JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

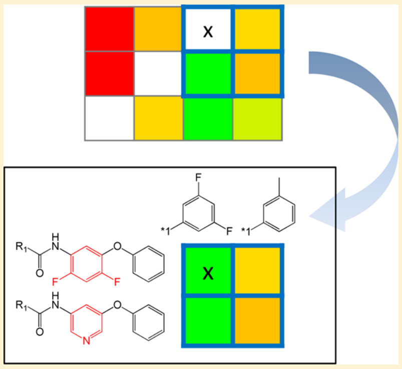
Neighborhood-Based Prediction of Novel Active Compounds from SAR Matrices

Disha Gupta-Ostermann,[†] Veerabahu Shanmugasundaram,[‡] 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, North Rhine-Westphalia, Germany

[‡]Computational Analysis & Design, Center of Chemistry Innovation & Excellence, Worldwide Medicinal Chemistry, Pfizer, Groton, Connecticut 06340, United States

ABSTRACT: The SAR matrix data structure organizes compound data sets according to structurally analogous matching molecular series in a format reminiscent of conventional R-group tables. An intrinsic feature of SAR matrices is that they contain many virtual compounds that represent unexplored combinations of core structures and substituents extracted from compound data sets on the basis of the matched molecular pair formalism. These virtual compounds are candidates for further exploration but are difficult, if not impossible to prioritize on the basis of visual inspection of multiple SAR matrices. Therefore, we introduce herein a compound neighborhood concept as an extension of the SAR matrix data structure that makes it possible to identify preferred virtual compounds for further analysis. On the basis of well-defined compound neighborhoods, the potency of virtual compounds can be predicted by considering individual contributions of core structures and substituents from neighbors. In extensive benchmark studies, virtual compounds have been prioritized in different data sets on the basis of multiple neighborhoods yielding accurate potency predictions.



INTRODUCTION

The conventional approach to the exploration of structure–activity relationships (SARs) in medicinal chemistry is the organization of compound series in R-group tables.[1] Such tables record R-groups at different substitution sites of a core structure common to a compound series and report activity values of analogs. With the aid of R-group tables, new compounds are designed on the basis of medicinal chemistry experience and intuition. Simple comparisons of molecular graphs typically provide a basis for studying similarities and differences between active compounds and for identifying key structural features that correlate with activity.[1,2] In addition to standard R-group tables, other hierarchical structural organization schemes have also been introduced for SAR exploration that are based upon maximum common core structures[3] or molecular scaffolds.[4,5]

Computationally, the design of new compounds on the basis of series information has traditionally been supported by Quantitative SAR (QSAR) analysis methods.[6,7] Following established QSAR principles, mathematical models are derived using descriptors of chemical structure and properties to predict the effects of chemical substitutions on compound potency. In addition, in recent years, graphical methods have been increasingly employed to visually analyze SARs, often going beyond individual compound series.[8,9] However, SAR visualization methods are generally descriptive in nature and do not directly lead to compound predictions.

In addition to scaffold-based approaches and graphical methods, the matched molecular pair (MMP) represents another emerging concept in medicinal chemistry[10,11] that is highly relevant for SAR analysis. An MMP is generally defined as a pair of compounds that differ only by a structural change at a single site, i.e., the exchange of a pair of substructures.[10] Applying the MMP formalism, a chemical change relating two compounds to each other can be directly associated with changes in different types of molecular properties including biological activity.[11,12] Even for large compound data sets, MMPs can be efficiently generated algorithmically,[13] hence providing a comprehensive pairwise substructure-based organization scheme. For SAR analysis, the MMP concept has been further extended by introducing matching molecular series (MMS),[14] which are defined as a series of compounds that only differ by chemical changes at a single site. MMS can be utilized to construct SAR networks and follow substructure changes within and across compound series that lead to SAR progression.[14]

As a data structure that combines intuitive SAR visualization with the MMS concept and further extends this concept, the SAR Matrix (SARM) has been recently introduced.[15] SARM provides an R-group table-like high-resolution view of compound data sets that are automatically organized into structurally related series, as illustrated in Figure 1. It consists of

