Communication

www.molinf.com

molecular informatics

DOI: 10.1002/minf.202000045

The SAR Matrix Method and an Artificially Intelligent Variant for the Identification and Structural Organization of Analog Series, SAR Analysis, and Compound Design

Atsushi Yoshimori[a] and Jürgen Bajorath*[b]

Abstract: The SAR Matrix (SARM) approach was originally conceived for the systematic identification of analog series, their structural organization, and graphical structure-activity relationship (SAR) analysis. For structurally related series, SARMs also produce virtual candidate compounds. Hence, SARM represents a unique computational approach establishing a direct link between SAR visualization and compound design. The SARM data structure is reminiscent of R-group tables and hence easily accessible from a chemical perspective, although the underlying algorithmic basis is complex. The SARM concept has been extended in different ways to further increase its analytical and design capacity. While the efforts were largely driven from a research perspective, they have also increased the utility for practical applications. Among others, extensions include approaches for SARM-based compound activity prediction, the generation of a large SARM database for analog searching, and the design of a deep learning architecture for advanced analog design taking chemical space information for target families into account. Herein, the SARM approach and its extensions are discussed within their scientific context.

Keywords: SAR Matrix · analog series · molecular fragmentation · SAR analysis · visualization · molecular grid maps · compound design · activity prediction · deep learning

In this contribution to the *Strasbourg Summer School in Chemoinformatics – 2020*, we present the SAR Matrix (SARM) concept and its methodological extensions.

SARM represents a unique computational methdology because it combines the *(i)* systematic detection of structural relationships and analog series in compound data sets with *(ii)* exploration and visualization of structure-activity relationships (SARs) and *(iii)* analog design and searching. In the following, different development stages are described and extensions of the SARM approach are rationalized. In addition, practical applications are discussed.

The SARM approach was originally conceptualized for the systematic identification, organization, and SAR analysis of analog series aided by SAR visualization.[1] SARM synonymously refers to the underlying computational methodology and matrix-based compound data structure. The SARM method systematically detects structural relationships between compounds in large data sets, extracts analog series, and organizes these series in matrices on the basis of related core structures.[1,2] The resulting SARMs are reminiscent of conventional R-group tables used in medicinal chemistry and hence easily interpretable, although their information content is much higher. The construction of SARMs is illustrated in Figure 1. SARMs are generated by applying an exhaustive dual-step compound fragmentation scheme adapted from matched molecular pair (MMP) analysis. An MMP is defined as a pair of compounds that are only distinguished by a chemical modification at a single site.[3] For MMP generation, a computationally efficient algorithm is available that deletes one to three exocyclic single bonds in compounds per iteration and stores the resulting fragments in an index table.[3] In the first step of SARM construction, all database compounds are subjected to standard MMP fragmentation and the first index table is created. This table contains key fragments representing core structures of compounds and smaller value fragments representing substituents (R-groups). Compounds sharing the same key (core) in first index table form

[a] *A. Yoshimori*
Institute for Theoretical Medicine, Inc.
26-1 Muraoka-Higashi 2-chome
Fujisawa, Kanagawa 251-0012, Japan
E-mail: yoshimori@itmol.com
[b] *J. Bajorath*
Department of Life Science Informatics
Bonn-Aachen International Center for Information Technology
Rheinische Friedrich-Wilhelms-Universität Bonn
Endenicher Allee 19c, D-53115 Bonn, Germany
Tel: +49-228-7369-100
Fax: +49-228-7369-101
E-mail: bajorath@bit.uni-bonn.de

