Methodology

*For reprint orders, please contact: reprints@future-science.com*

Future Drug Discovery

# Deep SAR matrix: SAR matrix expansion for advanced analog design using deep learning architectures

Atsushi Yoshimori[1] & Jürgen Bajorath*,[2]

[1]Institute for Theoretical Medicine, Inc., 26-1 Muraoka-Higashi 2-chome, Fujisawa, Kanagawa, 251-0012, Japan

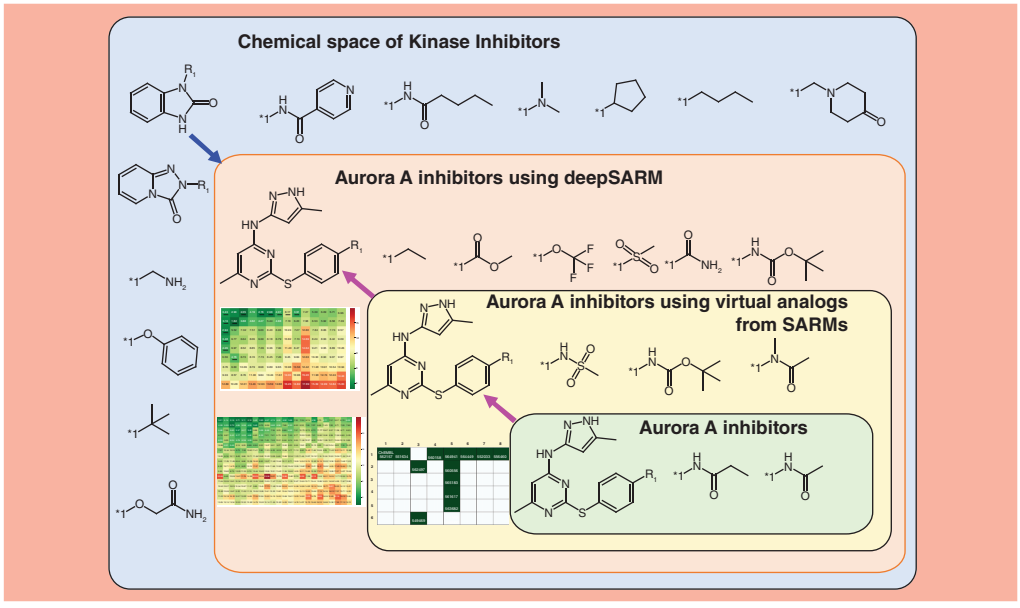
[2]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53113, Bonn, Germany

*Author for correspondence: Tel.: +49 228 7369100; Fax: +49 228 7369 101; bajorath@bit.uni-bonn.de

**Aim:** Enhancing the structure–activity relationship matrix (SARM) methodology through integration of deep learning and expansion of chemical space coverage. **Background:** Analog design is of critical importance for medicinal chemistry. The SARM approach, which combines systematic structural organization of compound series with analog design, is put into scientific context. **Methodology:** The new DeepSARM concept is introduced. The architecture of SARM-integrated deep generative models is detailed and the workflow for advanced analog design and matrix expansion described. **Exemplary application:** The DeepSARM approach is applied to design analogs of kinase inhibitors taking kinome-wide chemical space into account. **Future perspective:** Practical applications of DeepSARM will be a major focal point. Different applications are discussed. New computational features will be added to prioritize virtual candidate compounds.

**Lay abstract:** Compound optimization in medicinal chemistry depends on the design of new candidate molecules to improve biological activity and chemical properties. These molecules are usually structurally closely related to current leads and hence termed as analogs. Computational methods can support the design of analogs. The structure–activity relationship matrix (SARM) method systematically organizes series of compounds and suggests new analogs. The SARM approach has been extended through integration of machine learning. The resulting DeepSARM method further increases the coverage of biologically relevant chemical space with novel analogs.

**Graphical abstract:**




newlands press

First draft submitted: 2 February 2020; Accepted for publication: 24 February 2020; Published online: 30 March 2020

**Keywords:** analog design • compound optimization • computational medicinal chemistry • deep generative learning • DeepSARM concept • novel virtual analogs • SAR matrix method • SARM expansion

Compound optimization and development in medicinal chemistry relies on the generation of analogs of molecules having desired biological activity and other attractive features [1,2]. During hit-to-lead transformation and lead optimization (LO), medicinal chemists concentrate on the question which analogs to make next in order to explore and further evolve structure–activity relationships (SARs), improve compound potency and optimize other LO relevant molecular properties. LO efforts continue to evolve until compound series eventually hit insuperable roadblocks and must be abandoned or until the stage of clinical candidates is ultimately reached. In the practice of medicinal chemistry, analogs are mostly generated on the basis of chemical experience and intuition. This process is often perceived more as a form of science-driven art than a rigorous scientific exercise [2]. Given the central relevance of analog design for medicinal chemistry, it is not surprising that approaches capable of guiding and rationalizing analog generation are highly sought after. This explains the interest in computational methods for analog searching and design. In addition to conventional quantitative SAR methods for compound potency prediction, popular approaches for analog generation include scaffold and fragment searching, matched molecular pair (MMP) analysis, screening of large virtual libraries and computational analog design [3–10].

Among approaches for computational analog design, the SAR matrix (SARM) methodology is unique in that it combines the systematic extraction of analog series from compound collections with the generation of new virtual candidate compounds [10]. The SARM approach organizes related compound series in multiple matrices that are akin to, but conceptually distinct from standard R-group tables. The generation of these SARMs is based upon a systematic two-level fragmentation scheme to extract all core structures and substituents from available compounds, identify analog series and organize them into subsets of structurally related ones. Each individual SARM contains such a subset. In addition to existing analogs comprising a series, SARMs also contain all possible nonexplored core–substituent combinations, which represent virtual analogs for further exploration. These candidate compounds map analog space around related series on the basis of existing structural fragments that are systematically recombined. Hence, virtual analogs from SARMs can be envisioned to form an envelope around a given series in chemical space. The SARM approach has been successfully used in hit-to-lead projects to identify new active analogs and other practical applications [11,12]. To aid in prioritizing and selecting candidate compounds, the methodology has been extended through incorporation of local matrix-based quantitative SAR modeling [13], addition of molecular grid maps for integrated SARM display [14] and implementation of systematic analog searching in very large databases of SARMs [15].

A characteristic feature of the SARM approach is that the newly generated analog space around a given series is confined to recombination of existing structural fragments originating from the series. Thus, as a further improvement of the methodology, accessible analog space might be extended by generating candidate compounds with novel structural features. For example, this might be attempted by taking information from compounds active against related targets into account to obtain new structural fragments. While a systematic exploration of such information principally represents a nontrivial task, deep learning approaches, which are becoming increasingly popular in drug discovery and chemistry [16,17], offer potential solutions. Herein, we report a new concept for analog generation and the design, implementation and application of dedicated deep generative learning architectures to further increase the SAR information content and chemical space coverage of SARM-based analogs by taking compound information from target families into account.