Received: January 23, 2014
Published: March 4, 2014

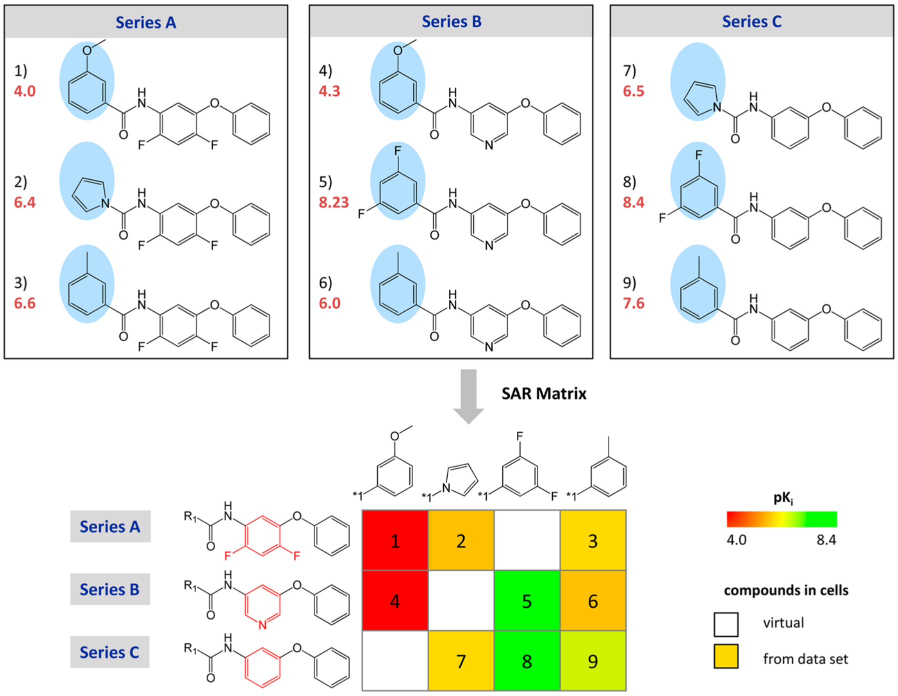


Figure 1. SAR matrix. Three model compound series (A, B, and C) containing three compounds each are shown with their respective pK_i values (red). Compounds in a series share a common core structure and differ by substitutions at a single site (highlighted in blue). The three series contain structurally related cores (bottom left; substructure differences between cores are highlighted in red). The SAR matrix is generated by combining structurally related analog series. Rows and columns represent compounds that share the same core and substituent, respectively. In each cell, the combination of a core and a substituent defines a unique compound. Compounds present in the data set are indicated by filled cells that are color-coded according to potency using a continuous spectrum from red (low potency) over yellow (intermediate) to green (high). In addition, empty cells indicate virtual compounds.

"real" data set and structurally analogous virtual compounds. Herein, we introduce an approach to predict virtual compounds and their actual potency values from SARMs on the basis of compound neighborhood information.

In the following sections, we describe the SARM data structure, introduce the prediction methodology, and report the results of systematic potency predictions of virtual compounds from SARMs of different data sets.

■ SAR MATRIX DESIGN PRINCIPLES

SARM generation involves a dual compound fragmentation scheme resulting in two-level MMP generation.[15] In the first step, compounds are subjected to MMP fragmentation applying the algorithm by Hussain and Rea,[13] leading to the generation of an index table with large key fragments (core structures) and smaller value fragments (substituents). In the second step, the cores in the index table are subjected to an additional round of fragmentation. The resulting fragments are stored in a new index table with the larger fragment as the key, analogously to the first step. Hence, this dual fragmentation scheme identifies compound series having structurally related cores termed "structurally analogous MMS" (A_MMS),[15] which further extends the MMS concept. In Figure 1, three series A, B, and C are shown, each of which contains a common core structure and differs at a single site (highlighted in blue). The second fragmentation step reveals that the core structures of these three series are related and also only differ at a single site (red substructures at the bottom). Hence, series A, B, and C form an A_MMS, which is represented in a unique SARM. As illustrated in Figure 1, the SARM is filled with structurally related cores resulting from the second MMP generation step and corresponding substituents resulting from the first step. Each row in the matrix contains an individual compound series, and each cell represents an individual compound (a unique combination of a key and value fragment). Hence, by design, SARMs are 2D matrices. The series comprising a SARM typically have overlapping yet distinct sets of substituents, giving rise to combinations of real (filled cells) and virtual compounds (empty cells). As shown in Figure 1, a continuous color spectrum is applied to capture the potency information of real compounds. Alternatively, ligand efficiency values can also be used.

Typically, a large compound data set yields multiple or many SARMs. Depending on the algorithmic fragmentation scheme,[13] single-cut matrices (i.e., one exocyclic bond in a compound is systematically deleted to yield key and value fragments), dual-cut (two exocyclic bonds are simultaneously deleted), and triple-cut matrices (three exocyclic bonds are deleted) are separately generated.[15] The resulting SARMs provide a high-resolution organization of a compound data set that accounts for all possible structural relationships between compound series. In addition, virtual compounds contained in a SARM represent as of yet unexplored key-value combinations and hence provide immediate suggestions for new analogs. As such, virtual compounds systematically captured by SARMs can be rationalized as a "chemical space envelope" that delineates regions surrounding a given data set in chemical space. The appearance of SARMs is akin to R-group tables, which makes them easily accessible to medicinal chemists and permits simultaneous exploration of structurally related compound series. In addition to SAR analysis, the SARM data structure has also been adapted to navigate multitarget activity space and study promiscuous compounds.[16]

The set of SARMs representing a larger data set typically contains many virtual compounds (as further detailed below). This raises the question which of these virtual compounds might be prioritized as design suggestions? SARM in its original conceptualization did not provide selection schemes for virtual compounds. Therefore, we introduce a compound and potency prediction approach to further improve the utility of SARMs for compound prioritization and design, as described in the following.