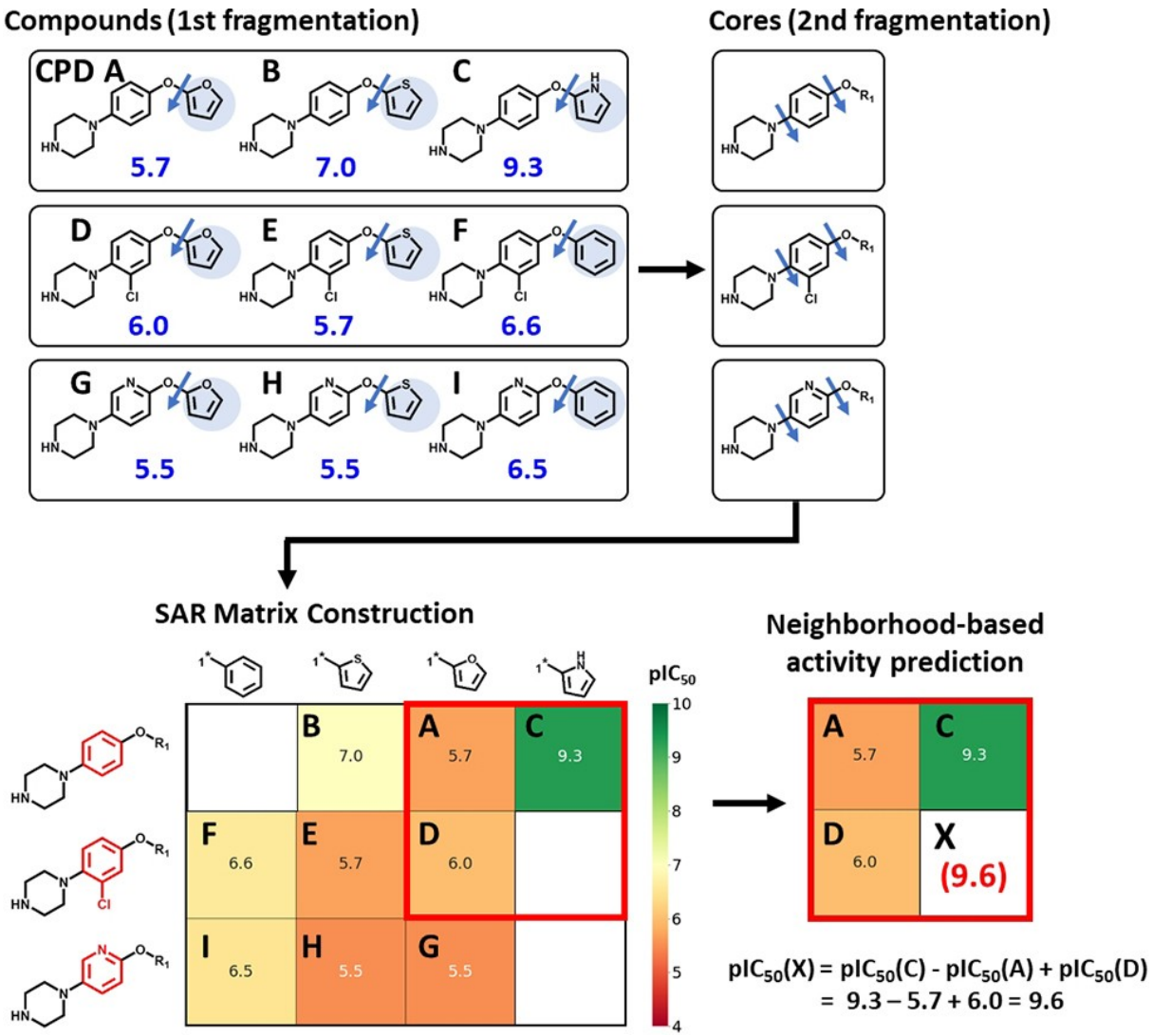

Figure 1. SARM generation and activity prediction. The construction of a SARM is illustrated using a small set of nine compounds (CPD A–C) whose pIC_{50} values are given in blue. Substituents (R-groups) distinguishing analogs with related yet distinct core structures are shown on a blue background. The SARM is obtained by applying a dual-step fragmentation scheme described in the text and contains analog series with structurally related cores. Substructures distinguishing these cores are colored red. Each filled cell in the SARM represents an existing compound (A–C) and an empty cell represents a virtual analog (i.e., a non-existing core-substituent combination). Cells representing real compounds are color-coded by potency (pIC50 value). Potencies of virtual analogs can be predicted on the basis of real compounds using local Free-Wilson QSAR models that depend on defined neighborhood constellations (illustrated on the lower right). X indicates the virtual analog whose potency is predicted.

a matching molecular series (MMS) defined as a series of compounds that only differ by a chemical modification at a single site.[4] Hence, an MMS represents an analog series with a single substitution site. Cores from the first index table are then subjected to a second round of MMP fragmentation, which again yields key and value fragments that are stored in the second index table. The second fragmentation step is unique to the SARM approach and identifies structurally analogous cores representing a so-called "key MMS". All analog series defined by a key MMS are then organized in an individual SARM, in which each row contains an analog series and each column analogs from structurally related yet distinct series carrying the same substituent. Each filled cell in a SARM contains a unique existing (real) compound. In addition, empty cells represent virtual analogs comprising currently non-existing core-substituent combinations. Cells of real compounds are color-coded by potency (or other molecular properties). Therefore, a SARM contains related analogs as well as varying numbers of virtual candidate compounds and visualizes SAR patterns. Depending on the number of analog series contained in a compound data set and their

structural relationships, a data set typically yields an ensemble of SARMs ranging from a few to many (hundreds or even thousands). SARMs from an ensemble can be prioritized indifferent ways, for example, on the basis of their SAR information they contain or the proportion of virtual candidate analogs.[2] SAR information content of individual SARMs can be quantified using numerical SAR analysis function. Furthermore, annotation of cells such as color-coding or labeling makes it possible to adopt the SARM data structure for other applications such as, for example, the systematic analysis of multi-target activities of closely related compounds.[2] Moreover, SARM ensembles can also be successively calculated for evolving compound series during lead optimization and their SAR information content can be determined as measure of SAR progression.[5] In this case, SARM ensembles serve as a diagnostic tool.

The SARM data structure bridges between SAR exploration and compound design. Accordingly, one would like to prioritize virtual compounds contained in SARMs for further evaluation. To these ends, SARM-based approaches for activity prediction were considered and local QSAR models focusing on compound neighborhoods of virtual candidates in SARMs developed,[6] as also illustrated in Figure 1. These local QSAR models were designed following Free-Wilson principles of additivity.[6,7] Accordingly, for a virtual analog, a qualifying neighborhood in a SARM consists of three real compounds with known potency and the virtual analog, which share cores and R-groups in a pairwise manner, thus permitting the prediction of the potency of the virtual candidate on the basis of individual core and substituent contribution applying the formula shown in Figure 1. For a given virtual candidate, multiple qualifying compound neighborhoods often exist in SARMs, which are systematically identified. These neighborhoods enable multiple potency predictions that can be compared for consistency. Local SARM-based QSAR models are applicable in hit-to-lead or lead optimization projects and have yielded accurate predictions on a variety of data sets,[6] despite their simplicity. However, given the general applicability domain of QSAR approaches, the application of these local models is confined to regions of SAR continuity in SARMs. In regions of SAR continuity, gradual changes in structure are accompanied by gradual changes in potency. In SARMs, such regions are easily identified when cells are color-coded by potency.

In addition, a conceptually distinct activity prediction approach has been devised for SARMs that is generally applicable to binary activity assignments (active/inactive) and any SAR environment. Binary activity assignments are typically obtained for compounds from single-concentration biological screens when a threshold of inhibition or residual activity is applied. The second predictive approach relies on the derivation of conditional probabilities of activity based upon the distribution of individual cores and substituents in pre-classified active and inactive compounds across SARM ensembles.[8] Conditional probabilities of cores and substituents from screening hits are combined to predict the activity of virtual analogs from SARMs and prioritize candidate compounds for hit expansion.[8] Given the analog series content of SARMs, this predictive approach is primarily applicable to target-focused screening libraries comprising subsets of structurally related compounds. In test calculations, activity predictions on screening data based on SARM-derived conditional probabilities met or exceeding the performance of machine learning models.[8] Furthermore, the approach was successfully applied prospectively to predict novel active compounds for cancer targets on the basis of cell-based screens.[9] In this case, focused libraries consisting of secondary structure mimetic compounds were screened in cell-based anticancer assays and subsequently analyzed in SARMs, which yielded virtual analogs. Activity predictions on the basis of the screening data prioritized limited numbers of virtual analogs. Twenty of these virtual candidates were synthesized and tested, confirming five new active compounds exhibiting dose-response behavior.[9]