## Methodology

### Methodological concept

DeepSARM was designed to expand the chemical space coverage of original SARMs using deep learning models for compound design. The concept of DeepSARM is illustrated in Figure 1, using kinase inhibitors as an example.

The key aspect of the approach is complementing analog design for a given set of active compounds (e.g., inhibitors of Aurora A kinase) through learning from compound information for related targets (e.g., inhibitors covering the human kinome), followed by fine-tuning of the design for the target of interest.

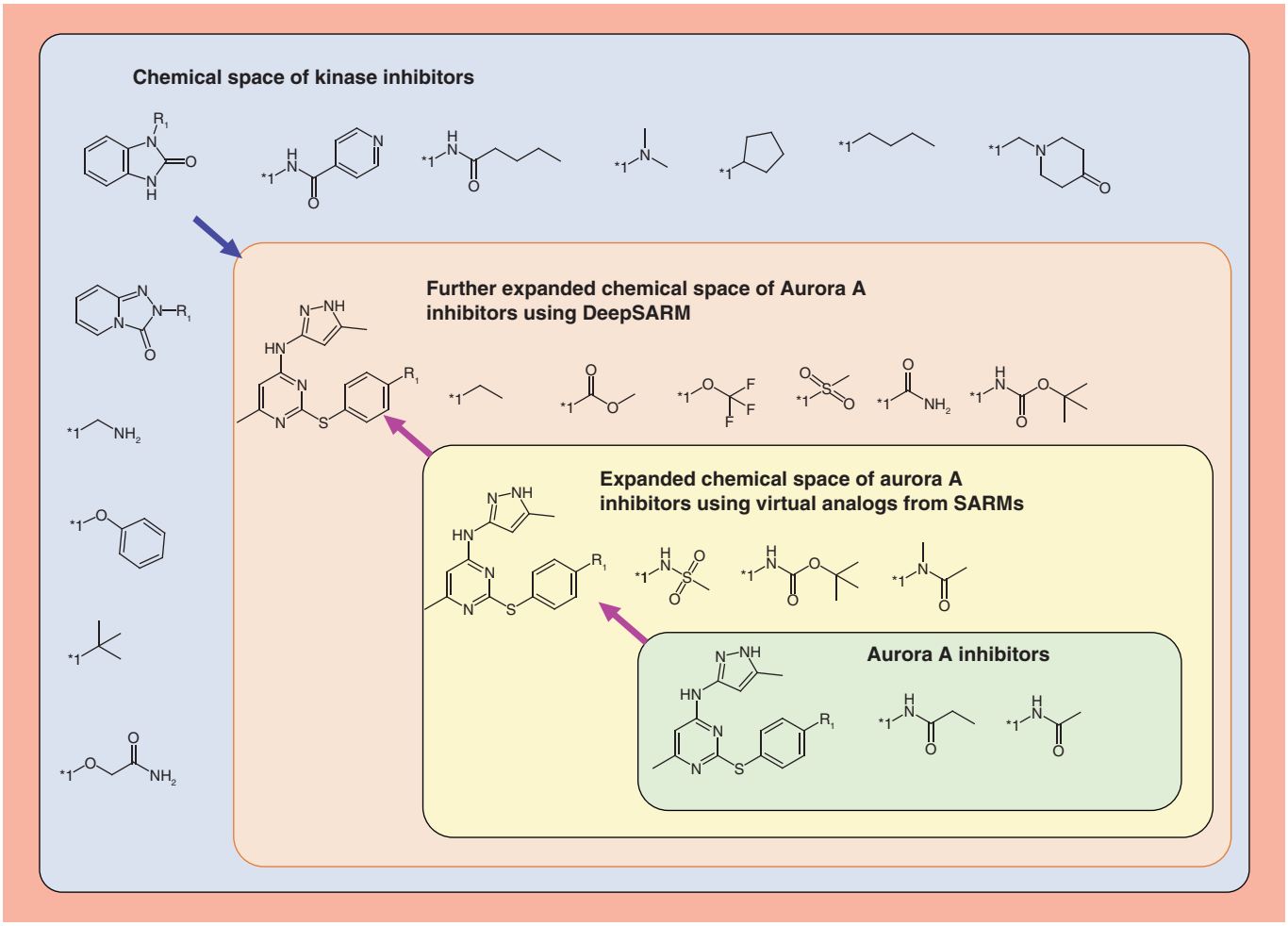



**Figure 1. Methodological concept of DeepSAR matrix.** SARMs are expanded through the use of deep generative models taking chemical space information from target families into account. As an example, the SARM method is applied to structurally organize Aurora A kinase inhibitors and generate close-in virtual analogs. In this case, the Kinase SARfari database is used to represent chemical space of kinase inhibitors covering the human kinome. DeepSARM generates additional candidate inhibitors through deep generative models taking information from SARMs and Kinase SARfari into account.
SARM: Structure–activity relationship matrix.

Through generative learning using deep molecular encoder–decoder models, expanded SARMs are constructed, which contain not only fragments from original SARMs, but also novel fragments obtained by target-directed learning from related compound information. Importantly, the newly derived predictive models enable learning from such compound information with a focus on individual targets and compatible structural fragments, as detailed below.

Hence, the DeepSARM concept can be rationalized as a data-driven expansion of confined analog space around compound series of interest.

## Construction of SARM index tables

The methodology for constructing SARMs has been detailed previously [10,11,15] and is briefly summarized here using a small set of nine exemplary molecules, shown in Figure 2A. SARMs are generated through systematic two-level MMP fragmentation [6,7]. An MMP is defined as a pair of compounds that only differ by a chemical change at a single site [6]. Two-level fragmentation is unique to the SARM approach. In the first step, all compounds are subjected to MMP fragmentation of exocyclic single bonds by application of the Hussain–Rea algorithm [7], which produces key (core structure) and smaller value (substituent, R-group) fragments. A first index table is generated from key and value fragments obtained from original compounds, as shown in Figure 2B. All compounds that are