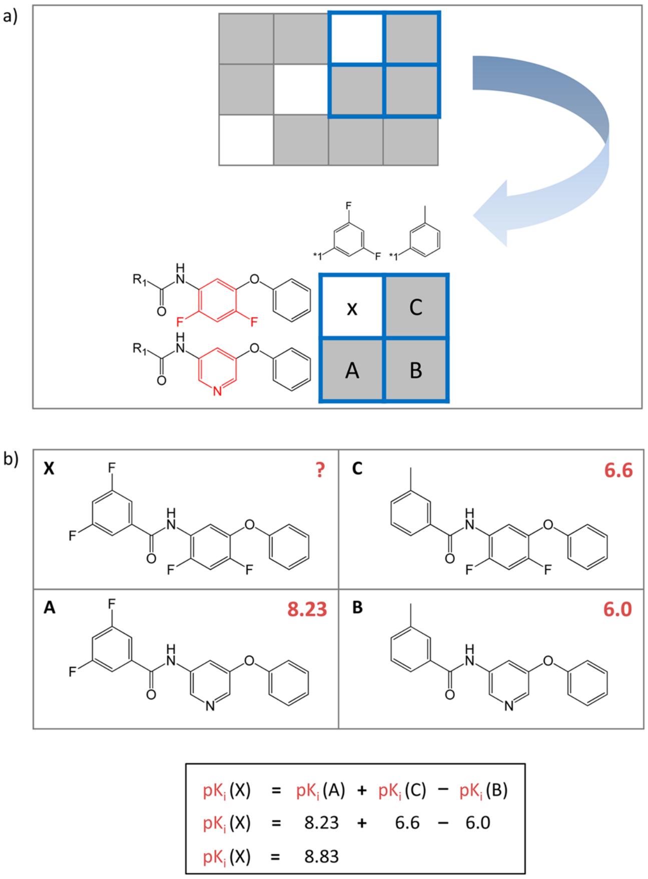

a)
X
C
A
B
R1
F
*1
b)
X ?
C 6.6
A 8.23
B 6.0
pKi (X) = pKi (A) + pKi (C) − pKi (B)
pKi (X) = 8.23 + 6.6 − 6.0
pKi (X) = 8.83

Figure 2. Compound neighborhood and potency prediction. (**a**) The NBH of virtual compound X is marked in blue in the given matrix and shown in more detail at the bottom. Cores and substituents comprising the three compounds forming the NBH are depicted. Compounds A and C share the same substituent and core with X, respectively, and the third neighbor B combines the core and R-group of neighboring compounds A and C, respectively. (**b**) Compound structures and pK_i values (red) of "real" neighbors are given, and calculations are reported to predict the potency of virtual compound X.

■ COMPOUND PREDICTION METHOD

A possible criterion for the selection of virtual compounds from SARMs is their proximity to real data set compounds for which target-specific activity information is available. The underlying idea is that close proximity of a virtual compound to multiple active compounds (i.e., the presence of close structural relationships) increases the probability that this virtual compound might also be active relative to other virtual compounds for which no active neighbors are available. This leads to the notion and assessment of defined compound neighborhoods (NBHs).

Neighborhood Definition. A preferred NBH of a virtual compound is defined as one that consists of three adjacent real compounds, as illustrated in Figure 2a. Because all real compound entries in a given column or row containing a virtual compound are equivalent as potential neighbors, many different three-compound NBHs might exist for a given virtual compound. In general, an NBH of a virtual compound is formed by any combination of three compounds present in the data set of which one shares the core of the virtual compound, the second its substituent, and the third the different core and substituent of these two neighbors. An NBH of this composition is depicted in Figure 2a. If one of these three neighboring compounds is missing, the NBH is incomplete and does not qualify for further analysis. NBHs within the same SARM consist of structurally closely related data set compounds, whereas NBHs from different SARMs might consist of structurally dissimilar compounds.

A virtual compound might already be prioritized based on the number of qualifying NBHs, which establish close structural relationships to known active compounds. However, given the composition of so-defined NBHs, one can go a step further and attempt to predict the potency of a given virtual compound on the basis of potency information of its neighbors.

Neighborhood-Based Potency Prediction. The presence of a three-compound NBH of a virtual compound makes it possible to predict its potency based on a local model utilizing the additivity assumption underlying Free-Wilson analysis,[17,18] as illustrated in Figure 2b. Following this approach, the potency of virtual compound X can be predicted from the sum of logarithmic potencies of compounds A and C, which share the same substituent and core with compound X, respectively, minus the logarithmic potency of compound B, which represents the combination of the core structure and substituent of compounds A and C, respectively:

$$pK_i(X) = pK_i(A) + pK_i(C) - pK_i(B)$$

Hence, subtracting the potency of B from the sum of potencies of A and B corrects for the contributions of the structural fragments of A and C that are not contained in X. Hence, in essence, the local models represent "mini-QSAR" models. These Free-Wilson-type predictions are applicable to any qualifying compound NBH. A virtual compound often (but not always) occurs in multiple SARMs obtained for a given data set. Therefore, all qualifying unique NBHs of each virtual compound are systematically identified and subjected to potency predictions, as illustrated in Figure 3. The number of qualifying NBHs might be increased in subsequent iterations by considering compounds with predicted potency for NBH definition (e.g., by generating NBHs consisting of data set and previously predicted compounds). However, prediction accuracy would likely be reduced in such cases.

Comparison with QSAR. The NBH-based prediction concept introduced herein represents a local prediction approach. Local Free-Wilson-type predictions are carried out for all matrices with qualifying NBHs for given virtual compounds. The method does not utilize chemical descriptors such as Fujita-Hansch QSAR or models based on training sets such as Fujita-Hansch and Free-Wilson QSAR. The NBH-based prediction scheme exploits characteristic features of SARMs. By design, all compounds contained in an individual SARM are structurally closely related and hence qualify for QSAR-like predictions of virtual compounds within the same