Since large SARM ensembles are hard to analyze interactively, it was attempted to generate a complementary global view of real and virtual SARM compounds and their relationships. These efforts led to the design of a new molecular grid map[10] illustrated in Figure 2. For generating grid maps, fingerprint similarity is applied as an alternative to MMP relationships because the grid map is required to combine structurally related and unrelated analog series from different SARMs. The generation of molecular grid maps is summarized in Figure 2A. For all existing and virtual compounds from a SARM ensemble, a selected fingerprint is calculated, hence encoding compounds in fingerprint space. Then, dimension reduction of fingerprint space is carried out to generate a 2D projection. For effective dimension reduction, different approaches are available. For example, a two-step procedure can be applied including pre-processing using principal component analysis (PCA) followed by t-distributed stochastic neighbor embedding (t-SNE),[11] as indicated in Figure 2A. Alternatively, one-step dimension reduction using generative topographic mapping (GTM)[12] has also been effective.[10] In the resulting 2D projection, increasing distance between compound positions scales with increasing dissimilarity. Compound positions are then transferred to points of a regular grid using the Jonker-Volgenant algorithm for linear assignment.[13] Accordingly, compound positions are iteratively mapped to grid points, alternative grid points are explored through combinatorial optimization, and a final grid positioning is derived attempting to preserve the clustering of compounds and inter-compound distances in the 2D projection. Then, SARM cells of mapped compounds containing compound structures and property range-based color coding are positioned on the assigned grid points. Through consistent indexing, the SARM location of compounds in the grid map is recorded making it possible to interactively

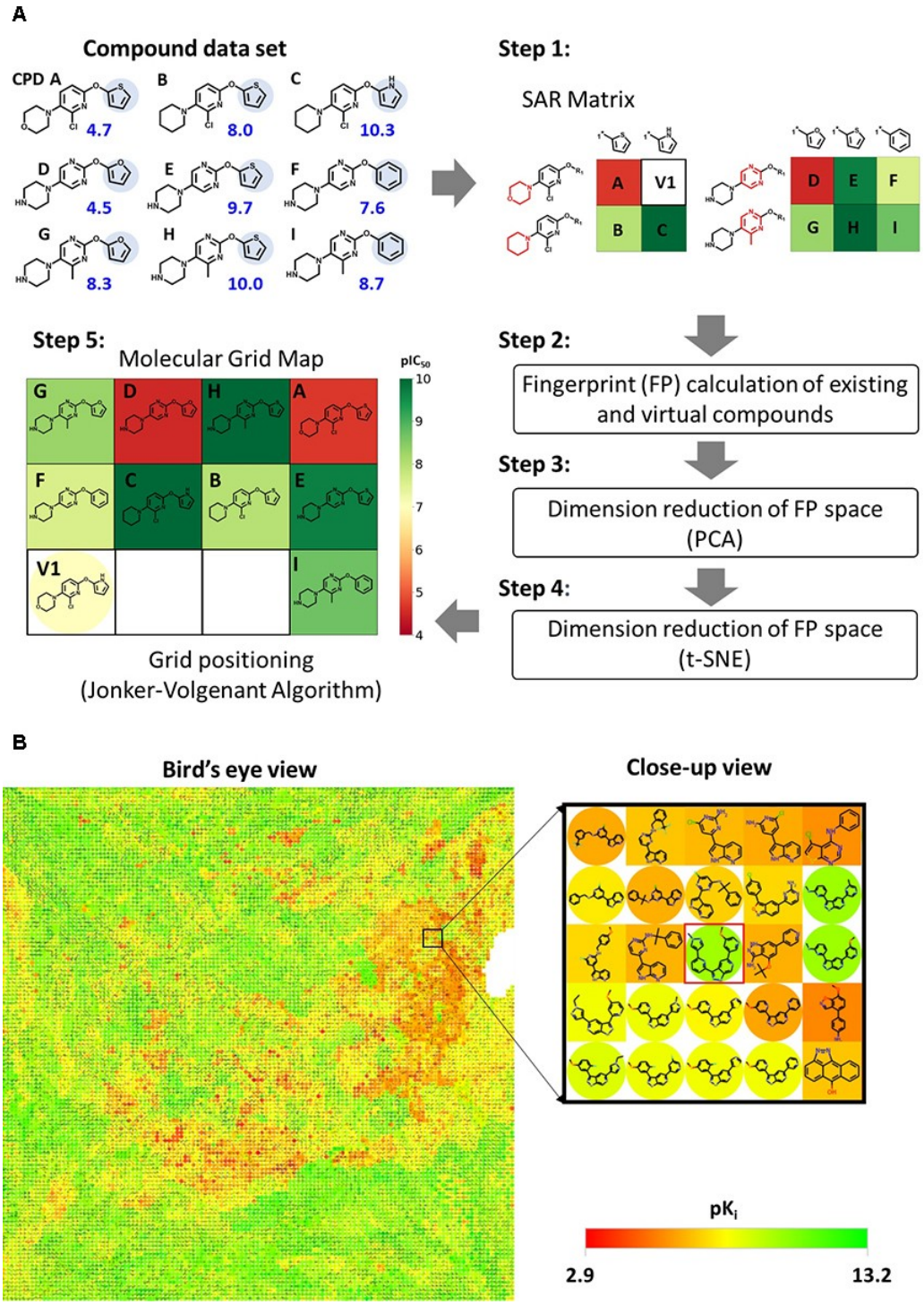

Figure 2. Molecular grid map. (**A**) The generation of a molecular grid map is illustrated using nine small compounds (CPD A–I). pIC_{50} values are given in blue. Substitutions distinguishing individual compounds are shown on a light blue background. Step 1: For the compound set, SARMs are generated with cells color-coded according to compound potency. V1 is a virtual analog. Steps 2–5 summarize different stages involved in constructing the grid map, as described in the text. If no SARM compound is assigned to a grid point in the map, the corresponding cell remains empty. In (**B**), a molecular grid map is shown for a set of 1772 PIM kinase inhibitors (ChEMBL[24] ID: 2147) and 14260 virtual analogs originating from SARMs constructed for the inhibitor set. A bird's eye view of the complete map is shown and a close-up view of a selected region. Squares indicate real and circles virtual compounds. Circles are color-coded according to potency values predicted using SARM-based local QSAR models. The white region on the right side of the large map consists of grid points to which no SARM compounds were assigned. The format of the figure was adapted from reference [10].

