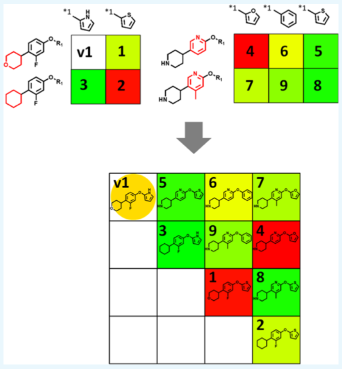
ACS OMEGA
Article
Cite This: ACS Omega 2019, 4, 7061–7069
http://pubs.acs.org/journal/acsodf

Integrating the Structure–Activity Relationship Matrix Method with Molecular Grid Maps and Activity Landscape Models for Medicinal Chemistry Applications

Atsushi Yoshimori,[†] Toru Tanoue,[‡] and Jürgen Bajorath*[,§]

[†]Institute for Theoretical Medicine, Incorporation, 26-1 Muraoka-Higashi 2-chome, Fujisawa, Kanagawa 251-0012, Japan
[‡]Infogram, Incorporation, 2-17-19 Yasuda Building No. 5 3F, Hakataekimae, Hakata-ku, Fukuoka-city, Fukuoka 812-0011, Japan
[§]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, Bonn D-53115, Germany



ABSTRACT: The structure–activity relationship (SAR) matrix (SARM) methodology was originally developed to systematically extract structurally related compound series from data sets of any composition, visualize SAR patterns, and generate virtual candidate compounds. The approach is based upon a dual fragmentation variant of the matched molecular pair formalism. Compound data sets typically yield multiple SARMs that contain unique subsets of structural analogs and virtual candidates complementing existing series. SARM-specific activity predictions make it possible to prioritize virtual analogs for synthesis. The SARM design is intuitive and reminiscent of conventional R-group tables, although the underlying data structure is more complex. Navigating multiple SARMs in parallel can be challenging, depending on the data sets under investigation. Therefore, in this work, we further extend the SARM approach through integration of matrices with newly designed molecular grid maps and activity landscape representations, which provide complementary views of compound relationships and SARs. Moreover, a grid map provides a global view of SARM information including existing compounds, virtual candidates, and associated properties. Grid maps preserve the origin of compounds such that corresponding SARMs can be concomitantly analyzed. In their current implementation, second-generation SARMs make it possible to comprehensively organize and explore large data sets, visualize SARs, and select candidate compounds for practical applications.

1. INTRODUCTION

The analysis of structure–activity relationships (SARs) is of central importance in medicinal chemistry to guide compound optimization.[1] A variety of computational methods are available to aid in SAR analysis and compound design. Classical quantitative SAR (QSAR) methods[2] are used to build linear SAR models for compound series and prioritize analogs. In addition, machine learning algorithms are applied to model nonlinear SARs and predict novel active compounds.[2,3] Furthermore, numerical SAR analysis methods have been introduced[4,5] to quantitatively describe SAR features in compound data sets. These functions can be used to globally characterize SARs or quantitatively describe local SARs for compound subsets. Numerical multi-objective optimization techniques were also adapted to utilize SAR information for compound optimization and design.[6,7] Another category of computational methods applies the scaffold concept to generate molecular hierarchies for SAR exploration and compound selection.[8,9] Moreover, different statistical and modeling approaches were introduced to monitor SAR progression of evolving compound series.[10−13] Going beyond individual analog series, computational methods and data structures have been used to extract SAR information from heterogeneous compound data sets.[14] Among these, approaches for graphical SAR analysis and SAR visualization play a particularly important role.[15] SAR visualization methods range from extensions of conventional R-group tables[1,16] and scaffold-based techniques[17−20] to molecular networks[20,21] and different types of activity landscape and activity cliff representations.[22,23]

The SAR matrix (SARM) method was designed to bridge between SAR visualization and compound design. It was originally introduced for the extraction and organization of related series of active compounds from data sets, elucidation of SAR patterns, and generation of virtual candidate compounds to further expand existing series.[24] Among SAR visualization approaches, the SARM method is unique because it also contains a compound design component and enables activity predictions. It was also adapted for studying structurally related compounds with multitarget activity[25] and—in combination with SAR analysis functions—for monitoring SAR progression of analog series.[11] In addition, SARMs were successfully used for expansion of screening hits from focused libraries.[26] Hence, SARMs provide a versatile

