into coordinate-free chemical space views. The resulting CSNs are shown to be informative for structure-property analysis.

Methods

Compound sets

For chemical space transformation, six previously reported compound sets taken from ChEMBL (version 20) [14] and DrugBank (version 4.1) [15] were used. These data sets were originally used to delineate drug-like subspaces in multi-dimensional chemical space [16]. Bioactive molecules were extracted from ChEMBL for which direct interactions with a human target at the highest level of assay confidence were reported. Only inhibition constant (K_i) values were considered as activity measurements. Compounds with multiple potency records that differed by more than one order of magnitude were not considered. If all values fell within the same order of magnitude, the arithmetic mean of all reported K_i values was used as the final activity value. Potency values were recorded as pK_i values, i.e., the negative decadic logarithm of K_i values. Compound sets from ChEMBL were complemented with drugs from DrugBank for which activity against the same target was reported and potency values were available in ChEMBL. Table 1 reports the composition of the six compounds sets.

Descriptors

Molecules were represented using a set of 13 numerical molecular descriptors reported in Table 2. These

descriptors were previously selected to account for chemically intuitive features with relevance for SARs and shown to yield resolved compound distributions in multi-dimensional space [16]. Hence, we had prior evidence that this descriptor set, albeit limited in size, was suitable for chemical space transformation. Among others, chemical features accounted for by these descriptors that were relevant for biological activity included hydrogen bond potential, flexibility/rigidity, solubility, and various molecular surface characteristics. All structural descriptors except a_ringR (ring content of a molecule) were calculated using the Molecular Operating Environment [17]. The a_ringR descriptor was calculated using an in-house script based upon the OpenEye toolkit [18].

Coordinate-based chemical space representation

The numerical descriptors were used to generate a 13-dimensional chemical space. For each compound, the descriptors were calculated and scaled to unit variance to ensure equivalent contributions to distance relationships in chemical space. Scaled descriptor values were then used as a coordinate vector to define the position of each compound in multi-dimensional space.

Two-dimensional chemical space projections

PCA and MDS were used to project multi-dimensional space onto a two-dimensional representation. They represent standard approaches for dimension reduction of multi-dimensional chemical space representations and are conceptually distinct. PCA creates a set of uncorrelated principal components from linear combinations of original descriptors that capture the variance within the original data set. The first two principal components were used as axes to generate a scatterplot representation. MDS accounts for pair-wise distances between data points in original space (rather than global data variance) and attempts to preserve these distance relationships in lower-dimensional space representations. MDS was used here to generate two-dimensional projections of multi-dimensional space. PCA and MDS calculations were carried out using in-house Python scripts based upon the scikit-learn package [19] using default parameter settings.

Table 2: Descriptors^a

Name	Description
a_acc	Number of hydrogen bond acceptors
a_aroR	Fraction of aromatic ring atoms
a_don	Number of hydrogen bond donors
a_ringR	Fraction of ring atoms
b_rotR	Fraction of rotatable bonds
chiral_u	Number of unconstrained chiral centers
Fcharge	Sum of formal charges
logP(o/w)	Log of partition coefficient (octanol/water)
logS	Log of aqueous solubility
PEOE_VSA_FHYD	Fraction of hydrophobic surface area
PEOE_VSA_FPNEG	Fraction of negative polar surface area
PEOE_VSA_FPPOS	Fraction of positive polar surface area
Weight	Molecular weight

^aThe 13 descriptors used to generate the multi-dimensional chemical space are defined.

Chemical space networks

CSNs are generated on the basis of pair-wise compound similarity relationships. Therefore, pair-wise Euclidian distances between compound descriptor vectors in multi-dimensional space were determined and transformed into similarity values within the range [0,1] using the following formula:

$$Similarity = 1 - \frac{distance}{Max(distance)}$$

So-defined similarity values correlated with fractions of the maximal distance in multi-dimensional descriptor space. From the complete similarity matrix, CSN adjacency matrices were calculated for varying similarity threshold values. CSNs were then generated using the Fruchterman-Reingold algorithm [20] and visualized using in-house Java programs based on the JUNG library [21]. The force-directed Fruchterman-Reingold layout algorithm places densely connected objects closely together and separates clusters from each other. As a consequence, edge lengths and distances between compounds and compound clusters in CSNs have no chemical meaning.

Network properties

CSNs were characterized using two major network properties including edge density and modularity [22, 23]. Density represents the fraction of all possible edges that are present in a network and is calculated as follows:

$$Edge\ density = \frac{2m}{n(n-1)}$$

where m is the number of edges and n the number of compounds.

CSNs are best compared and interpreted at low edge densities at which modularity is typically high [10]. Modularity is a measure for the cluster structure of a network. In CSNs, high modularity corresponds to the presence of well-defined compound communities [10]. Modularity values depend on the function $\delta(i, j)$, such that $\delta(i, j) = 1$

if molecules i and j belong to the same cluster and zero otherwise [23]. For modularity calculations, the cluster distribution within a CSN must be algorithmically determined, for which the Newman algorithm was used [24]. Modularity was then calculated as:

$$Modularity = \frac{1}{2m} \sum_{1 \leq i, j \leq n} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(i, j)$$

where m is the number of edges, a_{ij} is the value in the adjacency matrix for compounds i and j and k_i is the degree of compound i .

Results and discussion

Projections of coordinate-based chemical space

Initially, compound data sets were distributed in a 13-dimensional chemical space based on descriptor coordinates. In such coordinate-based representations, each chosen descriptor or feature adds a dimension to multi-dimensional space. These chemical space representations are widely used for applications in chemical informatics such as diversity analysis, compound classification, or activity prediction. In these multi-dimensional spaces, most computational operations are carried out numerically. However, multi-dimensionality, albeit suitable for numerical analysis, complicates graphical exploration of chemical space. In order to visualize coordinate-based chemical space representations, dimension reduction techniques such as PCA or MDS must be employed, which is another standard procedure in chemical informatics.

Figure 1 shows PCA and MDS projections of our 13-dimensional reference space for two exemplary compound sets. Dimension reduction is typically accompanied by a loss of information. For example, the first two principal components shown in Figure 1 account for only 49.9% (set 210) and 48.4% (1867) of the original data variance in multi-dimensional space. PCA and MDS projections yielded different compound distributions. For compound set 210, the PCA projection indicated that