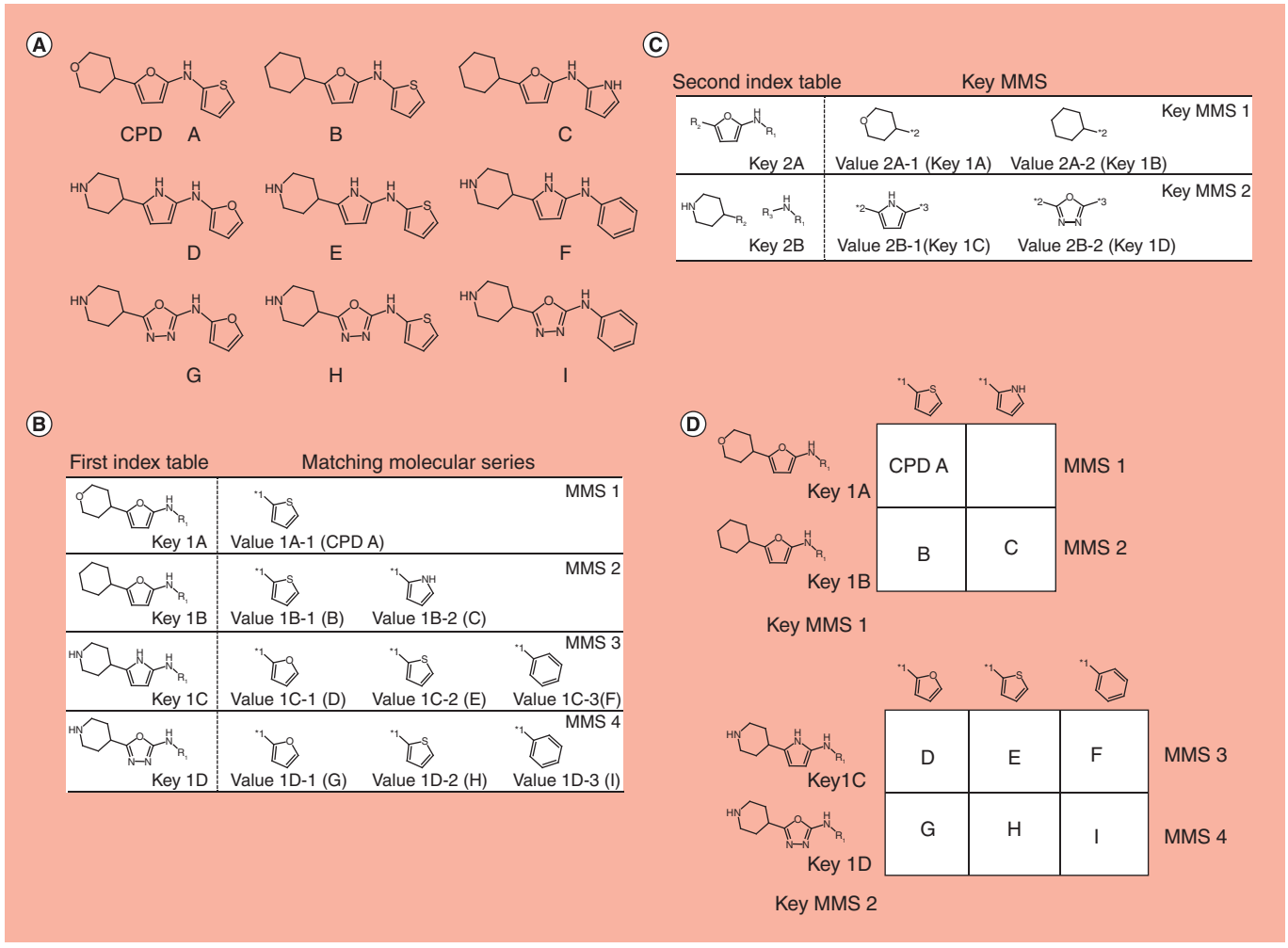



**Figure 2. Structure–activity relationship matrix data structure. (A)** Shown is a small model dataset for structure–activity relationship matrix (SARM) generation with nine compounds (CPD A–I). **(B)** Compound fragmentation through systematic deletion of exocyclic single bonds yields two types of fragments that are stored in the first index table. These fragments are termed key and value, respectively, following matched molecular pair terminology. Compounds containing the same key and different values form a MMS. **(C)** Structurally analogous cores are identified by subjecting all keys from the first index table to a second round of systematic bond fragmentation. Sets of structurally analogous cores form key MMSs and are stored in the second index table. **(D)** For each key MMS, an SARM is constructed by placing the keys in the leftmost column (y-axis of the SARM) and each value (from the first index table) from compounds forming the associated MMS in the corresponding row (x-axis). The empty cell in the SARM at the top represents a virtual analog (i.e., a currently unexplored key–value combination).
Adapted with permission from [15].
MMS: Matching molecular series.

associated with the same key fragment in the first index table form a matching molecular series (MMS), which is defined as a series of compounds that only differ by structural changes at a single site [18], hence representing an extension of the MMP concept.

Key fragments from the first index table are then subjected to a second fragmentation process. As illustrated in Figure 2C, a second index table is generated with the resulting fragments. All key fragments in the second index table represent core structures that are only distinguished by a modification at a single site. These structurally analogous cores form a so-called key MMS. Analog series (i.e., rows in the second index table) containing structurally related cores are then organized in individual SARMs, as shown in Figure 2D. In an SARM, each row contains an individual analog series and each column compounds from different series sharing the same substituent. Hence, each cell in an SARM represents a unique compound and an empty cell represents a virtual analog, in other words, a currently not yet explored key–value combination. Accordingly, SARMs comprehensively extract structural relationships from

compound datasets, organize compounds into structurally related series and generate virtual analogs from possible combinations of core structures and substituents, thereby expanding chemical space around a given set with newly designed compounds.

## Molecular representations

All compound structures were converted into canonical simplified molecular input line entry specification (SMILES) strings [19] using RDKit software [20]. For generative learning, SMILES are vectorized to one-hot encoded representations [21]. Initially, SMILES are tokenized based on a single character or multiple characters. The single character indicates atom types (e.g., 'C', 'c'), bond types (e.g., '-', '=' and '#') and ring closures (e.g., '1' and '2'). Two characters denote other atom types (e.g., 'Cl' and 'Br') and special environments are defined using square brackets (e.g., '[nH] and [*:1]'). Furthermore, start and stop tokens ('[START]' and '[END]', respectively) are added. Following tokenization, SMILES are transformed into one-hot encoded (binarized) representations.

## Generative models

### *Architecture*

Sequence-to-sequence (Seq2Seq) models represent a general-purpose encoder–decoder framework to translate one data sequence into another [22]. These models have been successfully applied in areas such as machine translation, text processing or image analysis [22]. Figure 3A illustrates a Seq2Seq encoder–decoder architecture that consists of two long short-term memory (LSTM) units [23]. This architecture represents a deep recurrent neural network [21] and is used to generate different Seq2Seq models. The encoder LSTM transforms input sequences into two-state vectors (h, c), and the decoder LSTM is trained to return the same sequences as target sequences on the basis of transformed SMILES. As an initial state, the decoder uses two state vectors from the encoder. The latent dimensionality of the encoding space of the LSTM is set here to 256.