Received: March 3, 2019
Accepted: April 8, 2019
Published: April 18, 2019

DOI: 10.1021/acsomega.9b00595

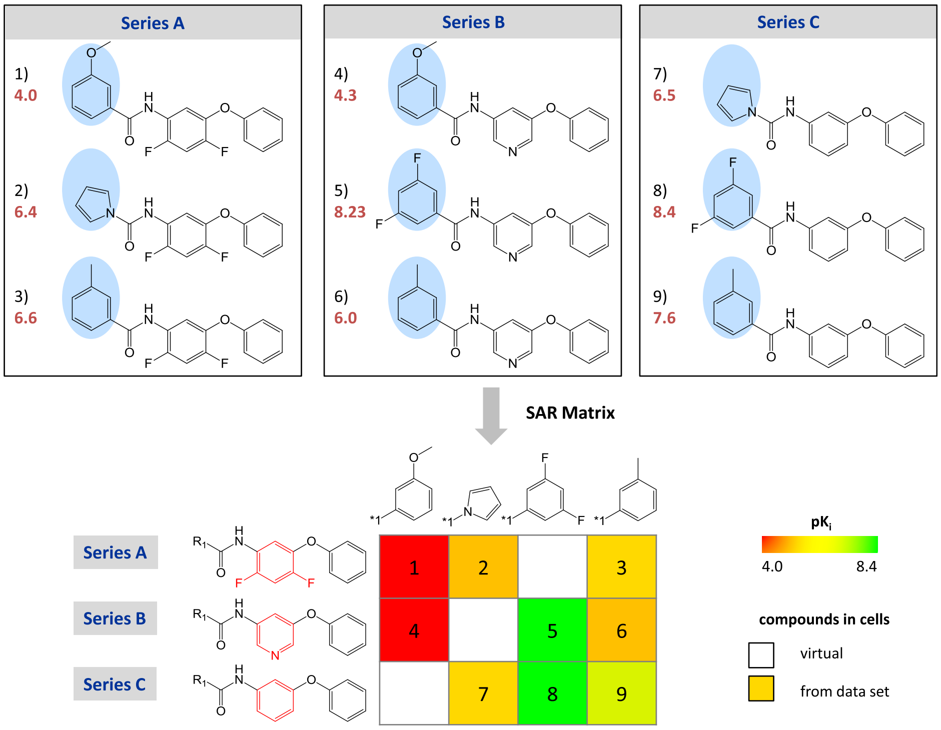


Figure 1. SARM. The design and generation of SARM is illustrated using three small compound series (A–C). Analogs from different series are consecutively numbered and their pK_i values are reported in red. Substitutions distinguishing analogs from individual series are shown on a blue background and substructures differentiating cores are colored red. The SARM combines these three series because they have structurally related cores. Each row contains a series and each column compounds from different series with the same substituent. Existing analogs are represented by cells that are color-coded by potency. In addition, empty cells represent virtual analogs. The figure was adopted from ref 31.

data structure and computational tool for medicinal chemistry applications. However, large ensembles of SARMs resulting from structurally complex data sets require thorough analysis, which benefits from matrix prioritization and additional operations.[25] Therefore, we have aimed at further expanding the SARM method and data structure to summarize SARM information content in a consistent manner, provide complementary visualization features, and enable fully interactive use. Herein, the integration of SARM with newly designed molecular grid maps and activity landscape models is reported, which further increases the utility for medicinal chemistry. Second-generation SARMs are applicable to very large compound data sets.

2. RESULTS AND DISCUSSION

In the first section, the SARM data structure is described. Subsequent sections report the design of molecular grid maps and activity landscape models and their integration with SARM. Key features of the extended approach are discussed. Additional methodological and computational details are provided in Materials and Methods. All compound data sets used in the following were obtained from the current release of ChEMBL.[27]

2.1. SARM Concept and Data Structure. SARMs are generated by systematically extracting compound series with clearly defined structural relationships from data sets and organizing them in matrices that are reminiscent of R-group tables. The identification and organization of structurally related analog series forming SARMs is based upon an extension of the matched molecular pair (MMP) concept. An MMP is defined as a pair of compounds that are only distinguished by a chemical modification at a single site[28] termed a chemical transformation.[29] MMPs are generated through systematic fragmentation of exocyclic single bonds in compounds.[29] Fragments are recorded in an index table as keys (core structures) and smaller values (substituents). The unique design principle underlying SARM generation is a dual fragmentation approach that generates MMPs at the level of compounds (first fragmentation) and core structures (second fragmentation).[24] In the first step, a series of analogs sharing a particular core are identified. In the second step, cores representing series are refragmented to identify all structurally related cores that only differ by a chemical change at a single site. Analog series with structurally related cores are then organized in an individual SARM, as illustrated in Figure 1. Each row contains an individual analog series and each column molecules from different series have the same substituent. Each cell in the matrix represents a unique compound. SAR information is conveyed by coloring cells according to compound potency. Empty cells represent virtual analogs consisting of currently unexplored combinations of cores and substituents. Thus, virtual compounds complement and further extend the analog space of related series. The potency of virtual candidates can be predicted on the basis of neighboring analogs in a matrix using local Free-Wilson-type QSAR models,[30,31] if actual data set compounds are available as neighbors.[31] Depending on the structural relationships contained in a given data set, varying numbers of SARMs are obtained, each of which comprises a unique subset of analog series with structurally related cores. Hence, the SARM data structure comprehensively captures structural relationships between compounds and series available in a given data set, organizes series into different subsets, and complements each subset with currently unexplored virtual analogs. Large and