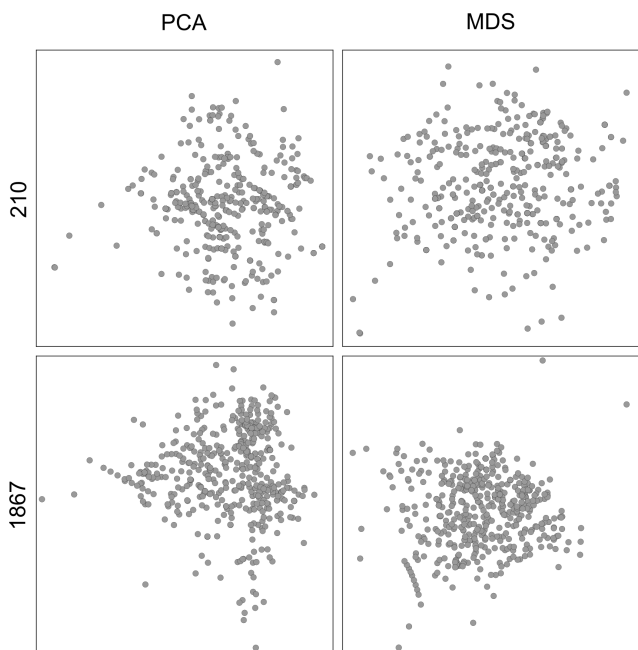


Figure 1: Exemplary chemical space projections. For two different compound sets, ligands of beta-2 adrenergic receptor (210, Table 1) and alpha-2a adrenergic receptor (1867), original multi-dimensional descriptor space was projected onto two dimensions using the first two components from principal component analysis (PCA) or via multi-dimensional scaling (MDS).

most compounds were more similar to each other than it appeared to be in the MDS view. In both projections of set 1867, the bulk of the compounds formed a densely populated central region. On the basis of these projections, it would essentially be impossible to interpret compound relationships in chemical space. Hence, PCA or MDS plots can only provide a rather approximate two-dimensional view of chemical space.

Transformation CSNs

The major aim of our study was to transform multi-dimensional chemical space into an alternative coordinate-free representation that can be directly visualized. A possible approach was the use of CSNs that can be constructed using different similarity measures. This required that pair-wise similarity relationships in CSNs accurately accounted for all possible compound relationships in multi-dimensional space. Importantly, relative compound positions in multi-dimensional space were determined by vectors of numerical property descriptors not taking compound structure directly – or any form of structural relationships – into account. Therefore, inter-compound distances in descriptor space were converted into scaled similarity relationships, which were then used to construct CSNs. The use of property distance-derived similarity (dd-sim) based upon numerical descriptor values in multi-dimensional space is a unique feature of these CSNs compared to earlier designs. Because these CSNs were used for the transformation of coordinate-based chemical space they were designated transformation CSNs (TRANS-CSNs). TRANS-CSNs are by design closely linked to the original multi-dimen-

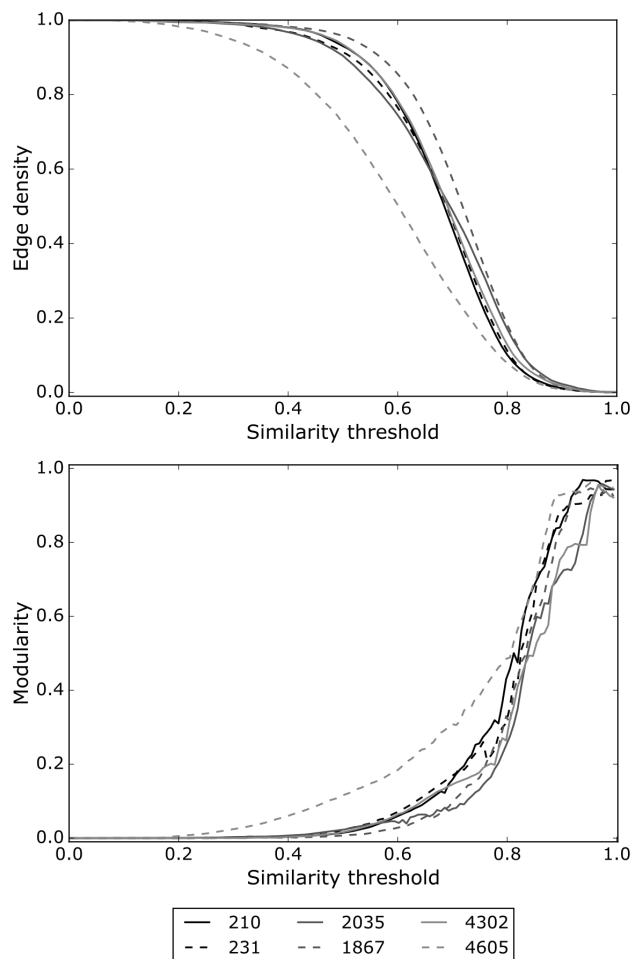


Figure 2: Network properties. Edge density (top) and modularity (bottom) are monitored for representative CSNs with increasing similarity threshold values.

sional representations and aim to convey the chemical information contained in the spaces in a conceptually different manner, making it accessible through immediate visualization.

TRANS-CSN properties

TRANS-CSNs are threshold networks, which have a different edge density for each chosen dd-sim threshold value. In the extreme case, a completely connected network with density of 1 is obtained for a similarity threshold value of 0 (i.e., all compounds are considered similar to each other). On the other hand, a minimally connected network would be obtained for a threshold of 1. In this case, edges would only be drawn between compounds having identical descriptor coordinates.

TRANS-CSNs were generated for our six compound sets from multi-dimensional chemical space under systematic variation of similarity thresholds. Density as a function of similarity thresholds is reported in Figure 2 (top graph). All density curves followed the same path and five of six curves were very similar. Density remained very high for threshold values up to 0.6 before notable reductions were detected. A characteristic feature of the density curves was that they reached a low level of edge density ($\leq 10\%$) at large threshold values (≥ 0.80), also