### *Model derivation*

Key and value designations for fragments used in the following are standard terminology of the MMP [7] and SARM [10] methodologies.

To generate expanded SARMs, three Seq2Seq models are trained as follows:

Model (key 2) using key 2 (input)/key 2 (target) pairs (with the same SMILES);
Model (value 2) using key 2 (input)/value 2 (target) pairs (from key MMSs);
Model (value 1) using key 1 (input)/value 1 (target) pairs (from MMSs). Training Seq2Seq models for the generation of expanded SARMs consists of the following steps that are summarized in Figure 3B: pretraining using large numbers of structures from a given target family or class, construction of the first and second index tables and fine-tuning of the model on the basis of compounds active against a target of interest from the family or class. Retraining involves the adjustment of internal model weights.

For training, the number of epochs is set to 5, batch size is set to 64 and the compound datasets are divided into training and validation sets (9:1).

Scripts for model derivation were written in Python (version 3.7.3), and the Seq2Seq models were built with Keras (version 2.2.4) [24].

### *Fragment sampling*

The trained Seq2Seq models generate fragments for key 2, value 2 and value 1 on the basis of SMILES strings. The fragment sampling procedure consists of the following steps: a fragment is submitted to the encoder and the resulting state vectors (h, c) are sent to the decoder, which uses them as initial states for the '[START]' token. The process is repeated until the decoder replies with an '[END]' token. Multinomial sampling on the probability distribution is applied and rescaled using temperature factors for each token as available in Keras [24]. Temperature factors for the key 2, value 2 and value 1 generator are set to 1.5, 1.5 and 1.2, respectively.

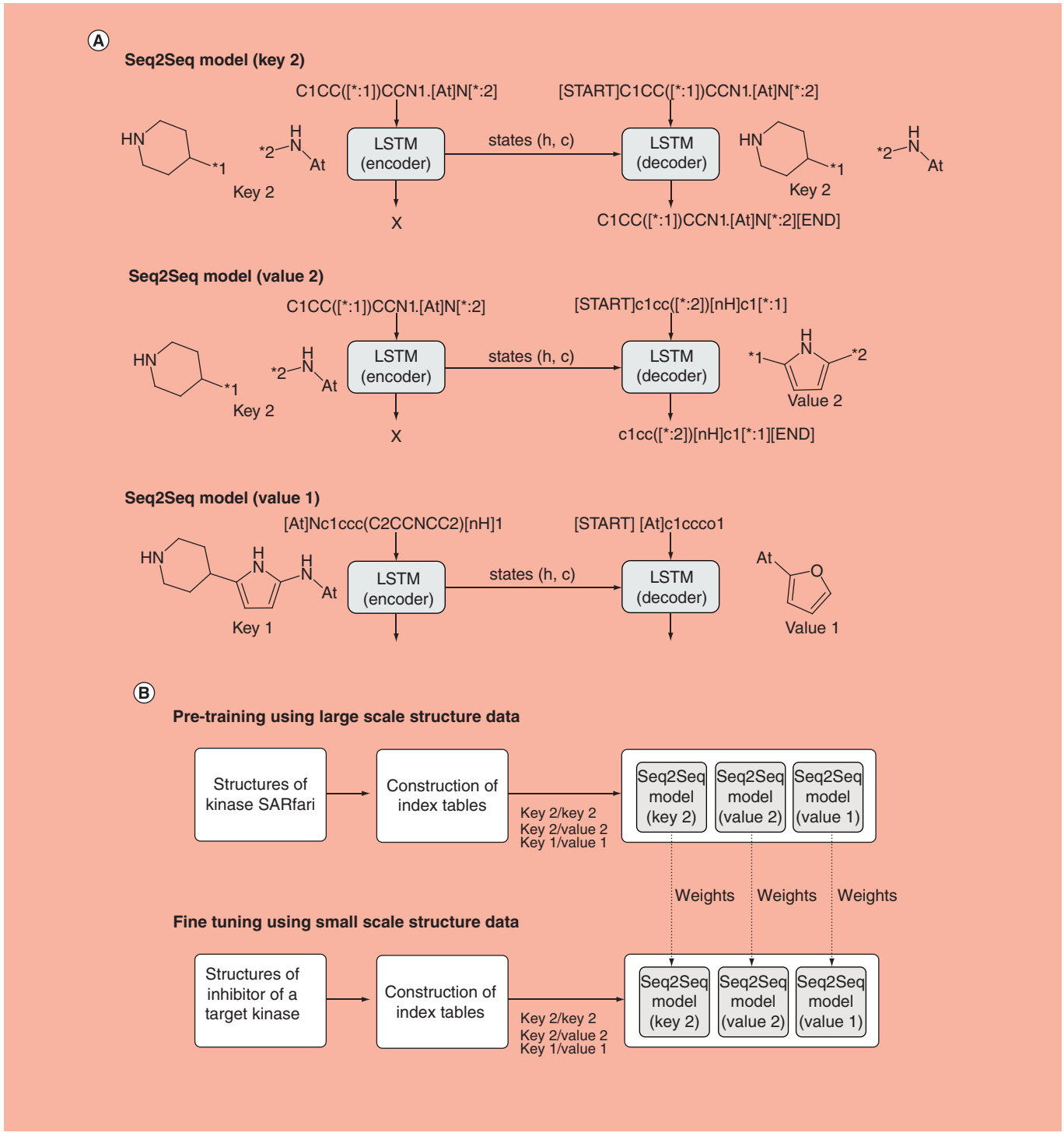



**Figure 3. Model derivation and DeepSARM workflow. (A)** Three sequence-to-sequence (Seq2Seq) models consisting of encoder–decoder long short-term memory units are derived. The Seq2Seq model for key 2 (i.e., the key 2 generator) is trained using input key 2/output key 2 pairs (with the same SMILES), the model for value 2 using key 2/value 2 pairs (from key matching molecular series; MMS), and the model for value 1 using key 1/value 1 pairs (from MMSs). Attachment points are indicated '*1', '*2' and 'At'. '*1', '*2' in input SMILES correspond to '*1', '*2' in output SMILES. **(B)** Seq2Seq models are first trained using fragments from large numbers of structures of a given target family or class (pretraining step) and then fine-tuned using corresponding fragments from compounds active against an individual target from that class. During retraining transferred model weights are further adjusted. **(C)** A new compound (CPD D) is assembled from fragments. Each Seq2Seq model generates a variety of fragments for structure–activity relationship matrix expansion that are prioritized on the basis of log_likelihood scores. Only one example is shown here. **(D)** Compounds comprising newly generated key fragments are added and systematically organized in the expanded structure–activity relationship matrix.
LSTM: Long short-term memory.