DOI: 10.1021/acsomega.9b00595
ACS Omega 2019, 4, 7061−7069

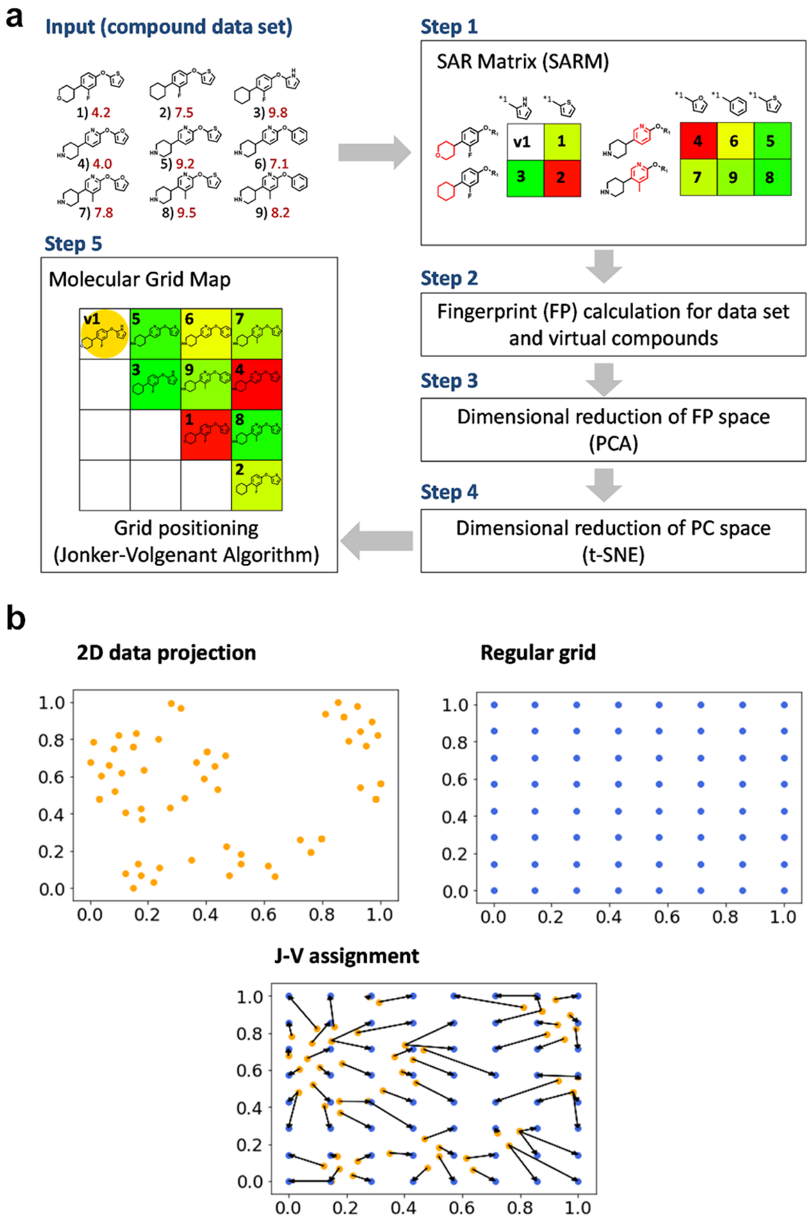

a
Input (compound data set)
1) 4.2
2) 7.5
3) 9.8
4) 4.0
5) 9.2
6) 7.1
7) 7.8
8) 9.5
9) 8.2
Step 1
SAR Matrix (SARM)
v1
1
3
2
4
6
5
7
9
8
Step 2
Fingerprint (FP) calculation for data set and virtual compounds
Step 3
Dimensional reduction of FP space (PCA)
Step 4
Dimensional reduction of PC space (t-SNE)
Step 5
Molecular Grid Map
Grid positioning (Jonker-Volgenant Algorithm)
b
2D data projection
Regular grid
J-V assignment
0.0
0.2
0.4
0.6
0.8
1.0

Figure 2. From SARMs to molecular grid maps. Panel (a) illustrates the generation of a grid map using compounds from SARMs. Input molecules are numbered and pK_i values are reported in red. In the SARMs (step 1), cells are color-coded according to compound potency and empty cells represent virtual analogs. Substructures distinguishing related cores are shown in red. After calculating compound fingerprints (step 2), dimension reduction is carried out using PCA (step 3) followed by *t*-SNE (step 4) to generate a 2D representation. Alternatively, dimension reduction is carried out via GTM. The 2D representations provide the basis for grid map generation using the J−V algorithm. In the final grid map, existing compounds are shown on a color-coded square background (experimental potency values) and virtual analogs on a circular background (predicted potency). If no SARM compound is assigned to a grid point in the map, the corresponding cell remains empty. Panel (b) shows an exemplary 2D representation resulting from reduction of fingerprint space and a corresponding regular grid and illustrates iterative J−V assignment of compounds to grid positions (indicated by black arrows).

structurally complex compound sets may yield tens to hundreds of SARMs that must be individually analyzed and compared.[25] Therefore, we have attempted to further expand the analytical capacity of the SARM approach and enable interactive prioritization of virtual candidate compounds.

2.2. Design of Molecular Grid Maps.

Two-dimensional (2D) molecular grid maps have been designed to provide a complementary view of a given compound data set organized in SARMs together with all resulting virtual analogs applying an alternative similarity measure (i.e., fingerprint similarity instead of MMP relationships). Hence, the grid map can be rationalized as a meta-level summary of SARM information content. Virtual compounds originating from SARMs cover chemical space around related analog series and connect SARMs to the grid map for the exploration of candidate compounds. From the grid map, compounds and virtual analogs of interest can be selected and traced back to their original SARM environment to consider other closely related candidates. This is made possible by consistently indexing compounds from SARMs in the grid map. Moreover, to provide a close link between compounds in grid maps and their original SARM environments, selecting a compound in a grid map also highlights all other compounds from the corresponding SARM in the map. This implementation aids in navigating the grid map and in comparing compounds from different SARMs together with their virtual analogs.

As a basis for grid map generation, real and virtual compounds are projected into a chemical descriptor (fingerprint) space that is transformed into a 2D representation through dimension reduction. In the resulting plane, compound positions are defined and increasing intercompound distance indicates increasing dissimilarity, without the need for explicit similarity calculations. Data points in the 2D representation are then mapped onto an appropriately sized regular grid by applying the Jonker–Volgenant (J–V) algorithm[32] to solve the associated linear assignment problem (LAP).[32]

Figure 2a provides an outline of the methodology using a simple example, leading from a SARM organization of a compound set to a 2D molecular grid map including virtual analogs. A small model data set comprising nine compounds (1–9) is shown, which yields two SARMs, each of which contains two series of analogs with structurally closely related cores. The first SARM also produces a virtual analog (v1) representing an unexplored core–substituent combination. For compounds 1–9 and v1, fingerprints are calculated and—as a complement to MMP-based organization—fingerprint space is projected through two-step or one-step dimension reduction onto an *x*,*y*-plane serving as a starting point for grid positioning using the J–V algorithms. Two-step dimension reduction is carried out using principal component analysis (PCA) followed by *t*-distributed stochastic neighbor embedding (*t*-SNE) and one-step dimension reduction using generative topographic mapping (GTM) (see Materials and Methods).