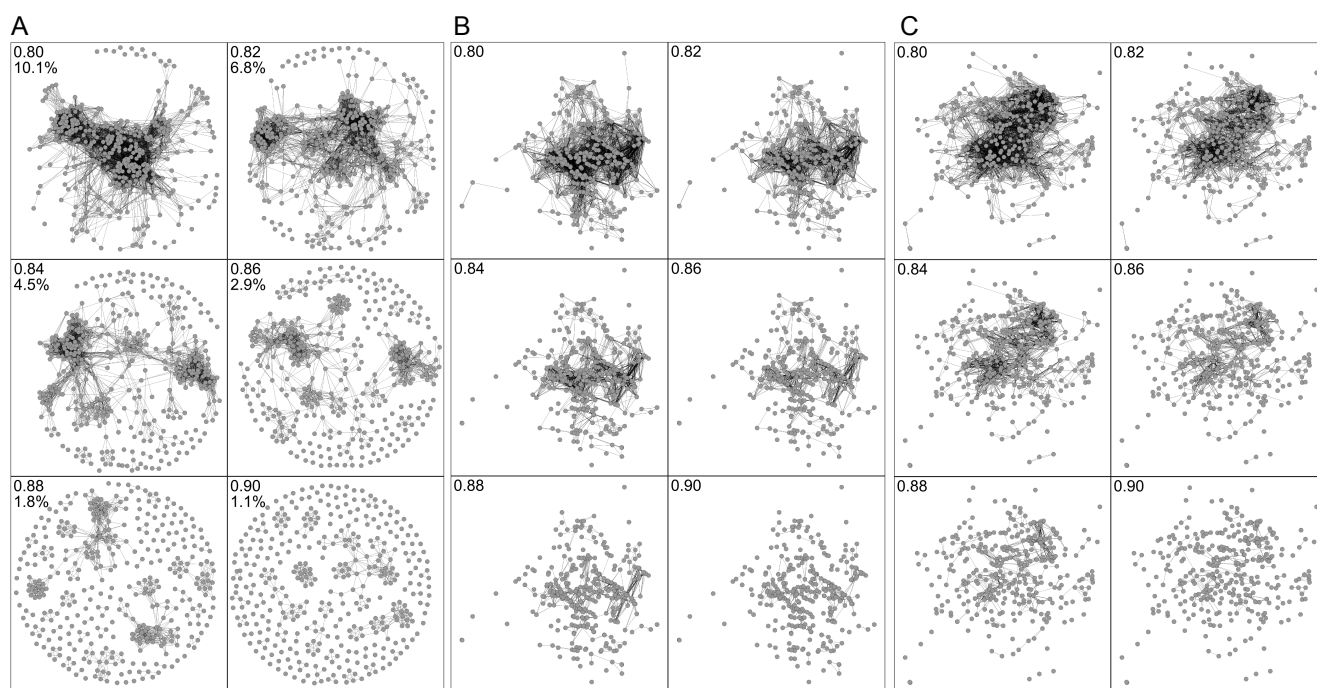


Figure 3: Comparison of TRANS-CSNs to PCA and MDS projections. For the beta-2 adrenergic receptor ligand set (210), (A) shows TRANS-CSNs at six different similarity thresholds. The threshold values and the edge density (in percent) of CSN are reported at the upper left of each graph. (B) shows the PCA projection of the compound set and (C) the MDS projection onto which TRANS-CSN similarity relationships obtained at the six different threshold values are mapped.

indicating that many compounds had similar descriptor coordinates.

The bottom graph in Figure 2 reports modularity as a function of *dd-sim* thresholds and complements the observations made for edge density. Modularity and density curves were essentially inverted mirroring the direct dependence of CSN modularity on edge density. Until a threshold of 0.6, modularity was very low and then began to increase. A sharp increase occurred for threshold values greater than 0.80. In this region, minute changes in density lead to small-magnitude fluctuations in modularity, leading to a more rugged appearance of the modularity than the density curves. At large *dd-sim* threshold values, when network density was less than 10%, modularity reached high values greater than 0.9.

Figure 3A shows corresponding TRANS-CSNs for an exemplary compound set, beta-2 adrenergic receptor ligands (set 210). These network views very well illustrate the relationships between *dd-sim* thresholds, density, and modularity and also reveal that TRANS-CSNs were capable of resolving relationships between compounds in chemical space much better than projections using dimension reduction techniques. At density levels of 10.1% and 6.8%, the CSNs still contained a large and densely connected central network component. However, at density levels of 4.5% and 2.6%—corresponding to *dd-sim* thresholds of 0.84 and 0.86, respectively—separate compound communities emerged and the networks became readily interpretable at a global level. At further decreasing density of 1.8% and 1.1%, communities were gradually dissolved, due to low connectivity. Hence, there was a delicate balance between threshold values, modularity, and global appearance for TRANS-CSNs. The relation between similarity threshold and density

values for sets 4605 and 231 was very similar to set 210. A similarity threshold value of 0.84 was suitable to create interpretable TRANS-CSN views of 3.6% and 4.3% density, respectively. For data sets 1867 and 4302, a slightly increased threshold value of 0.86 was required to yield equivalent resolution and interpretability of CSNs. At this marginally higher threshold value, TRANS-CSNs of data sets 1867 and 4302 had density values of 4.6% and 4.1%, respectively. For set 2305, the same threshold value of 0.86 was used and produced an edge density that was slightly larger (5.3%). Overall, there was only little variation in threshold values and ensuing edge densities of TRANS-CSNs for these compound sets, which displayed comparable resolution.

In Figure 3B, similarity relationships resulting from different *dd-sim* thresholds were mapped onto the PCA projection of the compound set, i.e., compound pairs with similarity meeting the threshold were connected. Mapping of similarity relationships showed that they could hardly be interpreted on the PCA background and, in addition, that relative compound positions often significantly differed in TRANS-CSNs and the PCA projection. Equivalent observations were made when mapping similarity relationships onto the MDS projection of the set, as shown in Figure 3C. By contrast, for TRANS-CSNs, it was possible to inspect network views at threshold values between 0.8 and 0.9 and select CSN views that best resolved compound relationships and were interpretable in a meaningful way.

Structure-activity relationships

One of the prime applications of chemical space representation, however they might be generated, is the analysis