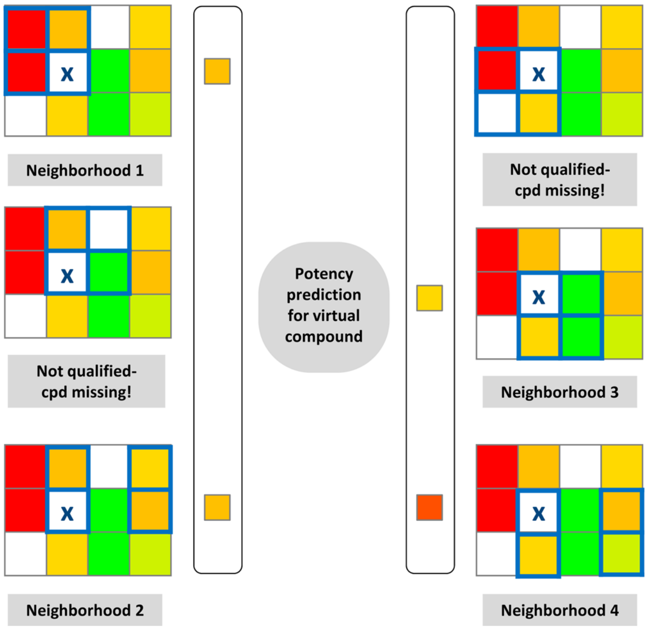


Figure 3. Neighborhood mining. For virtual compound X, the set of all NBHs (outlined in blue) in an SAR matrix is identified and neighborhoods qualifying for potency predictions are determined. In this example, four qualifying NBHs (1 to 4) are identified, and predicted activities are indicated by squares color-coded according to the spectrum in Figure 1.

matrix. In addition, NBH-based predictions do not depend on the population density of data set compounds in a given SARM, only on the presence of local NBHs. This represents an important difference compared to standard QSAR modeling. To predict virtual matrix compounds, QSAR models would need to be built for individual SARMs and would require the presence of sufficient numbers of data set compounds to assemble meaningful training sets.

Applicability Domain. The prediction methodology generally aims to prioritize virtual compounds as candidates for synthesis. In the context of our analysis, preferred virtual compounds would in principle be characterized by the presence of many different NBHs that yield consistent potency predictions. Most interesting would be compounds with consistently predicted high potency. The consistency of potency predictions can be assessed by calculating the standard deviation (SD) of multiple independent predictions; low SD values indicate consistent predictions. It should be noted that low SDs do not necessarily correlate with high prediction accuracy (i.e., predictions can be consistently incorrect). By contrast, high SDs are indicative of inconsistent predictions. Such predictions are likely to occur if virtual compounds have many structurally analogous neighbors with large differences in potency. In this case, virtual compounds map to regions of SAR discontinuity[8] including activity cliffs,[8,19] and local QSAR-type predictions are no longer applicable in a meaningful way.[19,20] Discontinuous compound NBHs yielding predictions with high SDs might still be attractive for the analysis of specific compound environments in SARMs, as further discussed below. However, they fall outside the applicability domain of QSAR modeling and are hence deprioritized in our systematic potency predictions (even if predictions from individual discontinuous NBHs would indicate high compound potency).

IMPLEMENTATION, DATA SETS, AND CALCULATION SETUP

Routines to generate SARMs were implemented with the aid of the OpenEye chemistry toolkit,[21] and potency prediction routines were implemented in Java. Statistical analyses were carried out with R.[22]

For benchmark calculations, six large sets of different G protein coupled receptor antagonists were extracted from ChEMBL (release 15)[23] for which K_i values were available as potency measurements. The targets, sizes, and potency ranges of these data sets are reported in Table 1. These data sets consisted of 1103−1850 compounds that formed SARMs. For all data sets, all possible SARMs were calculated, producing between 655 and 1109 matrices per set, as also reported in Table 1.

Table 1. Data Set Statistics[a]

target name	TID	# SAR matrix	# cpds	pK_i range
dopamine D2 receptor	217	700	1419	3.0 to 10.2
adenosine A1 receptor	226	1104	1825	4.2 to 10.5
adenosine A2a receptor	251	957	1850	4.0 to 11.0
adenosine A3 receptor	256	1109	1547	4.1 to 11.0
melanocortin receptor 4	259	669	1103	3.9 to 9.4
histamine H3 receptor	264	655	1718	4.4 to 10.5

[a] For each compound data set, the ChEMBL target ID (TID), the number (#) of SAR matrices and compounds, and their pK_i range are reported.

Following SARM generation, the following prediction protocol was established:

(i) One third of the compounds were randomly removed from each data set, and all corresponding matrix positions were converted into virtual cells.

(ii) All removed ("pseudovirtual") compounds were recorded as potential targets for predictions (original virtual compounds contained in unmodified SARMs could not be predicted in benchmarking).

(iii) For all "virtualized" cells from potential prediction targets, qualifying compound NBHs were systematically determined.

(iv) Compounds having at least three unique NBHs across SARMs were selected, and potency predictions were carried out on the basis of each NBH. Then, the consistency of predictions was assessed.

RESULTS OF SYSTEMATIC PREDICTIONS

Following the protocol reported above, pseudovirtual compounds having qualifying NBHs were identified across SARMs generated for different data sets and systematic potency predictions were carried out.

Prioritization of Pseudovirtual Compounds. Between 76 and 179 pseudovirtual compounds with at least three qualifying NBHs were identified in SARMs obtained for the different data sets, as reported in Table 2. For all of these compounds, potency predictions were carried out for individual NBHs, and SDs of the predictions were calculated. SD values falling into the first quartile of the distributions of all SDs within a data set were classified as low SDs, and the corresponding pseudovirtual compounds were designated low SD (L_SD) compounds. Figure 4 reports the distributions of SDs of predictions for all L_SD compounds. Median SD values were close to 0.5 pK_i units for all data sets. Hence, alternative

Table 2. Pseudovirtual Compounds and Potency Predictions[a]

TID	# pseudovirtual cpds	# prioritized cpds	# L Sd cpds
217	473	91	23
226	608	179	45
251	616	126	32
256	515	146	37
259	367	76	19
264	572	134	34

[a] For each data set (indicated by TID according to Table 1), the total number (#) of pseudovirtual compounds (i.e., data set compounds removed from matrices), number of prioritized virtual compounds with at least three qualifying compound neighborhoods, and number of prioritized virtual compounds yielding potency predictions with low standard deviations (L_SD) are reported.