Figure 2b shows a prototypic 2D representation resulting from dimension reduction and a corresponding regular grid and illustrates the J–V assignment procedure. Compounds are iteratively assigned to grid points and alternative grid points are explored through combinatorial optimization to converge on a final grid positioning. The algorithmic assignment of compound positions from the 2D representation of fingerprint space to points on the regular grid preserves the grouping of compounds and intercompound distances as much as possible (further details are provided in Materials and Methods). To construct the *m* × *m* square grid map for compound positioning, the size of the grid typically exceeds the number of projected compounds. Therefore, "dummy" molecules are introduced with zero fingerprint bit settings to complement the compound data set.

The final grid map yields a complementary view of the SARM results for entire data set. Cells corresponding to those used in SARMs are then positioned on grid points to display compound structures and are color-coded according to observed or predicted potency values (or other molecular properties), as illustrated in Figure 2a. Indices are introduced to record the SARM localization of projected compounds and thereby enable toggling between grid maps and SARMs.

2.3. Exemplary Grid Maps.

Figure 3a shows a representative grid map for a set of 92 cyclin-dependent kinase 1/cyclin B1 inhibitors, 156 SARM-based virtual analogs (for which potency values have been predicted), and eight dummy molecules (forming an empty region in the global map). The map was generated on the basis of PCA/*t*-SNE dimension reduction. Known inhibitors are displayed in cells with square backgrounds and virtual analogs in cells with circular backgrounds. For virtual analogs, potency values were predicted via SARM-based QSAR[31] (other predictive methods can also be applied). The grid map reveals clustering of compounds according to different potency levels and suggests selection of candidate compounds from the predominantly

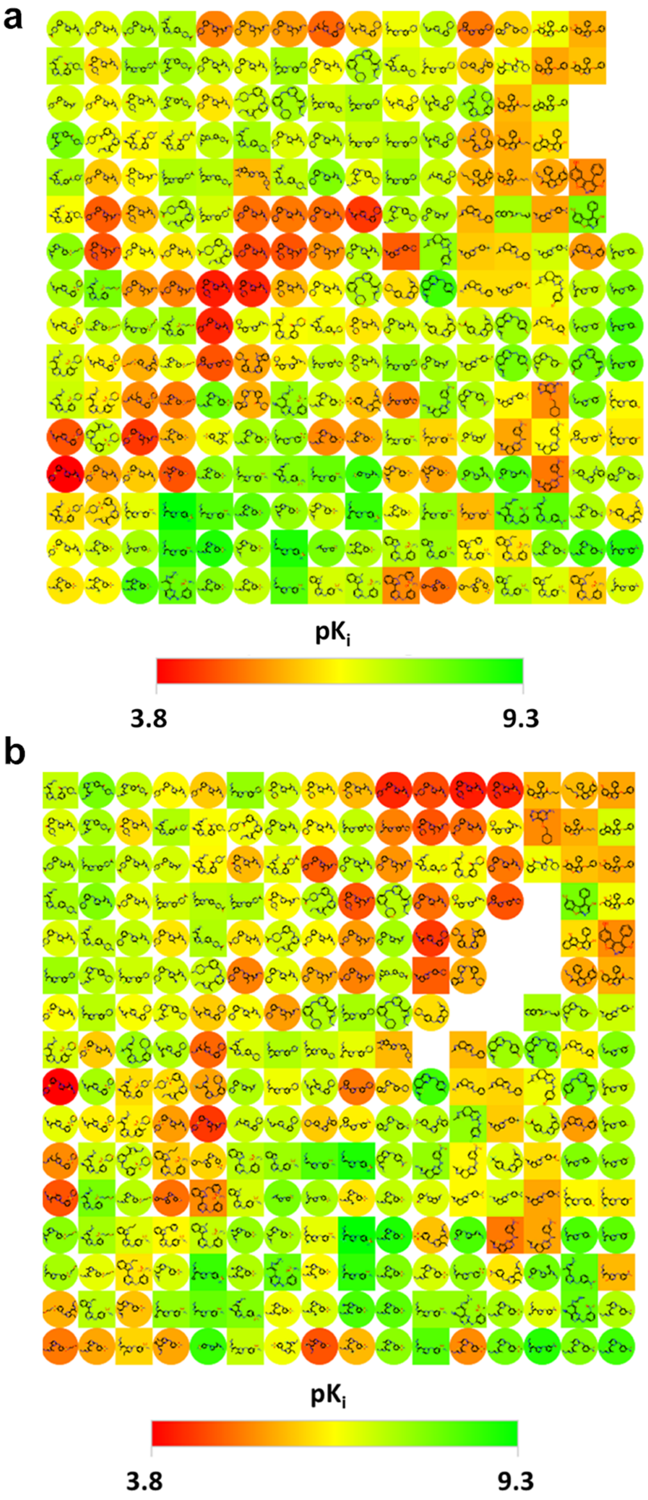


Figure 3. Small grid maps. Shown are exemplary grid maps for a set of 92 cyclin-dependent kinase 1/cyclin B1 inhibitors (ChEMBL ID: 1907602) and 156 virtual analogs resulting from SARM application. The representation of the map is according to Figure 2a. For virtual compounds, predicted potency values are used. The empty regions in the upper right of the maps results from grid points to which no SARM compounds are assigned. Grid maps were generated following dimension reduction using (a) PCA/*t*-SNE or (b) GTM.