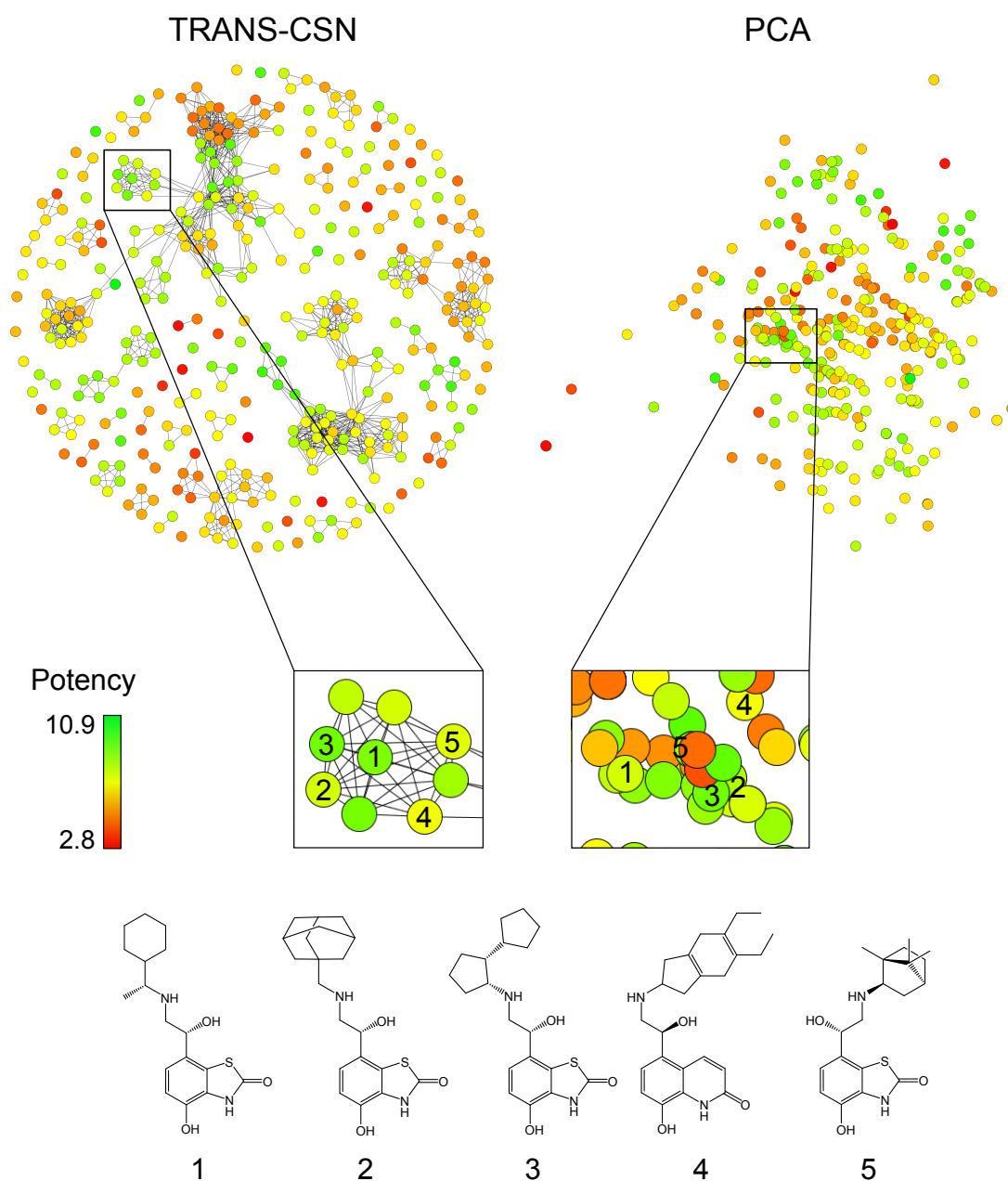


Figure 4: Compound mapping. Shown is a side-by-side comparison of a TRANS-CSN (dd-sim threshold of 0.88) and a PCA projection for the set of beta-2 adrenergic receptor ligands. Nodes are colored by compound potency using a continuous spectrum from red (lowest potency) over yellow to green (highest). Two corresponding regions in these graphs are delineated and enlarged. Five compounds are labeled (1-5) and their structures are shown.

of compound-property relationships including –first and foremost– structure-activity relationships (SARs). For SAR exploration, compounds are annotated with activity information and subsets of similar compounds are analyzed. For this purpose, the focus changes from a global view, required for studying compound distributions in chemical space, to a local view to identify compounds having different SAR characteristics. An example of a TRANS-CSN with potency information is shown in Figure 4. The TRANS-CSNs for the beta-2 adrenergic receptor ligand set at the threshold of 0.88 was color-coded according to compound potency. The annotated TRANS-CSN revealed a variety of compound communities with different potency distributions, which encoded different local SARs. A representative community is highlighted

in the TRANS-CSN that consisted of compounds having comparably high potency. The corresponding region was mapped in the PCA projection of the compound set where it had largely different composition, also including compounds with very low potency. Here, compound relationships were not resolved, as most compound positions overlapped, and local SARs were difficult, if not impossible to interpret. By contrast, compound communities in the TRANS-CSN generally displayed well-defined SAR characteristics, as shown in Figure 4. In many cases, compounds with similar (high or low) potency formed communities; in others, communities contained pairs of compounds with large potency differences (activity cliffs). Taken together, these observations indicated that the original chemical space representation using numer-

ical descriptors with emphasis on chemically intuitive properties was sensitive to SARs and, in addition, that TRANS-CSNs –as a coordinate-free representation– provided a high-resolution view of this space. A primary reason for this observation likely was the way in which molecular similarity or dissimilarity was accounted for. TRANS-CSNs are based upon compound similarity relationships that directly reflect distance relationships in original coordinate-based chemical space, whereas dimension reduction methods account for relative compound positions and distances in different, often more indirect ways. Thus, TRANS-CSNs are thought to be a meaningful alternative representation of multi-dimensional chemical space.

Conclusions

Chemical space can be rationalized in different ways and there is no generally applicable representation of chemical space. Visualization and navigation of chemical space continue to be challenging tasks in computational medicinal chemistry and chemical informatics. Multi-dimensional coordinate-based reference spaces have been used for many applications. As an alternative, coordinate-free space representations can also be considered. CSNs have recently been introduced as a paradigm for coordinate-free chemical space display. So far, coordinate-based and coordinate-free representations have been used in a mutually exclusive manner. Herein, we have introduced a computational approach to transform multi-dimensional chemical spaces into a new type of CSN using a similarity function that converts inter-compound distances in original space into scaled similarity values. Therefore, TRANS-CSNs directly capture relationships between compounds in multi-dimensional space and enable immediate visualization. Hence, these molecular networks provide a link between methodologically distinct ways of representing chemical space and further advance our ability to navigate and interpret biologically relevant sections of chemical space.

Future perspective

The development of methods for coordinate-free representation of chemical space is still in its early stages. CSNs are currently the most advanced representations and of particular interest for the analysis of bioactive compounds and their relationships. However, SAR applications are just beginning to be explored. In our view, the introduction of TRANS-CSNs is another important step forward because they are intimately linked to conceptually distinct coordinate-based chemical space representations and thus enable previously unconsidered applications. Of course, the ability to visualize multi-dimensional coordinate-based spaces without dimensionality reduction is an advance in itself, and it is anticipated that many graphical analyses of such spaces will be carried out. However, there may be more. For example, since most multi-dimensional chemical space constructs are based upon numerical property descriptor values,

they are only indirectly related to compound structure and activity. A key question in the generation of such reference spaces often is whether or not they might be activity-sensitive and suitable to explore SARs in a meaningful way. TRANS-CSNs make it possible to directly evaluate this question, as also discussed herein, because they can be used to analyze subsets of compounds that are closely related to each other in original property spaces –and emerge as compound communities in CSNs– and evaluate whether or not SAR information associated with such compounds is interpretable and makes sense. Do compounds that are close to each other in feature space often have similar activity? Are there well-defined activity cliffs? Or are activity values more or less randomly distributed over compounds that are similar to each other or dissimilar? Obtaining insights along these lines will likely provide important clues as to whether a given chemical reference space is suitable for a specific application at hand. Moreover, since TRANS-CSNs can be easily annotated with compound property information, such considerations are not limited to SAR analysis but can be extended to structure-property analysis in general. Hence, we expect that chemical space display using TRANS-CSN will find a variety of applications.

Conflict of interests

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties. No writing assistance was utilized in the production of this manuscript.

Acknowledgement

We thank Martin Vogt for helpful discussions and for providing routines for the analysis of CSNs. We also thank OpenEye Scientific Software for a free academic license.

References

- [1] C. M. Dobson. Chemical space and biology. *Nature* **432** (2004), 824–828.
- [2] J.-L. Reymond, R. van Deursen, L. C. Blum, L. Ruddigkeit. Chemical space as a source for new drugs. *MedChemComm* **1** (2010), 30–38.
- [3] R. S. Pearlman, K. M. Smith. “Novel software tools for chemical diversity”. In: *3D QSAR in Drug Design*. Dordrecht: Kluwer Academic Publishers, 1998, pp. 339–353.