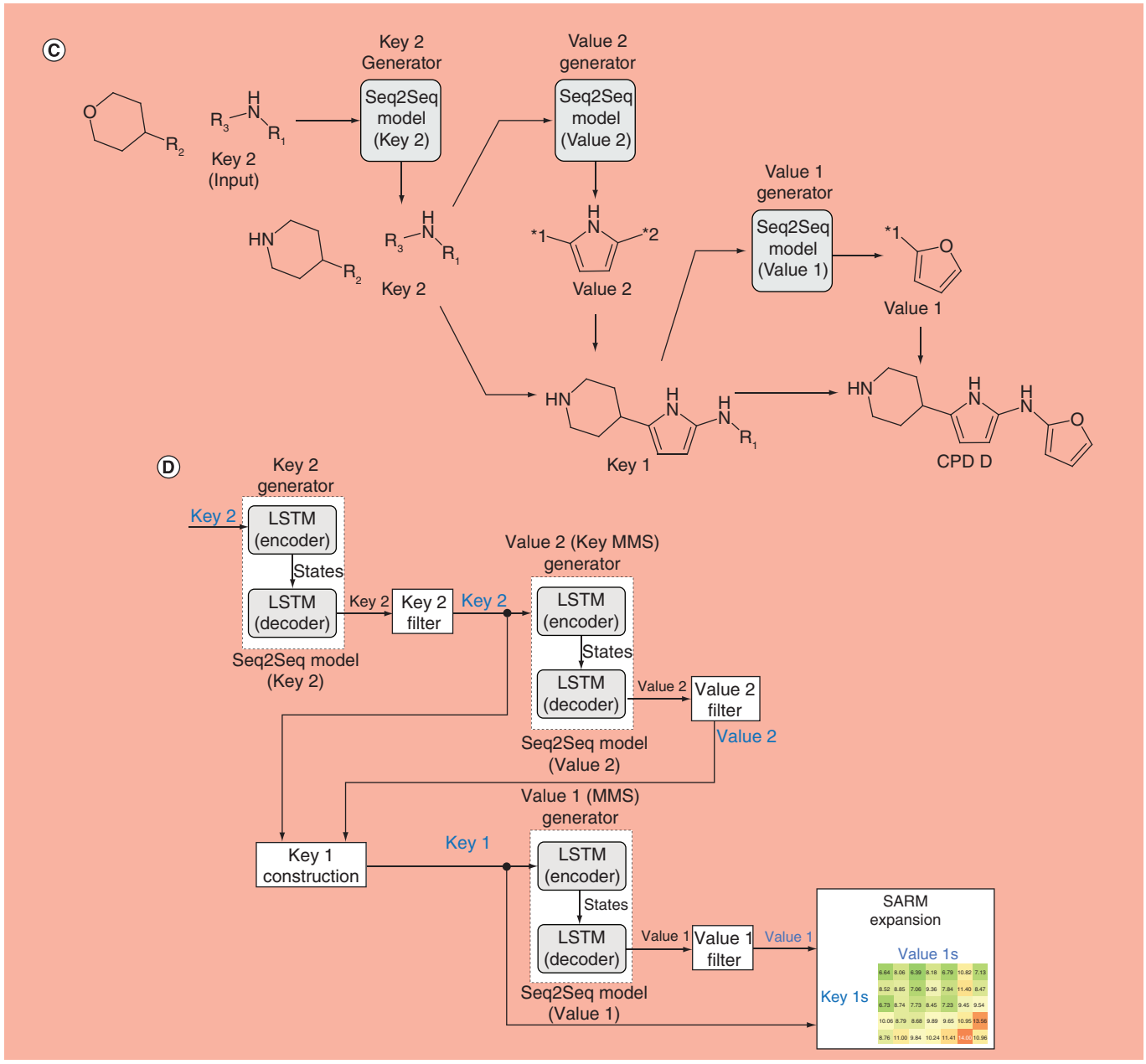



**Figure 3. Model derivation and DeepSARM workflow (cont.). (A)** Three sequence-to-sequence (Seq2Seq) models consisting of encoder–decoder long short-term memory units are derived. The Seq2Seq model for key 2 (i.e., the key 2 generator) is trained using input key 2/output key 2 pairs (with the same SMILES), the model for value 2 using key 2/value 2 pairs (from key matching molecular series; MMS), and the model for value 1 using key 1/value 1 pairs (from MMSs). Attachment points are indicated '*1', '*2' and 'At'. '*1', '*2' in input SMILES correspond to '*1', '*2' in output SMILES. **(B)** Seq2Seq models are first trained using fragments from large numbers of structures of a given target family or class (pretraining step) and then fine-tuned using corresponding fragments from compounds active against an individual target from that class. During retraining transferred model weights are further adjusted. **(C)** A new compound (CPD D) is assembled from fragments. Each Seq2Seq model generates a variety of fragments for structure–activity relationship matrix expansion that are prioritized on the basis of log_likelihood scores. Only one example is shown here. **(D)** Compounds comprising newly generated key fragments are added and systematically organized in the expanded structure–activity relationship matrix.
LSTM: Long short-term memory.

*Fragment scoring*

Generated fragments are evaluated on the basis of a log_likelihood score defined as:

$$\log\_likelihood\ score = -\sum_{t=1}^{T} \log P\left(x^t | x^{t-1}, \ \ldots, \ x^1\right)$$

where $P$ is a probability distribution of a decoder in the model and $T$ the number of tokens for a fragment. The score is used as a threshold value for filtering new key 2, value 2 and value 1 fragments. The minus sign effectuates that high probabilities yield small scores for fragment prioritization.

*Compound generation*

Figure 3C illustrates how new compounds are generated using sampled fragments. Key 2 is generated from input key 2 fragments using the first Seq2Seq model (key 2 generator). Value 2 is generated from key 2 using the second model (value 2 generator). Then, key 1 is assembled from key 2 and value 2. Value 1 is generated from key 1 using the third Seq2Seq model (value 1 generator). Finally, a new compound D is obtained by combining key 1 and value 1.

### SARM expansion

The workflow for expansion of SARMs is summarized in Figure 3D. New fragments are derived and filtered by log_likelihood score (applying a threshold value of 10 for key 2 and value 2, respectively, and a value of 5 for value 1). The log_likelihood score of a new compound is obtained by adding the individual scores of its fragments. When compounds are added to an SARM, unique key 1 fragments are placed on the vertical axis and value 1 fragment of the horizontal axis, resulting in new key–value combinations across the expanded SARM and new virtual analogs.