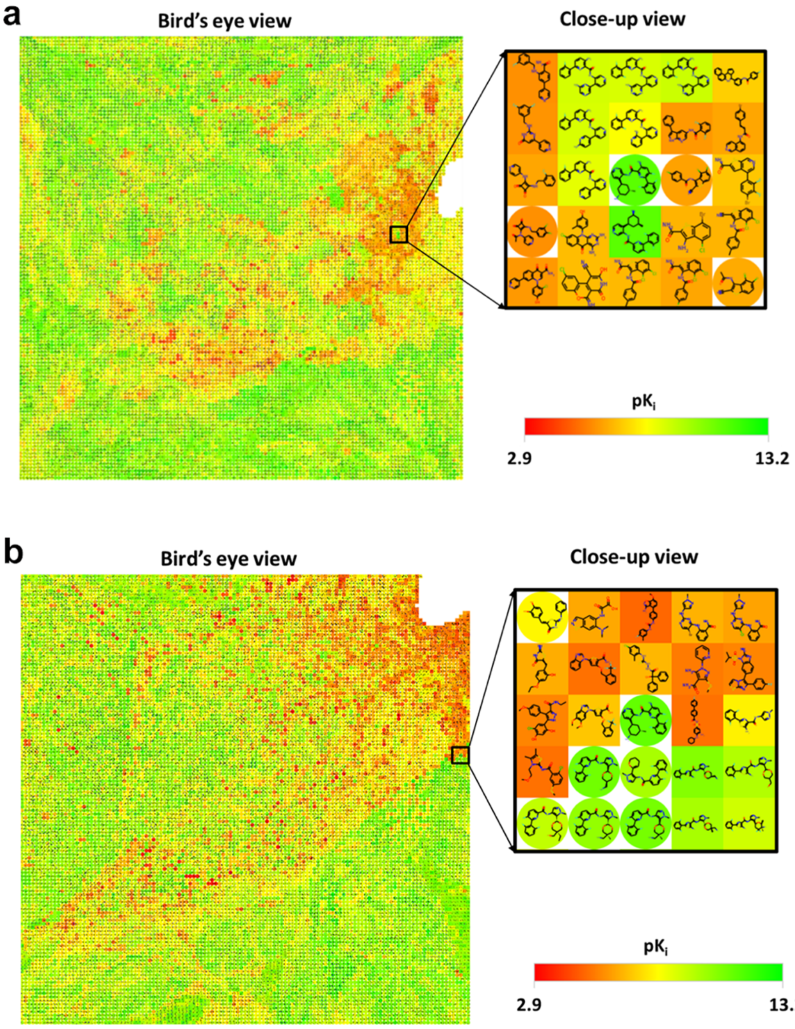


Figure 4. Large grid maps. Shown are exemplary grid maps for a set of 1772 kinase PIM inhibitors (ChEMBL ID: 2147) and 14 260 virtual analogs resulting from SARM application. On the left, a global view of the complete map is shown and, on the right, a close-up view of a selected region. For virtual compounds, predicted potency values are used. The empty region on the right side of the maps result from grid points to which no SARM compounds are assigned. Grid maps were generated following dimension reduction using (a) PCA/*t*-SNE or (b) GTM.

green regions emerging at the bottom of the map. Potent compounds and their neighbors in the map can then be viewed in their original SARM environments to provide additional SAR information. Hence, the grid map readily identifies attractive regions combining known compounds and virtual analogs that can be further explored. Figure 3b shows an alternative GTM-based grid map. As would be expected applying different dimension extension methods, relative compound positions in the PCA/*t*-SNE- and GTM-based maps partly differ. However, these maps reveal similar compound groupings and SAR patterns and are equally interpretable. In the GTM-based map, selection of candidate compounds would also focus on the green regions emerging at the bottom of the map, as discussed above. Thus, the comparison indicates that alternative dimension reduction approaches can be employed to yield interpretable grid maps, depending on methodological preferences and computational costs required for specific applications.

Figure 4a shows a PCA/*t*-SNE-based grid map comprising 16 129 compounds resulting from SARM processing of a set of 1772 PIM kinase inhibitors, yielding 14 260 virtual analogs (for which potency values have been predicted) and 97 dummy molecules (forming an empty region in the global map). Compound data sets of this size are usually difficult to analyze graphically. However, gird maps of large size are easily generated. A bird's eye view of the map of these kinase inhibitors and their virtual analogs, which occur in many different SARMs, reveals regions of predominantly high or low compound potency as well as regions of SAR discontinuity that are formed by compounds with varying potency levels. A close-up view is shown for an "island" of SAR discontinuity that is identified in the global map and contains four virtual candidates, one of which is predicted to be highly potent (green circular background in the center of the small map). Such views of compound subsets are straightforward to generate by delineating regions of interest in global maps, enabling interactive display of maps at different levels of resolution. Figure 4b shows the corresponding GTM-based grid map. Comparing these two maps, the global view of the PCA/*t*-SNE-based map in Figure 4a reveals a more extensive clustering of compounds according to different potency levels than the GTM-based map in Figure 4b, where compounds with different potency are more evenly distributed. On the other hand, the close-up view of the local environment of the