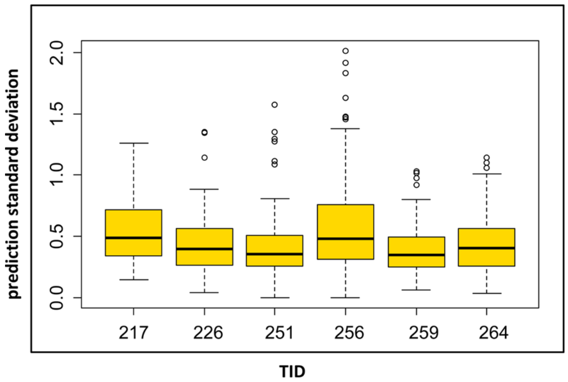


Figure 4. Standard deviations of predictions. Box plots report the distributions of standard deviations (y-axis) of potency predictions for single pseudovirtual compounds having at least three qualifying NBHs. Data sets are indicated by TIDs according to Table 1 (x-axis).

potency predictions for these compounds generally fell well within 1 order of magnitude. Between 19 and 45 L_SD compounds were obtained from ~400 to ~600 potential candidates for the different data sets, as also reported in Table 2. By increasing the SD threshold beyond the first quartile of the global distribution, additional compounds can be obtained. However, for our proof-of-principle investigation, the number of L_SD compounds in Table 2 was readily sufficient. On the basis of our above considerations, L_SD compounds fell within the applicability domain of local prediction models, and their potency predictions were further analyzed.

Prediction Performance. In Figure 5, distributions of prediction errors are reported for L_SD compounds from all data sets. With very few exceptions, prediction errors associated with potencies averaged over NBHs were well within 1 order of magnitude with a median of only ~0.25 order of magnitude. This is further illustrated in Figure 6 by a comparison of observed vs predicted potency values across the different potency ranges captured by the data sets. Regardless of the potency ranges, observed and predicted potency values were consistently very similar. Hence, for prioritized compounds, accurate potency predictions were obtained.

As control calculations, we also generated Free-Wilson QSAR models using R[22] for individual SARMs to predict L_SD compounds they contained (as discussed above). In these cases, data set compounds contained in a SARM were used as a training set for deriving a matrix-based Free-Wilson model. We then compared prediction errors of these Free-Wilson models for L_SD compounds with NBH-based predictions originating from the same SARM. In test calculations on individual SARMs, we observed very similar (low) prediction errors for matrix-based Free-Wilson and NBH-based predictions, indicating that NBH information was sufficient to achieve accurate predictions for L_SD compounds.

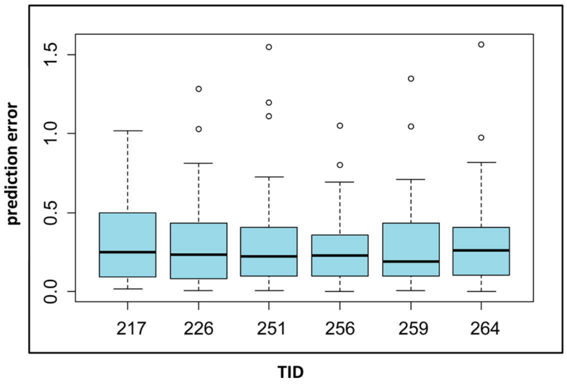


Figure 5. Prediction errors. Box plots report the distributions of prediction errors for individual pseudovirtual compounds yielding potency predictions with low standard deviations (L_SD compounds). Prediction errors were calculated on the basis of averaged predicted pK_i values and are reported on the y-axis as Δ pK_i units relative to observed compound potencies. On the x-axis, data sets are given.

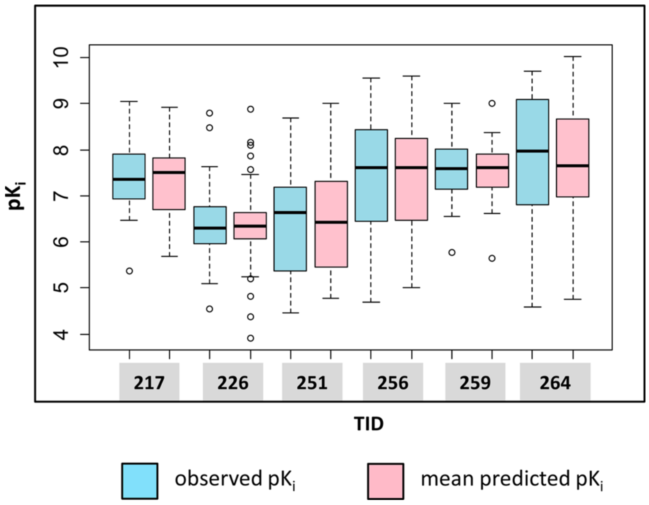


Figure 6. Observed vs predicted potency values. Box plots report the distribution of observed (blue) and mean predicted pK_i values (pink) for L_SD compounds for each data set.