- [4] M. Rupp, P. Schneider, G. Schneider. Distance phenomena in high-dimensional chemical descriptor spaces: Consequences for similarity-based approaches. *Journal of Computational Chemistry* **30** (2009), 2285–2296.
- [5] D. K. Agrafiotis, D. N. Rassokhin, V. S. Lobanov. Multidimensional scaling and visualization of large molecular similarity tables. *Journal of Computational Chemistry* **22** (2001), 488–500.
- [6] G. M. Maggiora, J. Bajorath. Chemical space networks: a powerful new paradigm for the description of chemical space. *Journal of Computer-Aided Molecular Design* **28** (2014), 795–802.
- [7] M. Wawer, L. Peltason, N. Weskamp, A. Teckenstrup, J. Bajorath. Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *Journal of Medicinal Chemistry* **51** (2008), 6075–6084.
- [8] N. Tanaka, K. Ohno, T. Niimi, A. Moritomo, K. Mori, M. Orita. Small-world phenomena in chemical library networks: application to fragment-based drug discovery. *Journal of Chemical Information and Modeling* **49** (2009), 2677–2686.
- [9] M. P. Krein, N. Sukumar. Exploration of the topology of chemical spaces with network measures. *The Journal of Physical Chemistry A* **115** (2011), 12905–12918.
- [10] M. Zwierzyzna, M. Vogt, G. M. Maggiora, J. Bajorath. Design and characterization of chemical space networks for different compound data sets. *Journal of Computer-Aided Molecular Design* **29** (2015), 113–125.
- [11] B. Zhang, M. Vogt, G. M. Maggiora, J. Bajorath. Comparison of bioactive chemical space networks generated using substructure- and fingerprint-based measures of molecular similarity. *Journal of Computer-Aided Molecular Design* **29** (2015), 595–608.
- [12] B. Zhang, M. Vogt, G. M. Maggiora, J. Bajorath. Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures. *Journal of Computer-Aided Molecular Design* **29** (2015), 937–950.
- [13] M. Wu, M. Vogt, G. M. Maggiora, J. Bajorath. Design of chemical space networks on the basis of Tversky similarity. *Journal of Computer-Aided Molecular Design* **30** (2016), 1–12.
- [14] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40** (2012), D1100–1107.
- [15] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, D. S. Wishart. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research* **42** (2014), D1091–D1097.
- [16] A. de la Vega de León, S. Kayastha, D. Dimova, T. Schultz, J. Bajorath. Visualization of multi-property landscapes for compound selection and optimization. *Journal of Computer-Aided Molecular Design* **29** (2015), 695–705.
- [17] Chemical Computing Group Inc. *Molecular Operating Environment (MOE)*. Montreal, 2011.
- [18] OpenEye Scientific Software Inc. *OEChem*. Santa Fe, NM, USA, 2012.
- [19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhoffer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12** (2011), 2825–2830.
- [20] T. M. J. Fruchterman, E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience* **21** (1991), 1129–1164.
- [21] J. O’Madadhain, D. Fisher, P. Smyth, S. White, Y.-B. Boey. *Analysis and visualization of network data using JUNG*.
- [22] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103** (2006), 8577–8582.
- [23] M. Newman. *Networks: an introduction*. Oxford University Press, 2010.
- [24] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E* **69** (2004), 066133.

Conclusions

A novel CSN was described that bridges coordinate-based property spaces and coordinate-free network representations. TRANS-CSNs were built using a similarity measure based on the Euclidean distance in high-dimensional property space. Compared to PCA and MDS, TRANS-CSN quickly identified communities of molecules with similar properties. TRANS-CSN visualizations were sensitive to the similarity threshold values used to create the networks. The analysis of networks with edge density lower than 0.05 and high modularity was straightforward. TRANS-CSNs provided novel insights into high-dimensional property spaces compared to dimensionality reduction techniques.

10 Conclusion

Structural relations between bioactive compounds can be analyzed in different ways. MMPs have become a very important tool in chemoinformatics. They are the preferred way to represent activity cliffs and several visualizations use them as a similarity criterion. They also provide comprehensive sets of chemical transformations, where the effect of these transformations on various properties can be studied.

This thesis has added to the chemoinformatics field by first measuring the effect of MMPs on two properties relevant for drug discovery: ionization state and ligand efficiency. For many targets, active molecules shared an IS-class. The chemical transformations found in MMPs did not often change the IS-class; more than 85% of the MMPs were ionization state conservative. Nevertheless, one out of three molecules had MMP partners with different IS-class. Additionally, MMP-cliffs generally encoded chemical changes that increased ligand efficiency between the weakly and the highly potent cliff partner. The ligand efficiency change was on average larger than that found for fingerprint-based activity cliffs.

Additionally, MMP extensions have been developed in this thesis. First, retrosynthetic rules have been used to create second generation MMPs. These new MMPs encode chemical transformations that originate from bonds created by common reactions. This makes it easier for medicinal chemists to apply these transformations for the synthesis of new compounds. Then, a novel methodology to map compounds to existing MMS has been presented. This method has been applied to obtain preliminary SAR information of confirmed hit compounds by mapping them to MMS obtained from bioactive compounds. The SAR information can be used to guide the optimization of hit compounds. Further, predictive SVR models of activity difference between MMP partners were developed. MMP-based

kernel functions enabled the accurate prediction of activity change using SVR models. KRR models that used these kernel functions were also able to achieve high prediction accuracy. This thesis has developed a number of extensions of MMPs that are useful for different tasks such as activity prediction, SAR analysis, or computer-aided compound design. They add to previous research and further emphasize the importance of MMPs in modern chemoinformatics applications.

Representations of chemical space have also been a focus of this thesis. First, principal component plots were combined with MMPA of property changes to guide and visualize compound optimization efforts. These plots allowed the identification of regions of space with favorable property values. Molecules with unfavorable property values were modified with MMP-derived chemical transformations, moving them along the space towards the favorable regions previously identified. Next, star coordinate and parallel coordinate plots have been adapted for medicinal chemistry applications. Star coordinate plots were used to visualize drug-like subspaces generated from a multi-objective optimization search. These subspaces were considered equivalent by the optimization procedure but generated distinct compound distributions. Parallel coordinates were used to study property value distributions in high-dimensional space. Finally, a novel visualization of high-dimensional space, TRANS-CSN, represented a first approximation of network representations to property spaces in chemoinformatics. Compared to dimensionality reduction techniques such as PCA and MDS, it was able to highlight important similarity relationships in high dimensional space. The techniques described in this thesis aid in the analysis of specific subspaces and of compound communities in property space.

Concluding, this thesis has expanded the applicability domain of MMPs with a focus on properties relevant for drug discovery. It has also introduced new visualizations to explore chemical space and aid in compound optimization.