DOI: 10.1021/acsomega.9b00595
ACS Omega 2019, 4, 7061−7069

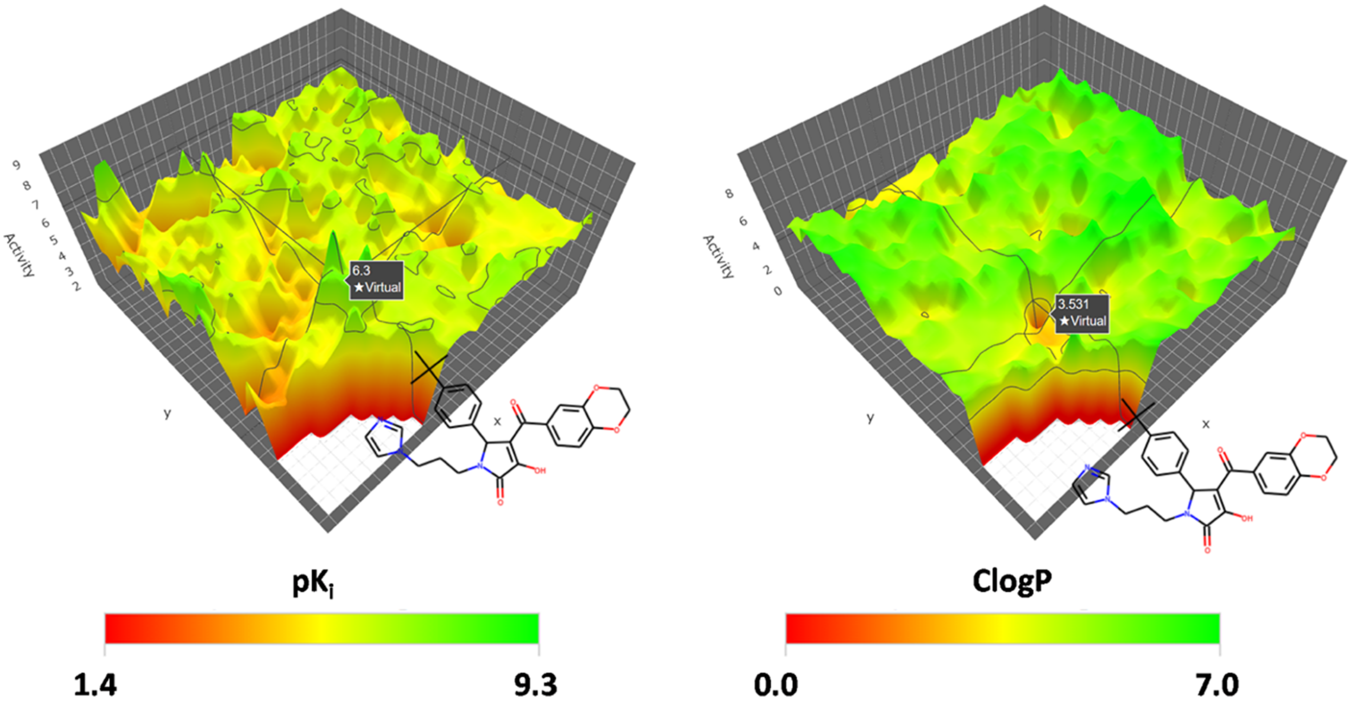


Figure 5. Property landscapes. Shown are two 3D property landscape models for a set of 130 E3 ubiquitin-protein ligase MDM2 inhibitors (ChEMBL ID: 1907611) and 551 virtual analogs from SARMs. In both landscapes, the same grid map forms the *x*,*y*-plane. On the left, an activity landscape is shown in which the third dimension (*z*-axis) is a potency surface. On the right, a corresponding model was generated with a Clog *P* surface instead. For a selected point on the landscape, the corresponding compound structure is displayed.

highly potent candidate compound that is also shown in Figure 4b displays a desirable enrichment with other potent analogs. However, since the PCA/*t*-SNE-based map is overall more structured than the corresponding GTM-based map we would, in this case, give preference to the former, which would make it easier to focus on local regions of interest. Of course, there is no a priori reason to select one map instead of the other since corresponding maps for different compound data sets can easily be compared and analyzed in context.

2.4. Property Landscapes.

Compound information provided by 2D grid maps is further complemented by 3D property landscape models generated on the basis of these maps. Instead of coloring grid map cells by potency or values of other molecular properties, these values are added as a third dimension to the map from which a coherent surface is interpolated (see Materials and Methods). Figure 5 (left) shows an activity landscape for a set of 130 × 10^3 ubiquitin-protein ligase inhibitors. In addition, a corresponding landscape is shown that was built on the basis of the same grid map but using calculated Clog *P* values (a measure of hydrophobicity) instead of potency values (right). These 3D landscape views provide both grid and elevation-dependent (color-coded) property information.

The topology of the activity landscape is rugged, revealing the formation of a number of activity cliffs. Thus, the 3D representation provides an alternative view of SAR discontinuity, and compounds involved in the formation of activity cliffs can be interactively selected. By contrast, the topology of the Clog *P* landscape is smooth, reflecting generally high hydrophobicity of the compound set. Nonetheless, small islands of low hydrophobicity also emerge in the landscape, which can be inspected for potential candidates such as the virtual compound shown in Figure 5. This compound originates from an SARM and is of particular interest because it maps to an activity cliff region in the activity landscape, is predicted to be highly potent, and also occupies a small pocket of low hydrophobicity in the Clog *P* landscape. Hence, in an optimization effort, it would be an attractive candidate for selection and synthesis.

2.5. Concluding Remarks.

SAR analysis and visualization methods play an important role in computational medicinal chemistry. A major task is translating the results of SAR analysis into compound design. During lead optimization, the key question is which compound(s) to make next to further advance a series. Going beyond classical QSAR, only few computational concepts have been introduced to aid in decision making during chemical optimization. The SARM approach was originally developed to bridge between structural analysis, SAR visualization, and compound design. This combination sets it apart from other SAR analysis and visualization methods. SARM identifies series of structurally related compounds, systematically organizes them, and generates virtual analogs. It can also be applied to given series to produce virtual candidates for optimization efforts. We have been interested in extending the SARM approach to further increase its analytic capacity, enable interactive application to large data sets, and aid in compound selection from matrices. Our study was inspired by the challenge to analyze increasingly large and heterogeneous data sets to complement lead optimization with SAR insights from external sources. In medicinal chemistry, this still represents largely uncharted scientific territory for which new computational concepts must be developed. SARM application makes it possible to combine compound series from different sources and generate virtual candidates from multiple series. However, to these ends, versatile analysis capabilities are required. Therefore, the SARM method has been integrated with newly designed 2D molecular grid maps and 3D property landscape models that are based on these maps. The SARM data structure, which is reminiscent of R-group tables, and molecular grid maps provide complementary reference frames for graphical analysis of compound data sets and SAR visualization. Molecular grid maps organize real and virtual compounds resulting from SARM application for entire data sets, regardless of their composition, and yield a global view of SARM information content. Furthermore, 2D grid maps and 3D property landscapes provide complementary SARM-associated graphical representations for the analysis of SARs and other structure—property relationships. Effective navigation and combination of SARMs, grid maps, and landscape models is facilitated through a new web-based interface, which also renders the approach applicable to very large compound data sets, as shown herein. Combining local and global SAR views provided by individual matrices and grid maps, respectively, adds another dimension to large-scale SAR exploration and helps to focus on promising candidate compounds. With its additional components, the second-generation SARM approach should be of considerable interest for practical applications.