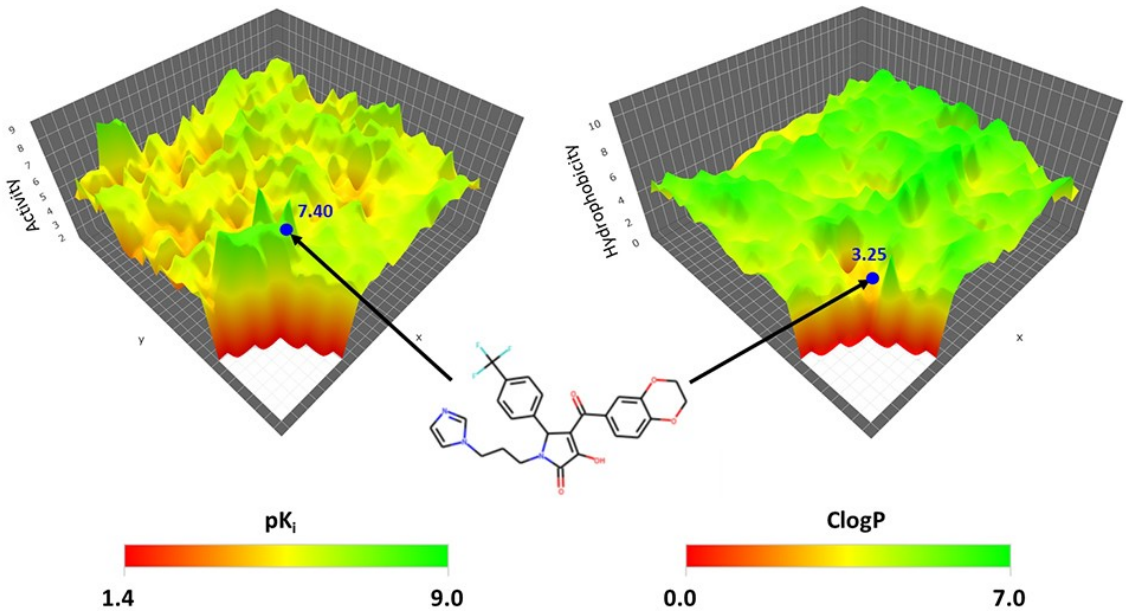

Figure 3. Property landscapes. Shown are two 3D property landscape models for a set of 130 E3 ubiquitin-protein ligase MDM2 inhibitors (ChEMBL ID: 1907611) and 551 virtual analogs from SARMs. The models include an activity landscape and a ClogP landscape accounting for compound hydrophobicity. In both cases, the molecular grid map of the compound data set formed the plane at the bottom to which a property surface was added in the third dimension. An exemplary virtual candidate compound with high predicted potency and low calculated hydrophobicity is shown and its positions in the landscapes are indicated. The figure was adapted from ref. [10] and modified.

navigate the grid map and trace compounds and their environments back to original SARMs. Figure 2B shows an exemplary grid map for a larger compound data set. The complementary global and local views provided by the map and corresponding SARMs and their interplay increase the capacity of graphical SAR analysis, as illustrated in Figure 2B. In addition, relationships between existing compounds and virtual analogs can be explored on the basis of alternative similarity measures.

The grid map of a given data set also provides a basis for the generation of 3D property landscape models.[14] In 3D property landscapes such as activity landscapes, compound activity values are added to a 2D projection of chemical reference space as a third dimension and a hypersurface is interpolated from these distributed values.[14] Figure 3 shows exemplary grid map-based activity and hydrophobicity landscapes for a compound data set and illustrates how their comparison aids in compound selection. The inclusion of grid map-based property landscapes enables the concomitant consideration of multiple optimization relevant criteria.

The algorithmic efficiency of the SARM approach has made it also possible to generate SARMs for the entire Leads Now subset of the ZINC database[15] comprising ~3.7 million compounds, thus yielding a very large collection of matrices termed Mega SARM.[16] For building Mega SARM, compound decomposition was not only carried out by systematic deletion of exocyclic bonds, but also by deletion of bonds according to retrosynthetic rules,[17] thereby further increasing the synthetic accessibility of virtual analogs. Conventional fragmentation and fragmentation according to retrosynthetic rules produced 204,825 and 47,786 unique SARMs, respectively. These SARMs contained a total of 1,531,669 virtual analogs of compound series extracted from ZINC. Hence, Mega SARM provides a very large resource structurally organized compound series and virtual analogs complementing these series.

To systematically identify existing and virtual analogs of query compounds in SARMs, an algorithm for SARM-based analog searching was developed.[16] Importantly, the algorithm not only identifies compounds with R-group replacements but also others with chemical changes in core structures. Thus, it further extends searchable analog space compared to conventional substructure-based methods. As illustrated in Figure 4, the search procedure partly differs depending on whether query compounds are contained in the database used to generate SARMs or not. For a novel query compound, the approach involves the generation of new SARMs, as further described below.

Figure 4A shows the search procedure if a query compound F is contained in the first index table for SARM construction. Initially, the index table is searched for keys generated from compound F, resulting in the identification of key C. In the next step, the second index table is searched for keys obtained from C, yielding key 2B. Then, the complete key MMS is assembled including keys C and D. In the fourth step, the MMSs of key C and D are extracted from the first index table. These MMSs contain all com-