Bibliography

- [1] B. H. Munos. Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug Discovery* **2009**, 8 (12), 959–968.
- [2] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, A. L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* **2010**, 9 (3), 203–214.
- [3] M. Abou-Gharbia, W. E. Childers. Discovery of innovative therapeutics: today's realities and tomorrow's vision. 2. Pharma's challenges and their commitment to innovation. *Journal of Medicinal Chemistry* **2014**, 57 (13), 5525–5553.
- [4] J. Knowles, G. Gromo. A guide to drug discovery. Target selection in drug discovery. *Nature Reviews Drug Discovery* **2003**, 2 (1), 63–69.
- [5] J. Kotz. Phenotypic screening, take two. *Science-Business eXchange* **2012**, 5 (15), doi:10.1038/scibx.2012.385.
- [6] W. P. Walters, M. Namchuk. A guide to drug discovery. Designing screens: how to make your hits a hit. *Nature Reviews Drug Discovery* **2003**, 2 (4), 259–266.
- [7] K. H. Bleicher, H.-J. Böhm, K. Müller, A. I. Alanine. A guide to drug discovery. Hit and lead generation: beyond high-throughput screening. *Nature Reviews Drug Discovery* **2003**, 2 (5), 369–378.
- [8] D. S. Wishart. Improving early drug discovery through ADME modelling. *Drugs in R & D* **2007**, 8 (6), 349–362.
- [9] L. Shargel, S. Wu-Pong, A. Yu. Drug product performance, in vivo: bioavailability and bioequivalence. In: *Applied Biopharmaceutics and Pharmacokinetics*. McGraw-Hill Professional Publishing, **2012**, 403–450.
- [10] P. Greaves, A. Williams, M. Eve. First dose of potential new medicines to humans: how animals help. *Nature Reviews Drug Discovery* **2004**, 3 (3), 226–236.

- [11] C. L. Meinert. *Clinical trials: design, conduct and analysis*. Oxford University Press, **2012**.
- [12] A. Hillisch, N. Heinrich, H. Wild. Computational chemistry in the pharmaceutical industry: from childhood to adolescence. *ChemMedChem* **2015**, 10 (12), 1958–1962.
- [13] F. K. Brown. Chemoinformatics: what is it and how does it impact drug discovery. In: *Annual Reports in Medicinal Chemistry* 33. Ed. by J. A. Bristol. Academic Press, **1998**, 375–384.
- [14] B. Testa, P. Crivori, M. Reist, P.-A. Carrupt. The influence of lipophilicity on the pharmacokinetic behavior of drugs: concepts and examples. *Perspectives in Drug Discovery and Design* **2000**, 19 (1), 179–211.
- [15] D. T. Manallack, R. J. Prankerd, E. Yuriev, T. I. Oprea, D. K. Chalmers. The significance of acid/base properties in drug discovery. *Chemical Society Reviews* **2013**, 42 (2), 485–496.
- [16] L. Peltason, J. Bajorath. Systematic computational analysis of structure–activity relationships: concepts, challenges and recent advances. *Future Medicinal Chemistry* **2009**, 1 (3), 451–466.
- [17] D. Stumpfe, J. Bajorath. Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry* **2012**, 55 (7), 2932–2942.
- [18] Y. Hu, D. Stumpfe, J. Bajorath. Advancing the activity cliff concept [version 1; referees: 3 approved]. *F1000Research* **2013**, 2, 199.
- [19] D. Stumpfe, Y. Hu, D. Dimova, J. Bajorath. Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *Journal of Medicinal Chemistry* **2014**, 57 (1), 18–28.
- [20] N. Furtmann, Y. Hu, M. Gütschow, J. Bajorath. Identification and analysis of the currently available high-confidence three-dimensional activity cliffs. *RSC Advances* **2015**, 5 (54), 43660–43668.
- [21] A. R. Leach, V. J. Gillet. Representation and manipulation of 2D molecular structures. In: *An Introduction To Chemoinformatics*. Springer Netherlands, **2007**, 1–25.
- [22] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, J. Laufer. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences* **1992**, 32 (3), 244–255.

- [23] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, 28 (1), 31–36.
- [24] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi. InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics* **2015**, 7 (1), 23.
- [25] R. Pruim, I. Wegener. Complexity theory: exploring the limits of efficient algorithms. Springer Berlin Heidelberg, **2005**.
- [26] M. A. Johnson, G. M. Maggiora. Concepts and applications of molecular similarity. Wiley, **1990**.
- [27] G. Downs. Molecular descriptors. In: *Computational Medicinal Chemistry for Drug Discovery*. Ed. by W. Langenaeker, H. De Winter, P. Bultinck, J. P. Tollenaere. CRC Press, **2003**, 515–538.
- [28] P. Willett, J. Barnard, G. Downs. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences* **1998**, 38 (6), 983–996.
- [29] Accelrys. *MACCS Structural Keys*. **2011**.
- [30] M. McGregor, S. Muskal. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *Journal of Chemical Information and Computer Sciences* **1999**, 39 (3), 569–574.
- [31] D. Rogers, M. Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **2010**, 50 (5), 742–754.
- [32] D. Horvath, G. Marcou, A. Varnek. Do not hesitate to use Tversky-and other hints for successful active analogue searches with feature count descriptors. *Journal of Chemical Information and Modeling* **2013**, 53 (7), 1543–1562.
- [33] A. M. Wassermann, M. Wawer, J. Bajorath. Activity landscape representations for structure-activity relationship analysis. *Journal of Medicinal Chemistry* **2010**, 53 (23), 8209–8223.
- [34] Y. Hu, D. Stumpfe, J. Bajorath. Lessons learned from molecular scaffold analysis. *Journal of Chemical Information and Modeling* **2011**, 51 (8), 1742–1753.
- [35] N. Brown, E. Jacoby. On scaffolds and hopping in medicinal chemistry. *Mini-Reviews in Medicinal Chemistry* **2006**, 6 (11), 1217–1229.
- [36] G. W. Bemis, M. A. Murcko. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* **1996**, 39 (15), 2887–2893.

- [37] Y. Xu, M. Johnson. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *Journal of Chemical Information and Computer Sciences* **2001**, 41 (1), 181–185.
- [38] A. R. Katritzky, J. S. Kiely, N. Hébert, C. Chassaing. Definition of templates within combinatorial libraries. *Journal of Combinatorial Chemistry* **2000**, 2 (1), 2–5.
- [39] A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch, H. Waldmann. The scaffold tree - visualization of the scaffold universe by hierarchical scaffold classification. *Journal of Chemical Information and Modeling* **2007**, 47 (1), 47–58.
- [40] S. Kayastha, D. Dimova, D. Stumpfe, J. Bajorath. Structural diversity and potency range distribution of scaffolds from compounds active against current pharmaceutical targets. *Future Medicinal Chemistry* **2015**, 7 (2), 111–122.
- [41] P. W. Kenny, J. Sadowski. Structure modification in chemical databases. In: *Chemoinformatics in Drug Discovery*. Ed. by T. I. Oprea. Wiley-VCH, **2005**, 271–285.
- [42] J. Hussain, C. Rea. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *Journal of Chemical Information and Modeling* **2010**, 50 (3), 339–348.
- [43] N. T. Southall, Ajay. Kinase patent space visualization using chemical replacements. *Journal of Medicinal Chemistry* **2006**, 49 (6), 2103–2109.
- [44] J. W. Raymond, I. A. Watson, A. Mahoui. Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *Journal of Chemical Information and Modeling* **2009**, 49 (8), 1952–1962.
- [45] A. G. Leach, H. D. Jones, D. A. Cosgrove, P. W. Kenny, L. Ruston, P. MacFaul, J. M. Wood, N. Colclough, B. Law. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *Journal of Medicinal Chemistry* **2006**, 49 (23), 6672–6682.
- [46] A. M. Birch, P. W. Kenny, I. Simpson, P. R. Whittamore. Matched molecular pair analysis of activity and properties of glycogen phosphorylase inhibitors. *Bioorganic & Medicinal Chemistry Letters* **2009**, 19 (3), 850–853.
- [47] X. Hu, Y. Hu, M. Vogt, D. Stumpfe, J. Bajorath. MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. *Journal of Chemical Information and Modeling* **2012**, 52 (5), 1138–1145.