For the practice of medicinal chemistry, the following conclusions can be drawn from our study:

(i) Analog series are organized in SARMs, which visualize SARs and generate virtual close-in analogs as candidates for synthesis.

(ii) Depending on the structural relationships between compound series, varying numbers of SARMs are obtained. The 2D grid map is designed to provide a global view of real and virtual compounds from SARMs and visualize them in context. For practical applications, the grid map is often easier to inspect than multiple SARMs to obtain an initial overview of available compounds and virtual candidates. From the map, compounds of interest can be immediately traced back to their original SARMs.

(iii) The grid map is closely connected to original SARMs and all compounds originating from a given SARM are also highlighted in the map when selecting a compound. In addition, 3D activity landscape representations can be viewed together with the grid map and corresponding SARMs to visualize global and local SAR characteristics in an intuitive manner.

(iv) Furthermore, the grid map summarizes all SAR features of given compound sets consisting of multiple series. Local SARs can be further inspected using the R-group table format of SARMs by selecting compounds of interest in the grid map. Given their R-group table like layout, SARMs provide an intuitive access to chemical modifications and local SAR features.

(v) On the other hand, compounds of interest identified in individual SARMs can also be traced in the grid map and inspected together with compounds from the same or other SARMs.

(vi) Thus, going back and forth between SARMs and their grid map highlights compounds in complementary SAR environments and enables the selection of virtual candidates for synthesis. Taken together, the second-generation SARM approach focuses on interactive analysis of analog series, associated SARs, and virtual candidates.

3. MATERIALS AND METHODS

SARMs, molecular grid maps, and 3D activity landscape models were implemented in Python with aid of RDKit,[33] scikit-learn,[34] lapjv,[35] and Plotly.js.[36]

3.1. Molecular Grid Map.

3.1.1. Dimension Reduction. For the generation of grid maps, all compounds (real and virtual) were represented as MACCS structural keys[37] (167 bits) calculated with RDKit. The resulting fingerprint space was subjected to two-step dimension reduction. Initially, PCA was carried out to generate 10 or 50 (orthogonal) principal components (PCs). Following this preprocessing step, *t*-SNE[38] was performed on PC space for 3000 iterations with an initial random seed and *t*-SNE parameter settings *n*_components = 2 and perplexity = 10, resulting in a 2D projection of fingerprint space.

As an alternative dimension reduction approach, GTM[39] was applied. GTM is a nonlinear dimension reduction method and represents an extension of the self-organizing map (SOM) neural network concept.[40] In SOMs, molecules are often mapped to the same cell. Since discrete 2D compound coordinates are required for grid positioning, SOM projections are not suitable for grid map generation. However, GTM yields discrete coordinates. Thus, for comparison with two-step dimension reduction, GTM calculations were carried out with ugtm.[41] To generate mean GTM 2D projections, eGTM implemented in ugtm was carried out with the following parameter settings: $m = 2$ (m is the square root of the number of RBF centers), and $k = 30$ (k is the square root of the number of GTM nodes), and mode = “mean”.

3.1.2. J—V Algorithm. Each data point in the *t*-SNE or GTM map was then assigned to a regular grid using the J—V algorithm in its lapjv implementation. The J—V algorithm was originally developed for solving an LAP through combinatorial optimization. For generating molecular grid maps, the LAP task consisted of optimally assigning *n* points in the *t*-SNE or GTM map to an evenly spaced grid.

LAP is generally defined as

$$\text{argmin} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{1}$$

subject to

$$\sum_{i=1}^{n} x_{ij} = 1, \qquad (j = 1, ..., n) \tag{1a}$$

$$\sum_{j=1}^{n} x_{ij} = 1, \qquad (i = 1, ..., n) \tag{1b}$$

$$x_{ij} = \{0, 1\} \qquad i, j = 1, ..., n \tag{1c}$$

where c_{ij} is the cost of assigning point i in the *t*-SNE map (p_i^{tSNE}) to point j in the regular grid (p_j^{grid}) and x_{ij} is the matching indicator. A setting of $x_{ij} = 1$ means that point i in the *t*-SNE map is assigned to j in the regular grid and $x_{ij} = 0$ indicates no assignment. The cost associated with minimization is given by

$$c_{ij} = \left\| p_i^{\text{tSNE}} - p_j^{\text{grid}} \right\|^2 \times a \tag{2}$$

where a is the scaling factor ($a = 100\,000/\max\{c_{ij}\}$).

Optimal assignment of $X = \{x_{ij}\}$ reveals the projection of p_i^{tSNE} to p_j^{grid}. The resulting molecular grid map displays the chemical graphs of assigned compounds with observed or predicted biological activities. To construct the $m \times m$ square grid map for compound positioning, dummy molecules are introduced with zero fingerprint bit settings to complement the compound data set.