Neighborhood Frequency. L_SD compounds can also be ranked according to the number of NBHs qualifying for prediction. Compounds frequently predicted to have high activity are generally the most interesting candidates for further exploration. In Figure 7, potency predictions and the frequency of NBHs are compared for individual L_SD compounds. Compound data points are scaled in size according to prediction errors. Depending on the data sets, individual compounds were found to have a maximum of ~10 to ~30 NBHs across SARMs. If prediction errors exceeding 1 order of magnitude were detected, they were exclusively observed for virtual compounds having only three or four NBHs. With further increasing numbers of NBHs, predictions became increasingly accurate (and prediction errors typically very small). These findings reflect a general relationship between

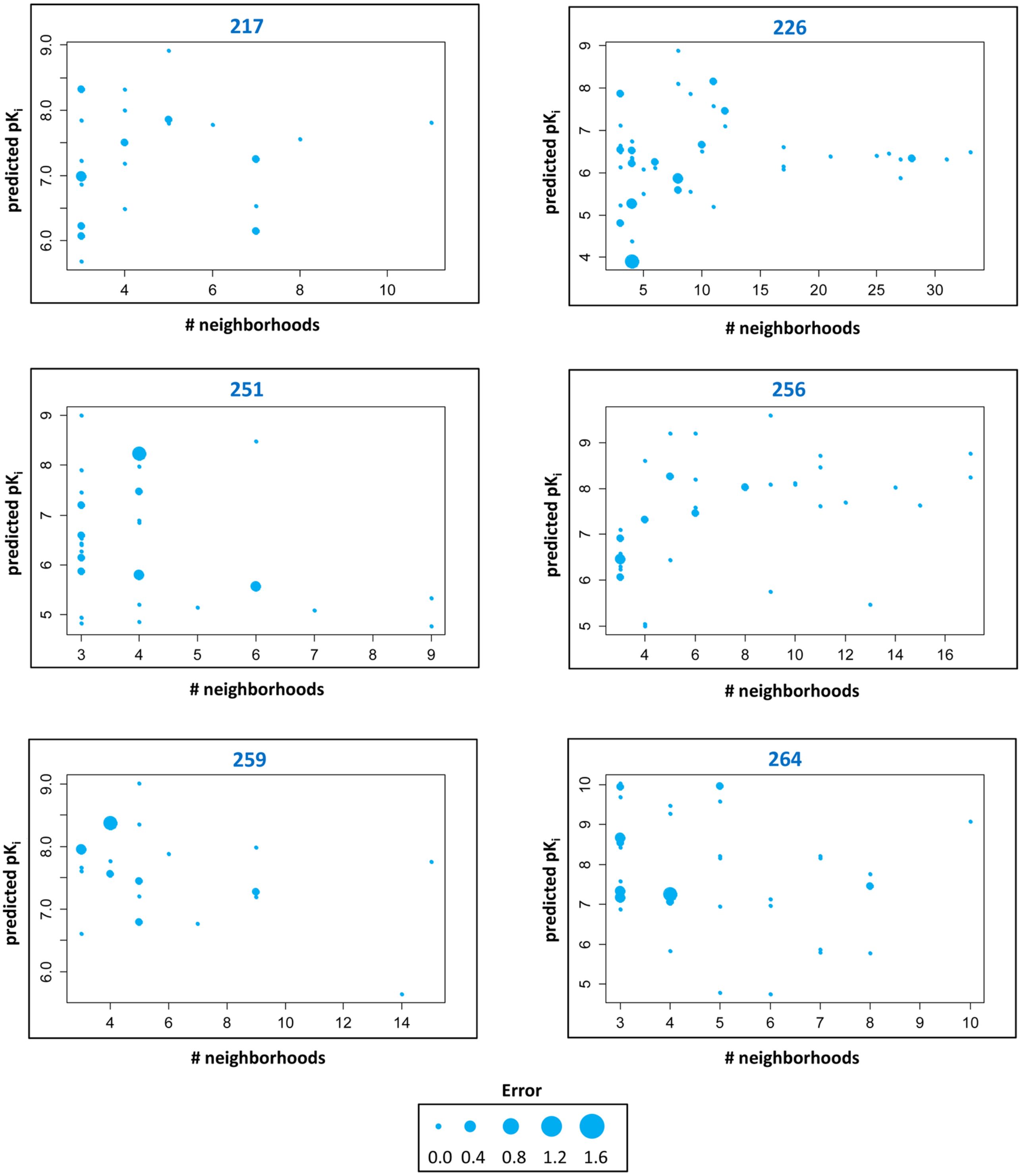


Figure 7. Neighborhood counts vs potency predictions. For L_SD compounds from each data set (TID at the top), the number of NBHs qualifying for predictions and the averaged predicted pK_i value is reported. Compounds are represented as dots that are scaled in size according to prediction errors (Δ pK_i units), as indicated at the bottom.

increasing numbers of qualifying NBHs and prediction accuracy.

Exemplary Compounds and Predictions. In Figure 8, potency predictions are reported for exemplary L_SD compounds together with their NBH information. In Figure 8a, predictions are reported for a histamine receptor antagonist for which 10 different NBHs were available. Despite different potency distributions of data set compounds across the NBHs, the nanomolar potency (pK_i 9.0) of this pseudovirtual compound was accurately predicted. Similarly, in Figure 8b, an adenosine A3 receptor antagonist for which six qualifying NBHs were available yielded an accurate potency prediction. By contrast, in Figure 8c, another adenosine A3 receptor antagonist is shown with only four available NBHs. Although a very low SD value (0.15 pK_i units) was also observed for individual predictions in this case, the potency of the