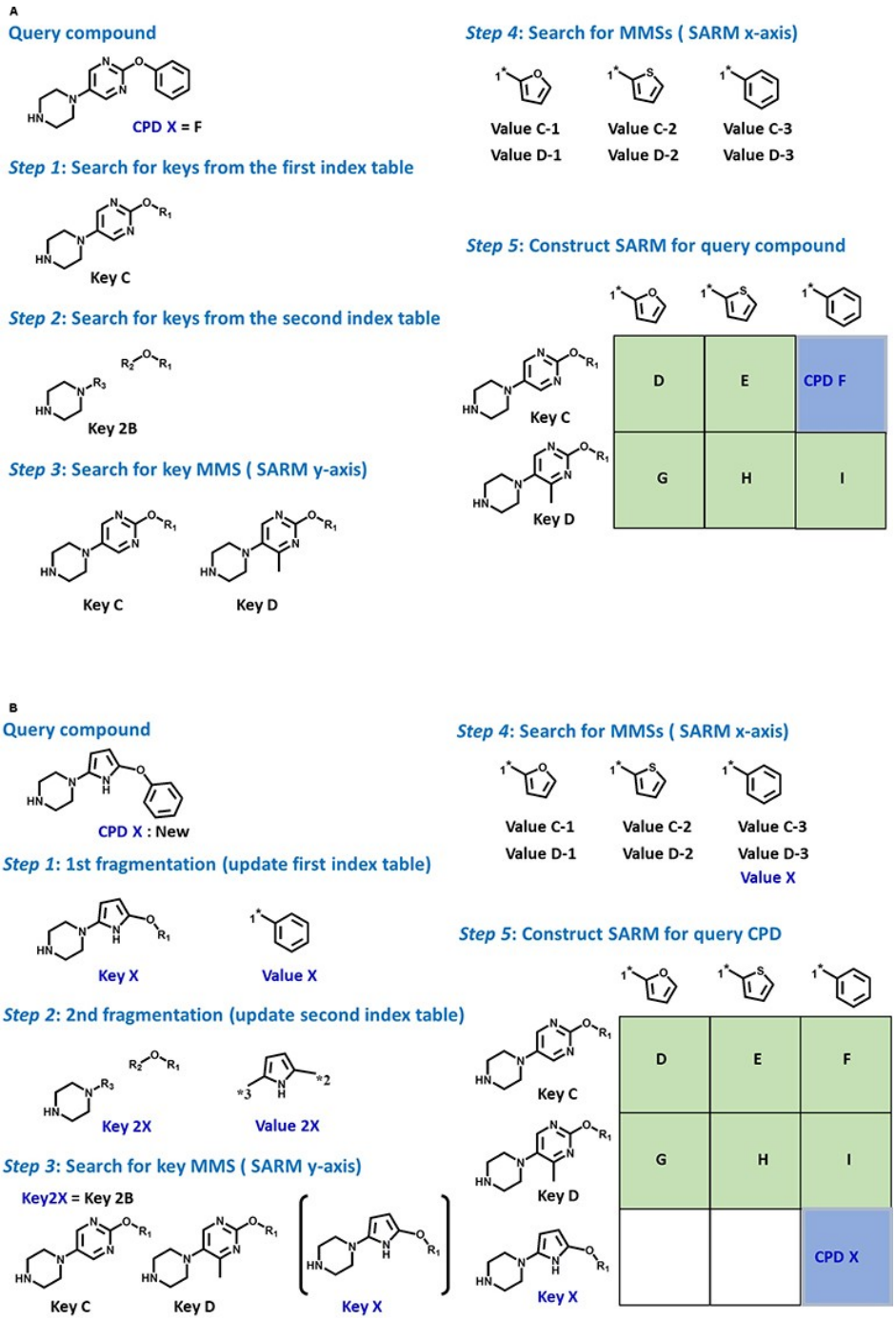

Figure 4. SARM-based analog searching. The search procedure for analogs of a query compound is summarized if the query is (**A**) present in compound data set used to generate the SARM ensemble or (**B**) not present in the data set. The figure was adapted from ref. [16] and modified.

pounds that qualify as structural analogs of query compound F and form a SARM. In the last step, the SARM is assembled. Figure 4B shows the corresponding search procedure if query compound X is not present in the SARMs. In this case, the first two steps of the search routine are replaced by compound fragmentation including the query in order to extend the first index table (yielding key X and value X) and the second index table (key 2X and value 2X). In the third step, the key MMS containing the new key X is extracted from the second index table. Then, MMSs with keys C, D, and X are obtained from the first index