We note that the primary application scenario of the DeepSARM approach is the expansion of structurally related analog series. The approach is readily scalable to series of increasing size as well as larger datasets. It is also applicable, for example, in the context of Mega SARM [15].

## Exemplary application

### Compound data

As an exemplary application, the DeepSARM methodology was applied to expand SARMs of human Aurora A kinase inhibitors. Therefore, a set of 43 Aurora A kinase inhibitors (in the following referred to as Aurora inhibitors) was taken from ChEMBL (id: CHEMBL1158437) [25]. In addition, inhibitors of other human kinases and their activity data were obtained from the public Kinase SARfari collection [26]. Compound entries were filtered to remove salts and other auxiliary molecules and select structures with molecular weight of 350–500 Da, resulting in 27,778 unique kinase inhibitors.

### SARMs of Aurora inhibitors

For the set of 43 Aurora inhibitors, an ensemble of 69 SARMs with dimensionality equal to or greater than 2 × 2 were obtained. Figure 4 illustrates the generation of one of these SARM. Figure 4A shows exemplary inhibitors and Figure 4B shows value 2 and value 1 fragments from the set of Aurora inhibitors for a given a key 2. Figure 4C shows the resulting SARM that contains 13 of 43 Aurora inhibitors (green cells) and 35 virtual analogs (empty cells).

### Expanded SARMs

Generation of new fragments using DeepSARM can complement existing SARMs with additional compounds or result in new SARMs that exclusively consist of new virtual analogs, depending on the structural relationships that are formed between compounds from original SARMs and newly generated molecules.

*Combining existing & new fragments*

Figure 5 provides an example for DeepSARM expansion leading to SARMs incorporating new fragments. Figure 5A shows key 2 structures produced by the key 2 generator on the basis of the input key 2 fragments from Figure 4B. The key 2 generator yielded 90 valid SMILES in 100 random trials. After removing duplicated SMILES, 13 unique

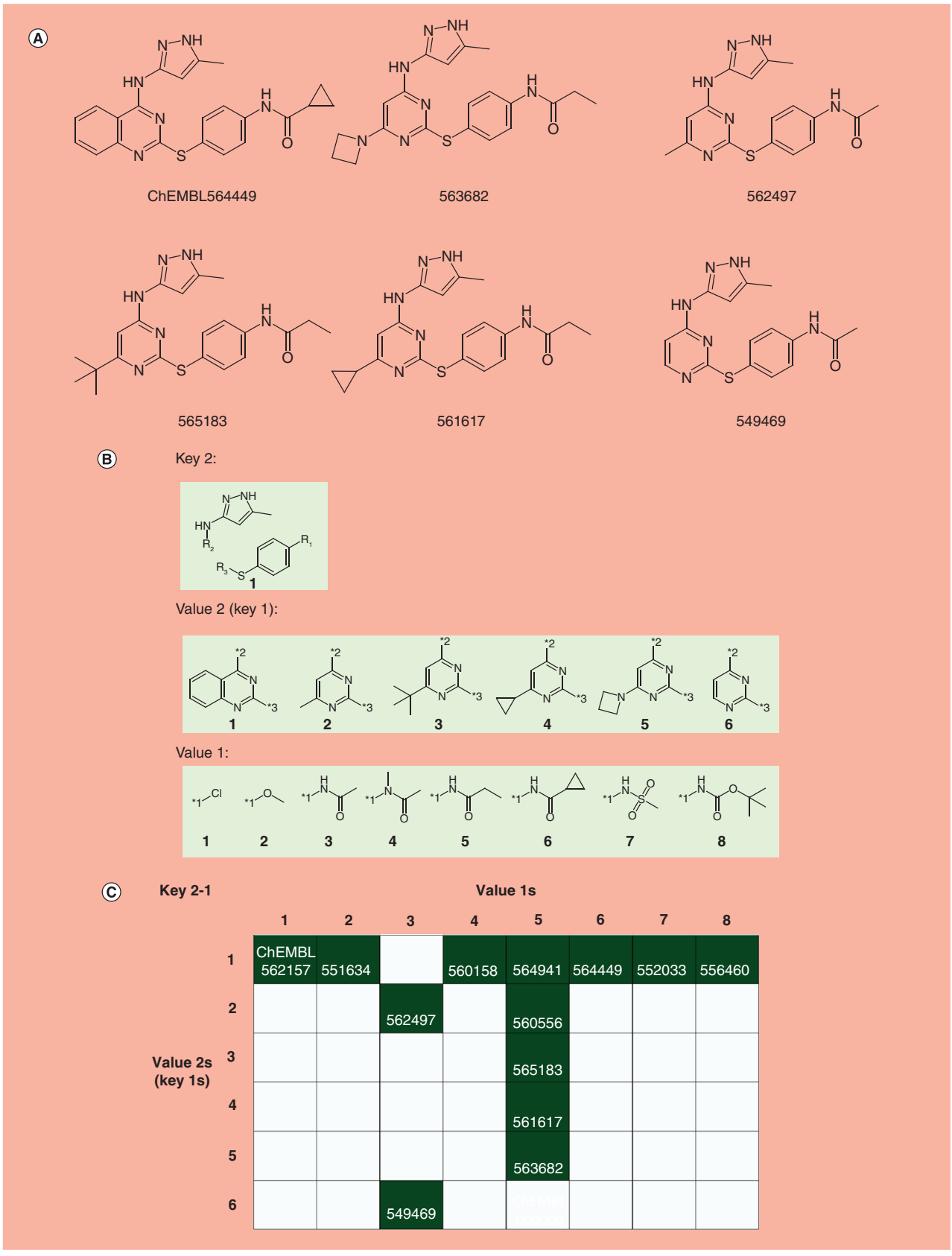



**Figure 4. Exemplary structure–activity relationship matrix. (A)** Six exemplary compounds from a set of 43 Aurora A kinase inhibitors used to generate structure–activity relationship matrices (SARMs) are shown. **(B)** Exemplary key 2, value 2 and value 1 fragments are depicted. 'Value 2 (key 1)' means 'value 2 for key 1 generation' and indicates that key 1 fragments are assembled from key 2 and value 2 fragments. Numbers are fragment identifiers (ID). **(C)** An exemplary SARM is assembled from fragments in **(B)** that contains 13 known inhibitors. 'Key 2-1' refers to the key 2 fragment with ID 1. Fragments contained in SARMs are labeled 's' (e.g., value 1s).