- [48] T. Geppert, B. Beck. Fuzzy matched pairs: a means to determine the pharmacophore impact on molecular interaction. *Journal of Chemical Information and Modeling* **2014**, 54 (4), 1093–1102.
- [49] J. Weber, J. Achenbach, D. Moser, E. Proschak. VAMMPIRE: a matched molecular pairs database for structure-based drug design and optimization. *Journal of Medicinal Chemistry* **2013**, 56 (12), 5203–5207.
- [50] J. Weber, J. Achenbach, D. Moser, E. Proschak. VAMMPIRE-LORD: a web server for straightforward lead optimization using matched molecular pairs. *Journal of Chemical Information and Modeling* **2015**, 55 (2), 207–213.
- [51] S. L. Posy, B. L. Claus, M. E. Pokross, S. R. Johnson. 3D matched pairs: Integrating ligand- and structure-based knowledge for ligand design and receptor annotation. *Journal of Chemical Information and Modeling* **2013**, 53 (7), 1576–1588.
- [52] M. Wawer, J. Bajorath. Local structural changes, global data views: graphical substructure-activity relationship trailing. *Journal of Medicinal Chemistry* **2011**, 54 (8), 2944–2951.
- [53] A. M. Wassermann, J. Bajorath. A data mining method to facilitate SAR transfer. *Journal of Chemical Information and Modeling* **2011**, 51 (8), 1857–1866.
- [54] N. M. O’Boyle, J. Boström, R. A. Sayle, A. Gill. Using matched molecular series as a predictive tool to optimize biological activity. *Journal of Medicinal Chemistry* **2014**, 57 (6), 2704–2713.
- [55] A. M. Wassermann, P. Haebel, N. Weskamp, J. Bajorath. SAR matrices: automated extraction of information-rich SAR tables from large compound data sets. *Journal of Chemical Information and Modeling* **2012**, 52 (7), 1769–1776.
- [56] D. Gupta-Ostermann, V. Shanmugasundaram, J. Bajorath. Neighborhood-based prediction of novel active compounds from SAR matrices. *Journal of Chemical Information and Modeling* **2014**, 54 (3), 801–809.
- [57] D. Gupta-Ostermann, J. Balfer, J. Bajorath. Hit expansion from screening data based upon conditional probabilities of activity derived from SAR matrices. *Molecular Informatics* **2015**, 34 (2-3), 134–146.
- [58] D. Stumpfe, D. Dimova, J. Bajorath. Composition and topology of activity cliff clusters formed by bioactive compounds. *Journal of Chemical Information and Modeling* **2014**, 54 (2), 451–461.

- [59] D. Dimova, D. Stumpfe, J. Bajorath. Method for the evaluation of structure–activity relationship information associated with coordinated activity cliffs. *Journal of Medicinal Chemistry* **2014**, 57 (15), 6553–6563.
- [60] E. Griffen, A. G. Leach, G. R. Robb, D. J. Warner. Matched molecular pairs as a medicinal chemistry tool. *Journal of Medicinal Chemistry* **2011**, 54 (22), 7739–7750.
- [61] A. M. Wassermann, J. Bajorath. Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. *Future Medicinal Chemistry* **2011**, 3 (4), 425–436.
- [62] A. M. Wassermann, J. Bajorath. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *Journal of Chemical Information and Modeling* **2010**, 50 (7), 1248–1256.
- [63] D. Dimova, D. Stumpfe, J. Bajorath. Specific chemical changes leading to consistent potency increases in structurally diverse active compounds. *MedChemComm* **2014**, 5 (6), 742–749.
- [64] C. Kramer, J. E. Fuchs, S. Whitebread, P. Geddeck, K. R. Liedl. Matched molecular pair analysis: significance and the impact of experimental uncertainty. *Journal of Medicinal Chemistry* **2014**, 57 (9), 3786–3802.
- [65] P. G. Polishchuk, T. I. Madzhidov, A. Varnek. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design* **2013**, 27 (8), 675–679.
- [66] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha. QSAR modeling: where have you been? Where are you going? *Journal of Medicinal Chemistry* **2014**, 57 (12), 4977–5010.
- [67] C. Hansch, P. P. Maloney, T. Fujita, R. M. Muir. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* **1962**, 194 (4824), 178–180.
- [68] T. Fujita, D. A. Winkler. Understanding the roles of the “two QSARs”. *Journal of Chemical Information and Modeling* **2016**, 56 (2), 269–274.
- [69] J. Balfer, J. Bajorath. Visualization and interpretation of support vector machine activity predictions. *Journal of Chemical Information and Modeling* **2015**, 55 (6), 1136–1147.

- [70] C. Cortes, V. Vapnik. Support-vector networks. *Machine Learning* **1995**, 20 (3), 273–297.
- [71] K. Heikamp, X. Hu, A. Yan, J. Bajorath. Prediction of activity cliffs using support vector machines. *Journal of Chemical Information and Modeling* **2012**, 52 (9), 2354–2365.
- [72] A. M. Wassermann, K. Heikamp, J. Bajorath. Potency-directed similarity searching using support vector machines. *Chemical Biology & Drug Design* **2011**, 77 (1), 30–38.
- [73] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, V. Vapnik. Support vector regression machines. In: *Advances in Neural Information Processing Systems 9*. MIT Press, **1996**, 155–161.
- [74] B. E. Boser, I. M. Guyon, V. N. Vapnik. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT '92*. ACM Press, **1992**, 144–152.
- [75] V. Shanmugasundaram, G. M. Maggiora. Characterizing property and activity landscapes using an information-theoretic approach. In: *Proceedings of 222nd American Chemical Society National Meeting, Division of Chemical Information*. American Chemical Society, **2001**, 77.
- [76] J. L. Medina-Franco, A. B. Yongye, J. Pérez-Villanueva, R. A. Houghten, K. Martínez-Mayorga. Multitarget structure-activity relationships characterized by activity-difference maps and consensus similarity measure. *Journal of Chemical Information and Modeling* **2011**, 51 (9), 2427–2439.
- [77] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **1901**, 2 (11), 559–572.
- [78] H. Abdi, L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* **2010**, 2 (4), 433–459.
- [79] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **1964**, 29 (1), 1–27.
- [80] L. Peltason, P. Iyer, J. Bajorath. Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *Journal of Chemical Information and Modeling* **2010**, 50 (6), 1021–1033.