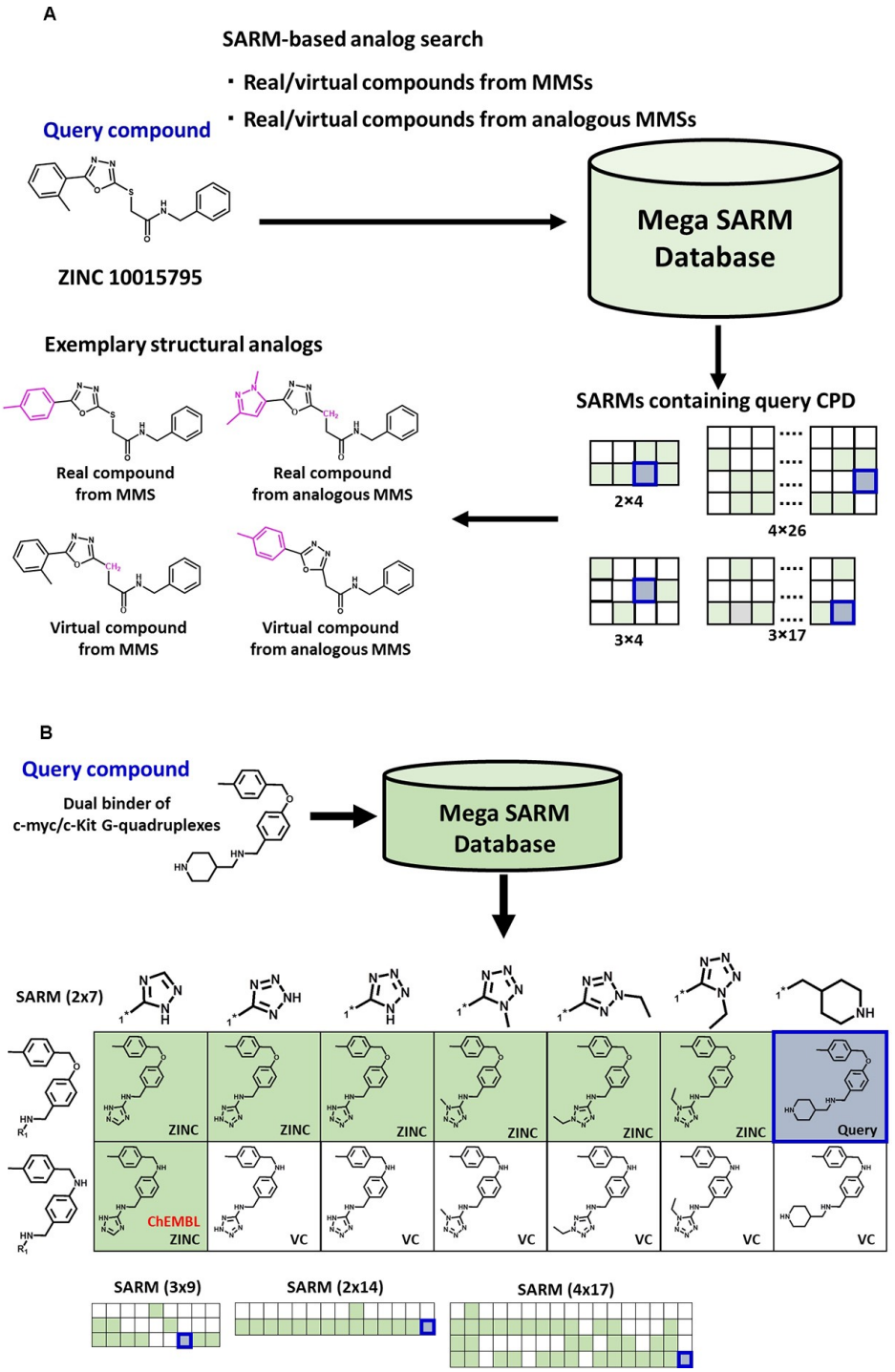

Figure 5. Analog searching in Mega SARM. In (**A**) and (**B**), results of an analog search in the Mega SARM database are summarized for a query compound from ZINC and another from ChEMBL, respectively. In identified SARMs, colored cells represent existing and empty cells virtual compounds. Cells containing the query are highlighted using blue borders. The figure was adapted from ref. [16] and modified.

table. Finally, the SARM of query X is generated, which contains six existing and two virtual analogs of X.

The algorithm also enables analog searching in the Mega SARM database.[16] Figure 5 shows representative examples. Depending on its structural relationships with database compounds, a given query might occur in multiple SARMs. In Figure 5A, a search is summarized for a query from the ZINC subset that was contained in four individual

SARMs from the Mega SARM collection having matrix dimensionality (rows×columns) of 2×4, 3×4, 3×17, and 4×26, respectively. These SARMs provide both existing and virtual analog candidates for further analysis. Figure 5B summarizes the search for a known active compound not contained in ZINC, requiring additional fragmentation. In this case, the query was found in four new SARMs with matrix dimensionality 2×7, 3×9, 2×14, and 4×17, respectively. The 2×7 SARM is shown in detail. This SARM contained seven existing and six virtual analogs of the query. One of the existing analogs was also detected in ChEMBL where it was annotated with multiple biological activities. These findings provided an indication of potential promiscuity of the query compound.

These examples illustrate that SARM-based searching for analogs of query compounds identifies qualifying analogs in varying structural contexts that are organized in SARMs.

Standard SARM-based design of virtual analogs is confined to recombining structural fragments extracted from known compounds. While these fragments and their systematic recombination yield many different virtual analogs, they do not include novel fragments. Hence, the resulting virtual compound populations can be envisioned to form an envelope in chemical space around given analog series. This envelope represents an attractive resource for the selection of close-in analogs for chemical optimization efforts, but it remains narrowly confined. Therefore, we have reasoned that it might be useful to further extend the analog design capacity of the SARM approach by adding novel compounds and fragments to the design cycles, thereby increasing chemical space coverage. For this purpose, deep generative learning[18] provides a suitable methodological framework. Therefore, the DeepSARM approach has been devised, which integrates deep generative learning for analog design into SARM.[19]

Figure 6A illustrates the DeepSARM concept. The basic idea is extending the design of virtual analogs of compounds that are against a given target through learning from other compounds that are active against related targets or the entire protein family. For example, given a set of inhibitors of a kinase of interest, information from inhibitors covering the human kinome can be taken into account while focusing the design on the particular kinase. Accordingly, new compounds with similar activity and their fragments enter analog design, going beyond the use of fragments from the original compound set.

To accomplish this goal, a computational architecture for generative deep learning was designed and implemented comprising three encoder-decoder "generators".[20] Each generator consists of two long short-term memory (LSTM) units.[21] An encoder-decoder generator derives sequence-to-sequence (Seq2Seq) models for transforming one sequence of data into another.[20] The architecture of the entire generative deep learning framework is depicted in Figure 6B. This architecture corresponds to a recurrent neural network for structure generation.[22] For input and output, fragments (keys and values) are represented as vectorized SMILES strings.[22,23] On the basis of the DeepSARM architecture, key and value fragments designated 1 or 2 are generated, which refer to the first and second index table, respectively (Figure 6C). Specifically, the Key 2 Generator (first Seq2Seq model) is trained to generate new key 2 fragments from input key 2 structures. Then, the Value 2 Generator (second model) uses key 2 fragments as input and derives new value 2 fragments. New key 1 fragments are assembled from the generated key 2 and value 2 fragments. These key 1 fragments provide the input for the Value 1 Generator (third model) that produces value 1 fragments. The newly derived key 1 and value 1 fragments are used to build an expanded SARM (where they are labeled key 1 s and value 1 s). Filters between Seq2Seq models prioritize fragments for the subsequent step on the basis of the probability distribution from the decoder, from which fragment-based log_likelihood scores are derived.

To facilitate target-specific expansion of SARMs on the basis of additional compound information, models are trained on fragments from target-specific compounds and then used to predict fragments extracted from compounds active against related targets on the basis of log_likelhood scores. New fragments can expand existing SARMs with new virtual analogs or yield additional SARMs that exclusively consist of new virtual analogs. For example, for a set of 43 inhibitors of Aurora A kinase, 69 SARMs were initially obtained (with dimensionality 2×2 or greater). Analog design was then extended using nearly 28,000 inhibitors covering the human kinome to predict new key and value fragments. These inhibitors were taken from the Kinase SARfari subset of the ChEMBL database.[24] New fragments passing the log_likelihood filters were mapped to existing SARMs. An exemplary SARM expansion is shown in Figure 6C. For a given key 2 fragment from Aurora inhibitors (shown on the upper left) nine value 2 fragments were available to produce key 1 fragments for the SARM (shown on the vertical axis). These value 2 fragments included six structures from Aurora inhibitors (with green numbers 1–6) and three from other kinase inhibitors (blue numbers 7–9). In addition, 14 corresponding value 1 fragments were obtained (horizontal axis) including eight from Aurora (green numbers) and six from other inhibitors (blue numbers). Recombination of the extended key 1 and value 1 sets produced 78 new virtual analogs and more than doubled the size of the original SARM. These new virtual analogs are contained in the area of the SARM in Figure 6C that is enclosed using a blue border. Green cells indicate compounds with most favorable log_likelihood scores including known Aurora inhibitors. As can be seen, a number of new virtual analogs have favorable log_likelihood scores comparable to analogs originating from Aurora inhibitors, especially those located in the upper right region of the SARM. Such new virtual analogs would be