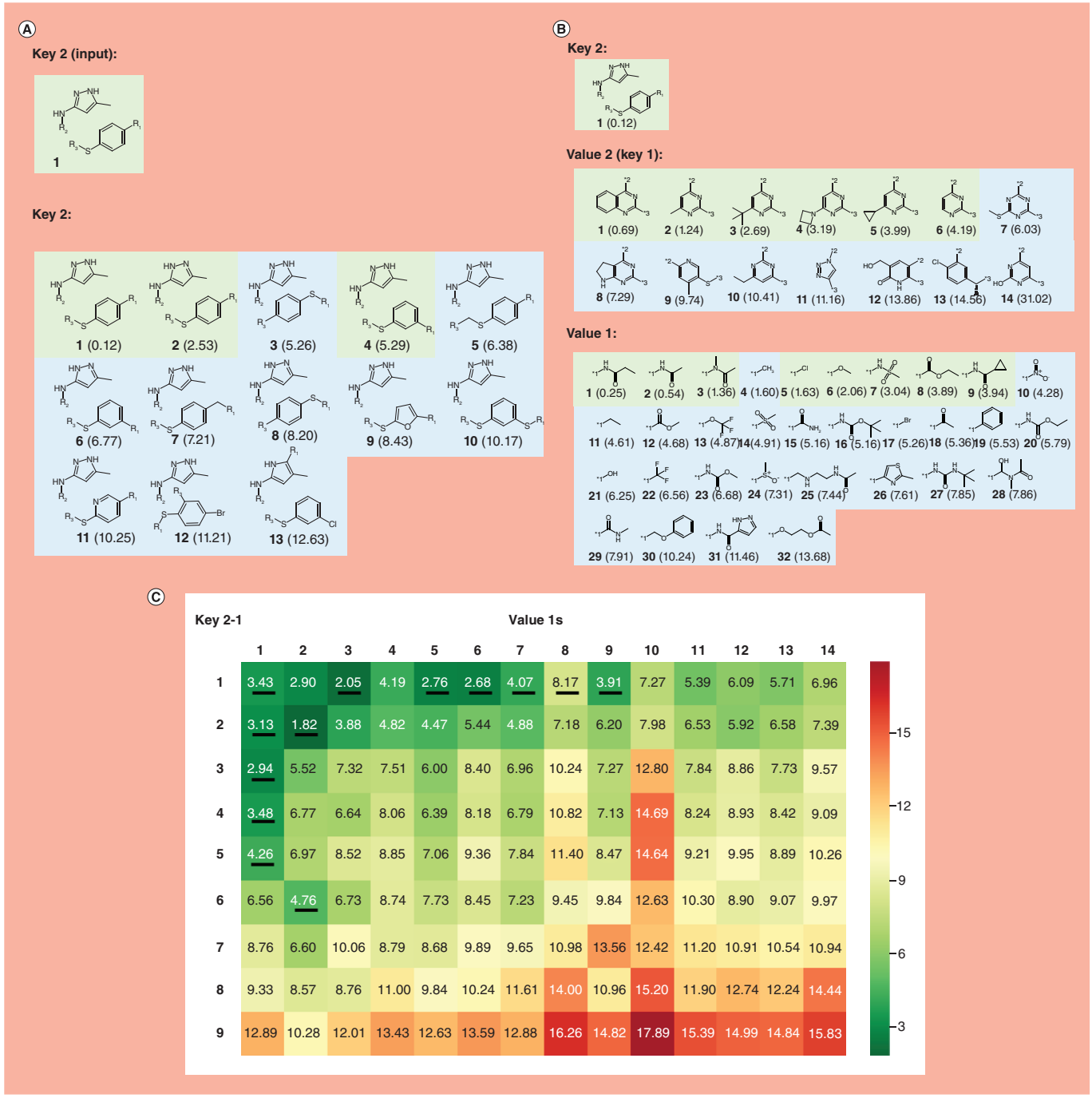



**Figure 5. Exemplary DeepSARM expansion. (A)** Shows key 2 fragments constructed with the key 2 generator based on input key 2-1 from Figure 4B. Key 2 structures on a green background are contained in original structure–activity relationship matrices (SARMs) of the set of Aurora A kinase inhibitors while key 2 structures on a blue background are novel. Numbers are fragment identifiers (ID) and numbers in parentheses report log_likelihood scores. **(B)** Shown are value 2 and value 1 fragments generated with DeepSARM on the basis of the SARM in Figure 4C. **(C)** The expanded SARM is displayed. Each cell in the SARM represents a unique compound and is color-coded by log_likelihood scores.

structures were obtained, including key 2 fragments from the original SARM as well as novel key 2 structures. Figure 5B shows value 2 and value 1 fragments generated using DeepSARM based on the SARM in Figure 4C. The value 2 generator produced 97 valid SMILES in 100 random trials, including 14 unique structures. In addition, 14 key 1 fragments were obtained from the 14 value 2 fragments and key 2-1. The value 1 generator then produced 32

unique structures in 10 random trials with each key 1. Figure 5C shows the expanded SARM. The SARM is color coded by log_likelihood scores. Cells with horizontal black bars contain the 13 compounds from the original SARM (Figure 4C). Most value 2 and value 1 fragments from the original SARM have low log_likelihood score (Figure 5B), indicating that DeepSARM learns structural information from Aurora inhibitors effectively. Novel value 2 and value 1 fragments are generated from other kinase inhibitors (originating from SARfari). Novel fragments with low log_likelihood score are located close to fragments from the original SARM in latent space of the decoder. Accordingly, as a results of fine-tuning, DeepSARM suggests novel fragments that are related to yet distinct from fragments in the original SARM.

### *Generating new SARMs*

DeepSARM expansion may also result in the formation of new SARMs exclusively containing novel fragments and analogs. Figure 6 shows a representative example. In Figure 6A, output structures from the value 2 generator are shown, which produced 93 valid SMILES in 100 random trials. After removing duplicated SMILES, 22 unique compounds with new value 2 fragments were obtained. In addition, 22 key 1 fragments were constructed from the 22 value 2 fragments and the designated as key 2–9. The value 1 generator then produced 69 unique molecules with value 1 fragments in 10 random trials with each key 1. A subset of 30 of these value 1 fragments is shown. Figure 6B illustrates the construction of a new SARM with new key 1 fragments obtained by combining key 2–9 and value 2 fragments. Key 1 fragments are placed in the leftmost column (vertical axis of the SARM) and value 1 fragments in corresponding rows. Each combination (cell) of a new key 1 and value 1 fragment yields a unique compound. For SARM expansion, key 1 and value 2 fragments were filtered and accepted on the basis of log_likelihood scores applying threshold values of $S \leq 10$ and $S \leq 5$, respectively. Through DeepSARM expansion, SARMs exclusively containing new structures can be systematically constructed for each newly generated key 2 fragment, as illustrated in Figure 6B. Thereby, analog space of original SARMs is substantially increased with structures related to yet distinct from original fragments.