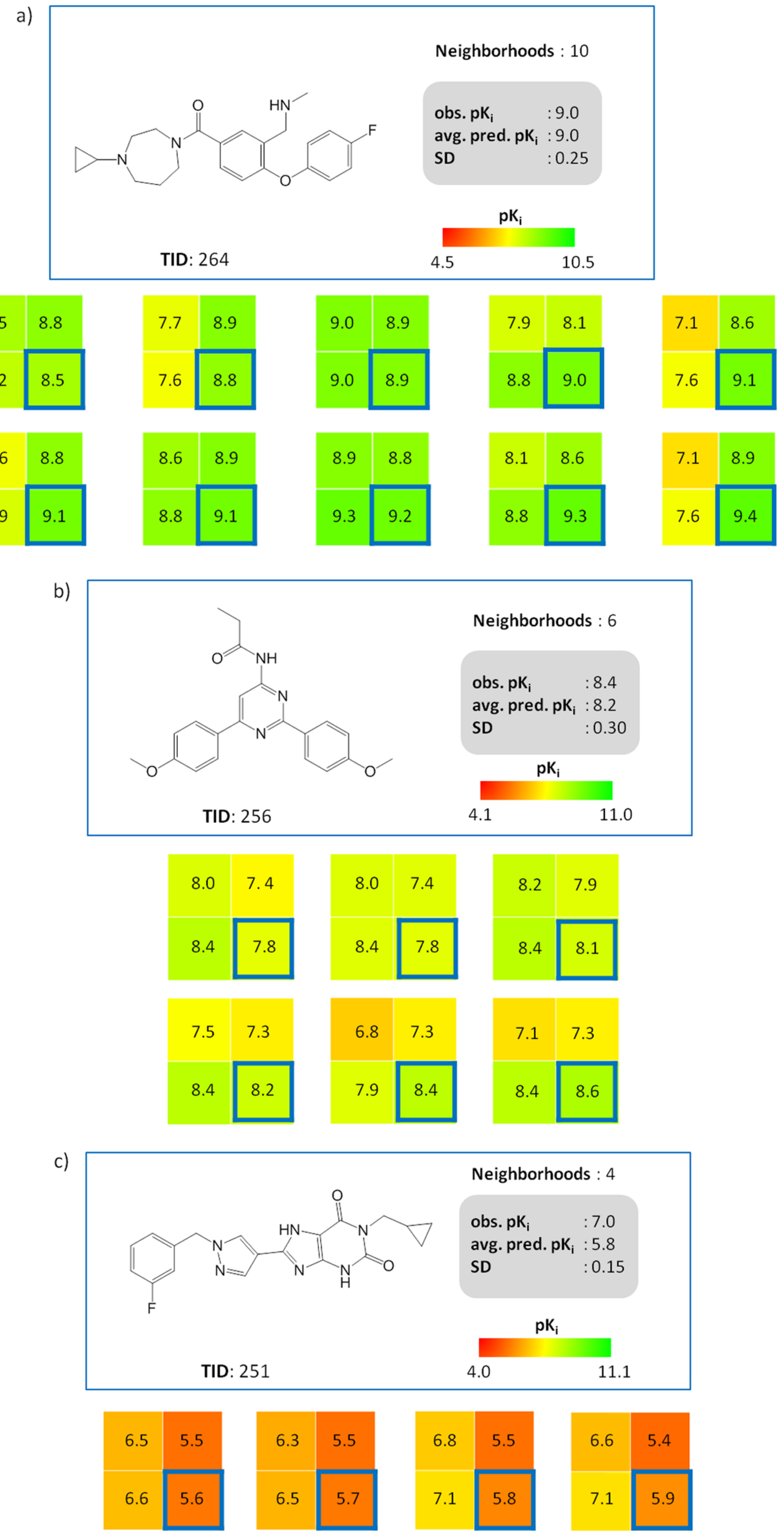


Figure 8. Exemplary predictions. In (**a**) and (**b**), the structures and NBHs of a histamine and adenosine A3 antagonist are shown, respectively, for which accurate potency prediction were obtained. In (**c**), adenosine A2 antagonist is shown for which a prediction error of more than 1 order of magnitude was observed. All three examples are L_SD compounds. Cells in NBHs are annotated with pK_i values of corresponding compounds. Cells framed in blue represent pseudovirtual compounds with predicted potency values.

compound was underpredicted in all four NBHs, yielding a final prediction error of 1.2 orders of magnitude, one of the largest errors observed; a rare case in our predictions of prioritized compounds. For increasing numbers of NBHs qualifying for prediction, only small errors were observed, as reported above.

For comparison, Figure 9 shows a compound yielding high SD values over multiple predictions (hence falling outside the

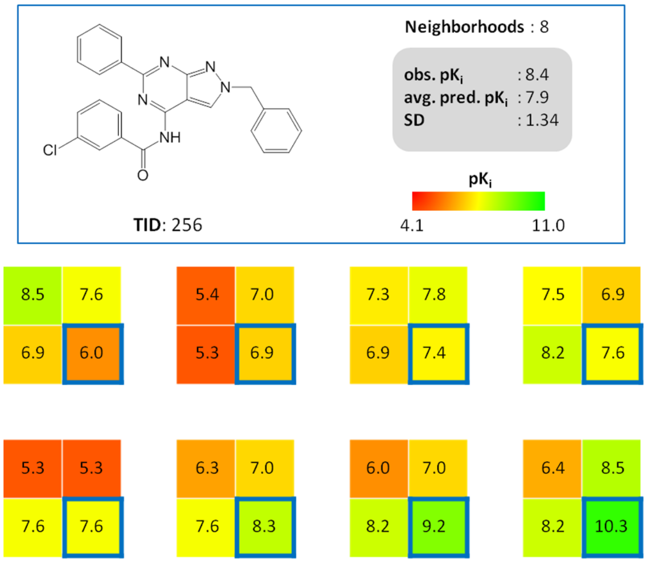


Figure 9. Neighborhoods in discontinuous SAR regions. Shown is an adenosine A3 antagonist for which NBHs map to discontinuous SAR regions and thus yield a wide range of potency predictions (falling outside the applicability domain of the NBH-based prediction concept). The representation is similar to Figure 8.

applicability domain of our approach). These predictions are characterized by the presence of NBHs falling into discontinuous local SAR regions. Although potency values cannot be accurately predicted for such NBHs, virtual compounds mapping to such regions might still be attractive prediction targets because large potency fluctuations are expected for such compounds (and one might hope to hit a potency "home run"). Therefore, high SD values can also be used as a diagnostic to systematically identify virtual compounds with NBHs in discontinuous SAR regions.