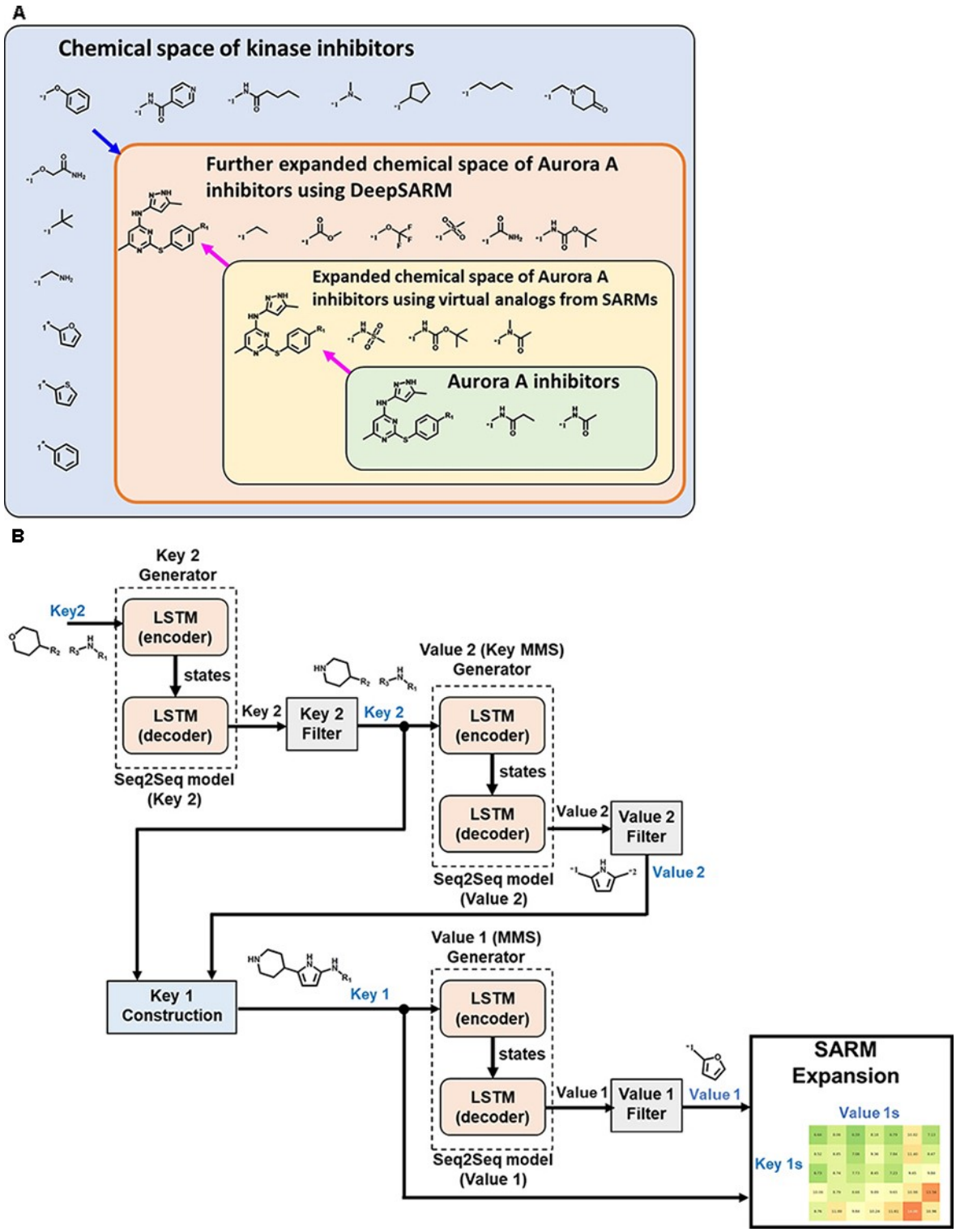

Figure 6. DeepSARM. In (**A**), the principle ideas underlying the DeepSARM approach are illustrated using inhibitors of Aurora A kinase as an example. (**B**) summarizes the computational architecture designed for deep generative models. In (**C**), results of an exemplary application are shown leading to SARM expansion, as discussed in the text. Key and value fragments labeled with green numbers originate from Aurora kinase inhibitors while fragments with blue numbers are from other kinase inhibitors. Cells are color-coded on the basis of inverted log_ likelihood scores (i.e., small scores are preferred). The figure was adapted from ref. [19] and modified.