3.2. Property Landscapes. On the basis of molecular grid maps, 3D property (activity) landscape models[42] were generated by adding compound property values as a third dimension (z-axis), followed by algorithmic interpolation of a coherent property (potency) surface.[42] From distributed data points, a coherent surface is interpolated using the kriging function[42,43] or related approaches. Landscape calculations were carried out using Plotly.js.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de. Phone: 49-228-7369-100 (J.B.).

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Author Contributions

The study was carried out and the manuscript written with contributions of all authors. All authors have approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Yuichi Horita for valuable contributions to the development of a graphical user interface for the molecular grid map.

REFERENCES

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

(2) 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.; Terfloth, 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.

(3) Varnek, A.; Baskin, I. Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis? *J. Chem. Inf. Model.* **2012**, *52*, 1413−1437.

(4) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571−5578.

(5) Guha, R.; Van Drie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.

(6) Nicolaou, C. A.; Brown, N.; Pattichis, C. S. Molecular Optimization Using Computational Multi-Objective Methods. *Curr. Opin. Drug Discov. Dev.* **2007**, *10*, 316−324.

(7) Segall, M. Advances in Multi-Parameter Optimization Methods for De Novo Drug Design. *Expert Opin. Drug Discovery* **2014**, *9*, 803−817.

(8) 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.

(9) Agrafiotis, D. K.; Wiener, J. J. M. Scaffold Explorer: An Interactive Tool for Organizing and Mining Structure−Activity Data Spanning Multiple Chemotypes. *J. Med. Chem.* **2010**, *53*, 5002−5011.

(10) Munson, M.; Lieberman, H.; Tserlin, E.; Rocnik, J.; Ge, J.; Fitzgerald, M.; Patel, V.; Garcia-Echeverria, C. Lead Optimization Attrition Analysis (LOAA): A Novel and General Methodology for Medicinal Chemistry. *Drug Discovery Today* **2015**, *20*, 978−987.

(11) Shanmugasundaram, V.; Zhang, L.; Kayastha, S.; de la Vega de León, A.; Dimova, D.; Bajorath, J. Monitoring the Progression of Structure−Activity Relationship Information during Lead Optimization. *J. Med. Chem.* **2016**, *59*, 4235−4244.

(12) Maynard, A. T.; Roberts, C. D. Quantifying, Visualizing, and Monitoring Lead Optimization. *J. Med. Chem.* **2015**, *59*, 4189−4201.

(13) Vogt, M.; Yonchev, D.; Bajorath, J. Computational Method to Evaluate Progress in Lead Optimization. *J. Med. Chem.* **2018**, *61*, 10895−10900.

(14) 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 Discov. Today* **2010**, *15*, 630−639.

(15) Stumpfe, D.; Bajorath, J. Methods for SAR Visualization. *RSC Adv.* **2012**, *2*, 369−378.

(16) Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A new SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926−5937.

(17) Renner, S.; van Otterlo, W. A. L.; Dominguez Seoane, M.; Möcklinghoff, S.; Hofmann, B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T. I.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-guided Mapping and Navigation of Chemical Space. *Nat. Chem. Biol.* **2009**, *5*, 585−592.

(18) 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.

(19) Ertl, P. Intuitive Ordering of Scaffolds and Scaffold Similarity Searching Using Scaffold Keys. *J. Chem. Inf. Model.* **2014**, *54*, 1617−1622.

(20) Varin, T.; Schuffenhauer, A.; Ertl, P.; Renner, S. Mining for Bioactive Scaffolds with Scaffold Networks: Improved Compound Set Enrichment from Primary Screening Data. *J. Chem. Inf. Model.* **2011**, *51*, 1528−1538.

(21) Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure-Activity Relationship Trailing. *J. Med. Chem.* **2011**, *54*, 2944−2951.

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

(23) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.

(24) 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.

(25) Gupta-Ostermann, D.; Bajorath, J. The "SAR Matrix" Method and its Extensions for Applications in Medicinal Chemistry and Chemogenomics. *F1000Research* **2014**, *3*, No. e113.

(26) Gupta-Ostermann, D.; Hirose, Y.; Odagami, T.; Kouji, H.; Bajorath, J. Prospective Compound Design Using the "SAR Matrix" Method and Matrix-Derived Conditional Probabilities of Activity. *F1000Research* **2015**, *4*, No. e75.

(27) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.

(28) 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.

(29) 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.

(30) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399.

(31) Gupta-Ostermann, D.; Shanmugasundaram, V.; Bajorath, J. Neighborhood-based prediction of novel active compounds from SAR matrices. *J. Chem. Inf. Model.* **2014**, *54*, 801−809.

(32) Jonker, R.; Volgenant, A. A Shortest Augmenting Path Algorithm for Dense and Sparse Linear Assignment Problems. *Computing* **1987**, *38*, 325−340.

(33) RDKit: Cheminformatics and Machine Learning Software. 2013. http://www.rdkit.org (accessed Jan 2, 2019).

(34) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(35) lapjv: Linear Assignment Problem Solver Using Jonker-Volgenant Algorithm 2018. https://github.com/src-d/lapjv (accessed Jan 2, 2019).

(36) Plotly.js: JavaScript Graphing Library. https://plot.ly/javascript (accessed Jan 2, 2019).

(37) *MACCS Structural Keys*; Accelrys: San Diego, CA, 2011.

(38) Van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(39) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215−234.

(40) Kohonen, T. *Self-Organizing Maps*; Springer: Heidelberg, 2001.

(41) ugtm: A Python Package for Generative Topographic Mapping (GTM). https://ugtm.readthedocs.io (accessed March 10, 2019).

(42) Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 1021−1033.

(43) Cressie, N. *Statistics for Spatial Data*; Wiley: New York, 1993.