DISCUSSION AND CONCLUSIONS

The SARM data structure was originally designed to organize compound sets on the basis of core structures and substituents and all possible structural relationships between cores. This was accomplished through the systematic generation of A_MMS and their organization in a matrix format. A characteristic feature of SARMs is that they contain large numbers of virtual compounds that represent as of yet unexplored core structure and substituent combinations and hence offer suggestions for compound design. However, the SARM data structure does not enable a direct prioritization of such virtual compounds. Rather, visual analysis of SARMs is required to study virtual compounds. Therefore, we have developed a methodology to predict novel active compounds from SARMs. The central idea underlying this approach is the compound neighborhood concept. For each virtual compound in SARMs, NBHs exclusively consisting of known active compounds are systematically assessed. Virtual compounds can be ranked according to numbers of such NBHs. Prioritization on this basis is akin to a "guilt by association" approach, which assumes that the likelihood of a virtual compound to be active increases with the number of neighboring structural analogs. However, as shown herein, one can go a step further and utilize the NBH concept for potency predictions. These predictions are facilitated by applying a Free-Wilson-like additivity principle to individual neighborhoods. This leads to the prediction of the potency of a virtual compound on the basis of differential core and substituent contributions from active neighbors. Of course, the approach is distinct from classical Free-Wilson analysis that derives a mathematical model for the activity of a series of analogs by additively accounting for contributions from all R-groups. However, adapting the additivity principle essentially limits NBH-based potency predictions to the applicability domain of QSAR approaches, thus requiring the presence of SAR continuity and the absence of activity cliffs and cooperative SAR effects. A distinguishing feature of our NBH-based prediction approach is that predictions over multiple NBHs are prioritized. Then, one can assign confidence to consistent predictions resulting in low SD values. As demonstrated herein, accurate potency predictions were obtained in such cases across different data sets, with prediction accuracy further increasing with the number of qualifying NBHs. Depending on the composition of NBHs, virtual compounds with higher potency than known active neighbors can be predicted, and these predictions can also be easily prioritized. Moreover, potency predictions over multiple NBHs can be used as a diagnostic for local SAR environments. For example, predictions yielding high SD values are indicative of discontinuous SAR regions surrounding virtual compounds in which structurally analogous neighbors might have very different potencies. Although these regions usually fall outside the applicability domain of potency predictions employing an additivity principle, they are nonetheless interesting for compound design. This is the case because one might hope to hit a potency "home run" in discontinuous regions that are rich in NBHs, which are easily identified using the approach introduced herein.

In conclusion, neighborhood-based matrix analysis and potency predictions enable the prioritization of virtual compound from SARMs and are thus anticipated to further increase the attractiveness and utility of the SARM data structure for medicinal chemistry applications.

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.

ACKNOWLEDGMENTS

The authors would like to thank Dilyana Dimova for help with data sets and Liying Zhang for helpful discussions. D.G.-O. is supported by Boehringer Ingelheim.

REFERENCES

(1) *The Practice of Medicinal Chemistry*, 3rd ed.; Wermuth, C. G., Ed.; Academic Press-Elsevier: Burlington, San Diego, USA, London, UK, 2008.

(2) Kubinyi, H. Similarity and dissimilarity. A medicinal chemist's view. *Perspect. Drug Discovery Des.* **1998**, *9–11*, 225–252.

(3) Cho, S. J.; Sun, Y. Visual exploration of structure-activity relationship using maximum common framework. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 571–578.

(4) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The scaffold tree — visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

(5) Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Introducing the LASSO graph for compound data set representation and structure-activity relationship analysis. *J. Med. Chem.* **2012**, *55*, 5546−5553.

(6) Martin, Y. C. A practitioner's perspective of the role of quantitative structure-activity analysis in medicinal chemistry. *J. Med. Chem.* **1981**, *24*, 229−237.

(7) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure-activity relationship paradigm. *Methods Mol. Biol.* **2004**, *275*, 131−214.

(8) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

(9) Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discovery Today* **2010**, *15*, 631−639.

(10) Kenny, P. W.; Sadowski, J. Structure modification in chemical databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271−285.

(11) 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.

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

(13) 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.

(14) Wawer, M.; Bajorath, M. Local structural changes, global data views: graphical substructure-activity relationship trailing. *J. Med. Chem.* **2011**, *54*, 2944−2951.

(15) Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR matrices: automated extraction of information-rich SAR tables from large compound data sets. *J. Chem. Inf. Model.* **2012**, *52*, 1769−1776.

(16) Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Systematic mining of analog series with related core structures in multi-target activity space. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 665−674.

(17) Free, S. M.; Wilson, J. W. A mathematical contribution to structure-activity studies. *J. Med. Chem.* **1964**, *7*, 395−399.

(18) Kubinyi, H. Free Wilson analysis. Theory, applications and its relationships to Hansch analysis. *Quant. Struct.-Act. Relat.* **1988**, *7*, 121−133.

(19) Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.

(20) Maggiora, G. M. On outliers and activity cliffs − why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

(21) OEChem, version 1.7.7; OpenEye Scientific Software, Inc.: Santa Fe, NM, USA, 2012.

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

(23) 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.