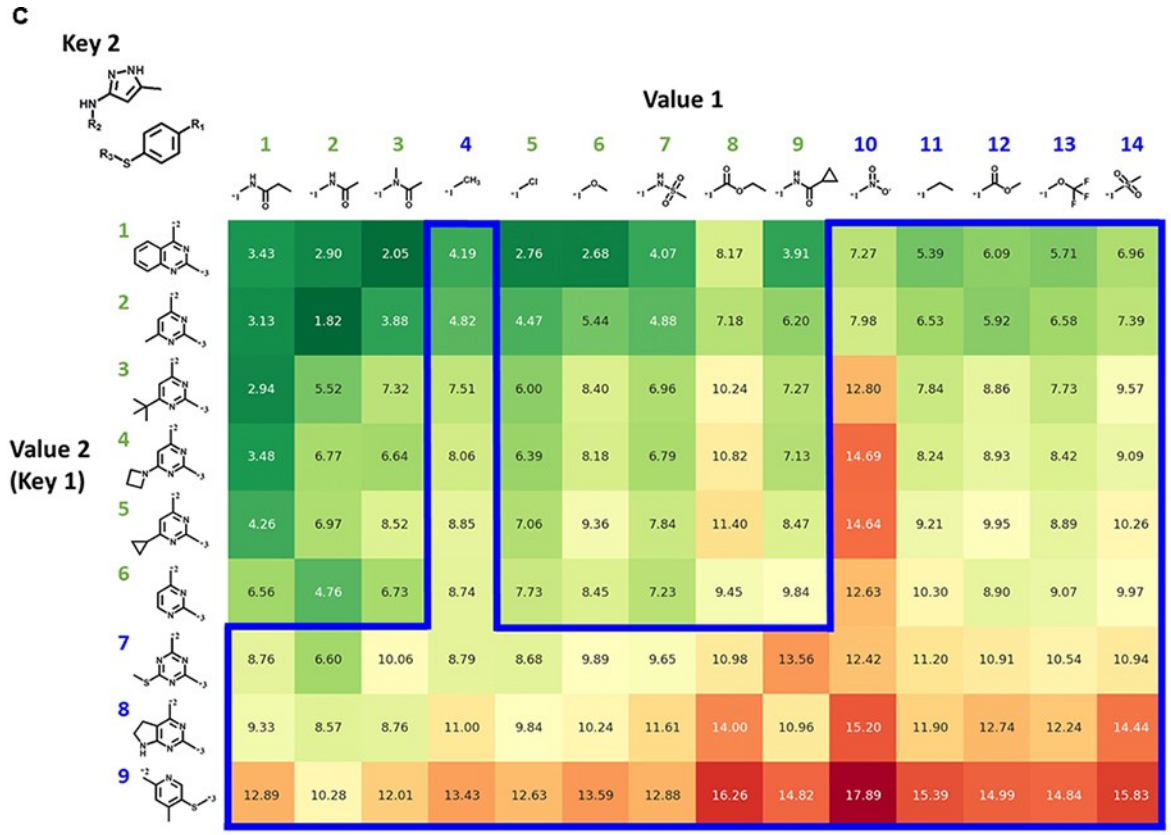

Figure 6. Continued

prioritized as additional candidate compounds for chemical exploration.

In conclusion, we have presented the methodological concept of the SARM approach and its extensions. The discussion mirrors the evolution of the methodology enabling applications with increasing focus on medicinal chemistry. A unique feature of the SARM concept is that it bridges between SAR analysis and compound design. The original implementation included the systematic extraction of structurally related analog series from large compound data sets and their organization in matrices reminiscent of R-group tables and provided a basis for intuitive SAR visualization. Moreover, it also yielded virtual analogs of series organized in SARMs representing unexplored combinations of compound core structures and substituents. These virtual analogs represented candidate compounds for chemical optimization and there was a need to prioritize them. This requirement led to the subsequent development of SARM-based approaches for activity prediction. Moreover, the analysis of SARM ensembles obtained from large data sets was then substantially supported through the introduction of molecular grid maps, which provided a global view of SARM information content and enabled an integrated interactive analysis of interesting regions in a grid map and corresponding SARMs. Another logical extension of the SARM concept has been its large-scale application to generate collections of SARMs as a source of existing and virtual analogs, resulting in the development of Mega SARM. The availability of Mega SARM also motivated the development of a computationally efficient method for SARM-based analog searching. Finally, the interest in further extending SARM-based analog design led to the incorporation of generative deep learning into the SARM framework, hence providing novel structural fragments for SARM-based analog design and expanding chemical space coverage around related series of active analogs. Taken together, these developments were initially driven from a research perspective but have substantially increased the practical utility of the approach. In summary, the evolution of the SARM methodology presented herein is thought to provide an instructive example of iterative methodological developments originating from chemoinformatics with increasing potential for practical applications in medicinal chemistry.

Conflict of Interest

A. Yoshimori is the CEO of the Institute for Theoretical Medicine, Inc. (ITM) that develops scientific software for drug discovery. J. Bajorath is a consultant to ITM.

Acknowledgement

The authors thank Dr. Hiroyuki Kouji for helpful discussions. In addition, the authors are grateful to former members of the Bajorath research group for their contributions to the development of the SARM approach. Furthermore, the authors thank Toru Tanoue and Yuichi Horita for valuable contributions to the development of the Molecular Grid Map and Mega SARM. Open Access funding enabled and organized by Projekt DEAL.

References

[1] A. M. Wassermann, P. Haebel, N. Weskamp, J. Bajorath, *J. Chem. Inf. Model.* **2012**, *52*, 1769–1776.
[2] D. Gupta-Ostermann, J. Bajorath, *F1000Research* **2014**, *3*, e113.
[3] J. Hussain, C. Rea, *J. Chem. Inf. Model.* **2010**, *50*, 339–348.
[4] M. Wawer, J. Bajorath, *J. Med. Chem.*, **2011**, *54*, 2944–2951.
[5] V. Shanmugasundaram, L. Zhang, S. Kayastha, A. de la Vega de León, D. Dimova, J. Bajorath, *J. Med. Chem.* **2016**, *59*, 4235–4244.
[6] D. Gupta-Ostermann, V. Shanmugasundaram, J. Bajorath, *J. Chem. Inf. Model.* **2014**, *54*, 801–809.
[7] S. M. Free, J. W. Wilson, *J. Med. Chem.* **1964**, *7*, 395–399.
[8] D. Gupta-Ostermann, J. Balfer, J. Bajorath, *Mol. Inf.* **2015**, *34*, 134–146.
[9] D. Gupta-Ostermann, Y. Hirose, T. Odagami, H. Kouji, J. Bajorath, *F1000Research* **2015**, *4*, e75.
[10] A. Yoshimori, T. Tanoue, J. Bajorath, *ACS Omega* **2019**, *4*, 7061–7069.
[11] L. Van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
[12] C. M. Bishop, M. Svensen, C. K. I. Williams, *Neural Comput.* **1998**, *10*, 215–234.
[13] R. Jonker, A. Volgenant, *Computing* **1987**, *38*, 325–340.
[14] L. Peltason, P. Iyer, J. Bajorath, *J. Chem. Inf. Model.* **2010**, *50*, 1021–1033.
[15] T. Sterling, J. J. Irwin, *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
[16] A. Yoshimori, Y. Horita, T. Tanaoue, J. Bajorath J, *J. Chem. Inf. Model.* **2019**, *59*, 3727–3734.
[17] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
[18] A. C. Mater, M. L. Coote, *J. Chem. Inf. Model.* **2019**, *59*, 2545–2559.
[19] A. Yoshimori, J. Bajorath, *Future Drug Discov.* **2020**, *2*, in press.
[20] I. Sutskever, O. Vinyals, Q. V. Le, *Advances in Neural Information Processing Systems 27 (NIPS 2014)* **2014**, 3104–3112..
[21] S. Hochreiter, J. Schmidhuber, *Neur. Comput.* **1997**, *9*, 1735–1780.
[22] S. Zheng, X. Yan, Q. Gu, Y. Yang, Y. Du, Y. Lu, J. Xu, *J. Cheminf.* **2019**, *11*, e5.
[23] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
[24] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J. P. Overington, *Nucleic Acids Res.* **2014**, *42*, D1083-D1090.

Received: March 16, 2020
Accepted: April 9, 2020
Published online on April 20, 2020