- [81] A. de la Vega de León, J. Bajorath. Design of a three-dimensional multitarget activity landscape. *Journal of Chemical Information and Modeling* **2012**, 52 (11), 2876–2883.
- [82] M. Wawer, L. Peltason, N. Weskamp, A. Teckentrup, J. Bajorath. Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *Journal of Medicinal Chemistry* **2008**, 51 (19), 6075–6084.
- [83] T. M. J. Fruchterman, E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience* **1991**, 21 (11), 1129–1164.
- [84] L. Peltason, Y. Hu, J. Bajorath. From structure-activity to structure-selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem* **2009**, 4 (11), 1864–1873.
- [85] P. Iyer, D. Stumpfe, J. Bajorath. Molecular mechanism-based network-like similarity graphs reveal relationships between different types of receptor ligands and structural changes that determine agonistic, inverse-agonistic, and antagonistic effects. *Journal of Chemical Information and Modeling* **2011**, 51 (6), 1281–1286.
- [86] D. Dimova, M. Wawer, A. M. Wassermann, J. Bajorath. Design of multitarget activity landscapes that capture hierarchical activity cliff distributions. *Journal of Chemical Information and Modeling* **2011**, 51 (2), 258–266.
- [87] D. Gupta-Ostermann, Y. Hu, J. Bajorath. Introducing the LASSO graph for compound data set representation and structure-activity relationship analysis. *Journal of Medicinal Chemistry* **2012**, 55 (11), 5546–5553.
- [88] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **2001**, 46 (1-3), 3–26.
- [89] C. Abad-Zapatero. A Sorcerer's apprentice and The Rule of Five: from rule-of-thumb to commandment and beyond. *Drug Discovery Today* **2007**, 12 (23/24), 995–997.
- [90] J. Clardy, C. Walsh. Lessons from natural molecules. *Nature* **2004**, 432 (7019), 829–837.
- [91] A. L. Hopkins, C. R. Groom, A. Alex. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, 9 (10), 430–431.
- [92] I. D. Kuntz, K. Chen, K. A. Sharp, P. A. Kollman. The maximal affinity of ligands. *Proceedings of the National Academy of Sciences* **1999**, 96 (18), 9997–10002.

- [93] A. L. Hopkins, G. M. Keserü, P. D. Leeson, D. C. Rees, C. H. Reynolds. The role of ligand efficiency metrics in drug discovery. *Nature Reviews Drug Discovery* **2014**, 13 (2), 105–121.
- [94] D. Tanaka, Y. Tsuda, T. Shiyama, T. Nishimura, N. Chiyo, Y. Tominaga, N. Sawada, T. Mimoto, N. Kusunose. A practical use of ligand efficiency indices out of the fragment-based approach: ligand efficiency-guided lead identification of soluble epoxide hydrolase inhibitors. *Journal of Medicinal Chemistry* **2011**, 54 (3), 851–857.
- [95] P. D. Leeson, B. Springthorpe. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery* **2007**, 6 (11), 881–890.
- [96] C. Abad-Zapatero, J. T. Metz. Ligand efficiency indices as guideposts for drug discovery. *Drug Discovery Today* **2005**, 10 (7), 464–469.
- [97] K. Miettinen. Nonlinear multiobjective optimization. Springer US, **1999**.
- [98] V. Pareto. Manuale di economia politica. Società Editrice Libreria, **1906**.
- [99] H. W. Kuhn, A. W. Tucker. Nonlinear programming. In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, **1951**, 481–492.
- [100] C. A. Nicolaou, N. Brown. Multi-objective optimization methods in drug design. *Drug Discovery Today: Technologies* **2013**, 10 (3), e427–e435.
- [101] Y. Collette, P. Siarry. Multiobjective methods using metaheuristics. In: *Multiobjective Optimization: Principles and Case Studies*. Springer Berlin Heidelberg, **2004**, 109–134.
- [102] J. Kennedy, R. Eberhart. Particle swarm optimization. In: *Proceedings of ICNN'95 - International Conference on Neural Networks*. IEEE, **1995**, 1942–1948.
- [103] V. Namasivayam, J. Bajorath. Multiobjective particle swarm optimization: automated identification of structure-activity relationship-informative compounds with favorable physicochemical property distributions. *Journal of Chemical Information and Modeling* **2012**, 52 (11), 2848–2855.
- [104] D. T. Manallack. The pK(a) distribution of drugs: application to drug discovery. *Perspectives in Medicinal Chemistry* **2007**, 1, 25–38.
- [105] D. T. Manallack, R. J. Pranker, G. C. Nassta, O. Ursu, T. I. Oprea, D. K. Chalmers. A chemogenomic analysis of ionization constants-implications for drug discovery. *ChemMedChem* **2013**, 8 (2), 242–255.

- [106] P. S. Charifson, W. P. Walters. Acidic and basic drugs in medicinal chemistry: a perspective. *Journal of Medicinal Chemistry* **2014**, 57 (23), 9701–9717.
- [107] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2012**, 40 (Database issue), D1100–1107.
- [108] H. N. Po, N. M. Senozan. The Henderson-Hasselbalch equation: its history and limitations. *Journal of Chemical Education* **2001**, 78 (11), 1499–1503.
- [109] S. Kayastha, A. de la Vega de León, D. Dimova, J. Bajorath. Target-based analysis of ionization states of bioactive compounds. *MedChemComm* **2015**, 6 (6), 1030–1035.
- [110] OpenEye Scientific Software Inc. *OpenEye*. **2010**.
- [111] K. Godula, D. Sames. C-H bond functionalization in complex organic synthesis. *Science* **2006**, 312 (5770), 67–72.
- [112] M. Lobell, M. Hendrix, B. Hinzen, J. Keldenich, H. Meier, C. Schmeck, R. Schoe-Loop, T. Wunberg, A. Hillisch. In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem* **2006**, 1 (11), 1229–1236.

Additional Publications

V. Shanmugasundaram, L. Zhang, S. Kayastha, A. de la Vega de León, D. Dimova, J. Bajorath. Monitoring the progression of structure-activity relationship information during lead optimization. *Journal of Medicinal Chemistry*, in press.

A. Hameed, K. Khan, S. T. Zehra, R. Ahmed, Z. Shafiq, S. M. Bakht, M. Yaqub, M. Hussain, A. de la Vega de León, N. Furtmann, J. Bajorath, H. A. Shad, M. N. Tahir, J. Iqbal. Synthesis, biological evaluation and molecular docking of N-phenyl thiosemicarbazones as urease inhibitors. *Bioorganic Chemistry* **2015**, 61, 51-57.

D. Stumpfe, A. de la Vega de León, D. Dimova, J. Bajorath. Follow up: Advancing the activity cliff concept, part II [version 1; referees: 2 approved]. *F1000Research* **2014**, 3, 75.

Y. Hu, A. de la Vega de León, B. Zhang, J. Bajorath. Matched molecular pair-based data sets for computer-aided medicinal chemistry [version 2; referees: 4 approved]. *F1000Research* **2014**, 3, 36.