## Conclusion

The SARM methodology was originally developed to combine the identification and structural organization of compound series with the design of new analogs. Individual SARMs, which are reminiscent of R-groups tables and thus easily accessible by medicinal chemists, represent the basic data structure. For different sets of active compounds, variably sized ensembles of SARMs are usually obtained. The SARM method was further extended through incorporation of functions for compound activity prediction, combined SARM display and systematic analog searching. Original SARM-based analog design was confined to recombination of structural fragments extracted from known inhibitors, thus narrowly charting chemical space around series of interest. It has been our intention to further enhance analog design including novel fragments and compounds, thereby increasing coverage of series-centric chemical space. To these ends, the DeepSARM concept was developed. DeepSARM is an analog design approach that is specific to the SARM context. There currently is no comparable analog design methodology available. The underlying idea is refining and expanding analog design by taking compound information for a target family into consideration, followed by fine-tuning of the design toward for a target of interest. This makes it possible to complement a population of virtual analogs derived from compounds active against a target of interest with analogs having novel structural features. Therefore, an integral part of DeepSARM is a deep learning architecture, comprising three Seq2Seq models for fragment generation following the key–value-based compound design strategy specific to the SARM methodology. DeepSARMs expands existing SARMs (individual compound matrices) through the incorporation of novel fragments and generates new SARMs exclusively consisting of new fragments and virtual analogs. As an exemplary application, SARM expansion through generative modeling was carried out for kinase inhibitors, illustrating the DeepSARM approach.

## Future perspective

Given the critically important role of analog design for SAR analysis, hit-to-lead and LO projects in medicinal chemistry, computational support is highly desirable. Analog searching and design can be attempted in different ways at varying levels of computational sophistication. The SARM method combines structural analysis of compound datasets and organization of analog series with the generation of new virtual analogs, which sets it apart from other computational approaches. DeepSARM further extends SARM-based design through expansion of matrices with structurally novel fragments and analogs. To achieve this goal, an SARM-specific architecture for Seq2Seq learning

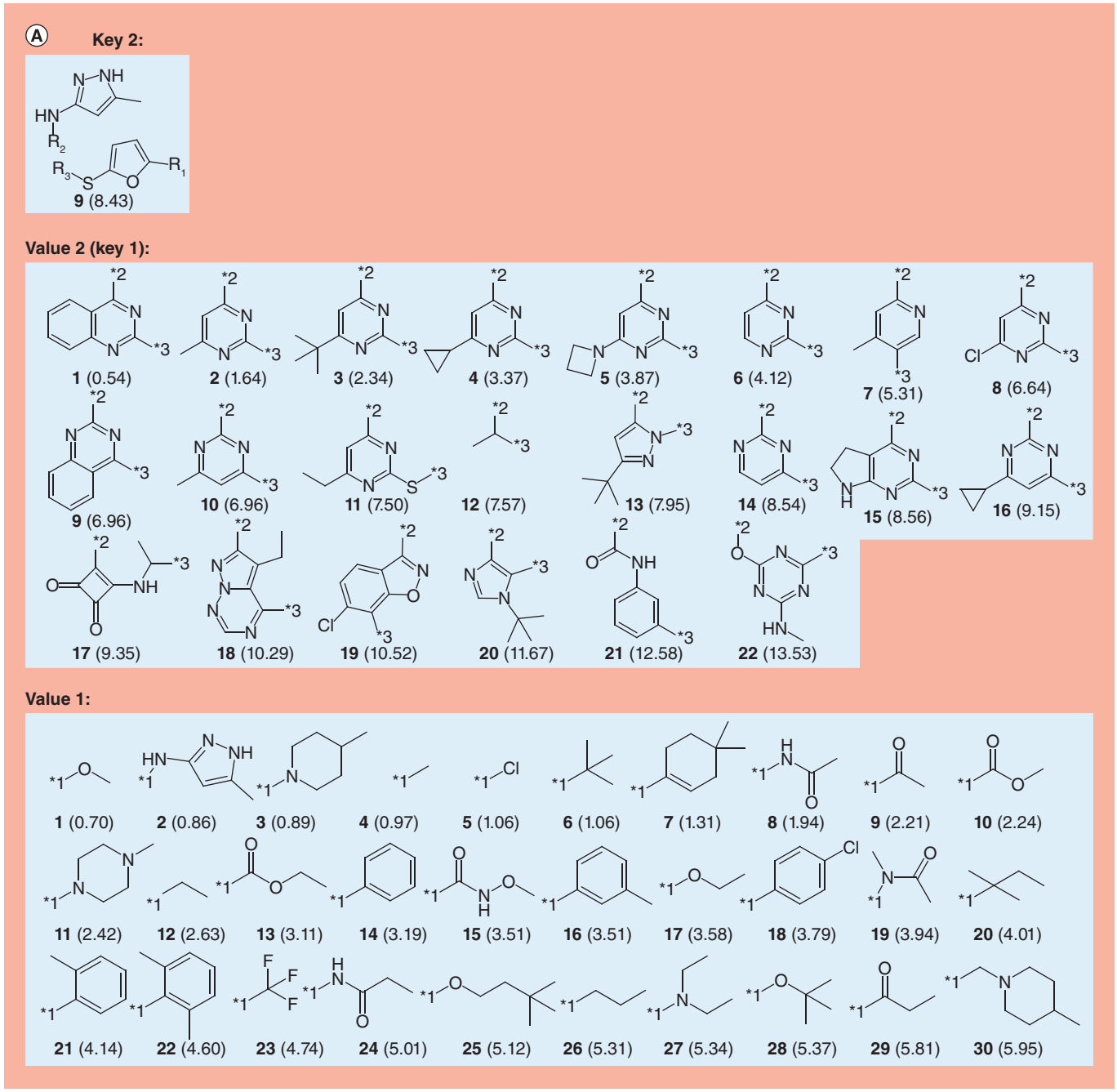



**Figure 6. Novel structure–activity relationship matrix. (A)** Novel key 2, value 2 and value 1 structures are shown used for structure–activity relationship matrix (SARM) expansion. **(B)** The resulting SARM exclusively contains virtual analogs not present in original SARMs. The representation is according to Figure 5C.

and generative modeling was designed and implemented. The DeepSARM methodology has been extensively tested and is readily applicable to different targets and families, as illustrated by an exemplary application presented herein. Going forward it is intended to use the DeepSARM framework for practical applications in hit-to-lead and LO projects. For example, an application scenario highly suitable for DeepSARM, given its characteristic features, is its use on large and related parallel compound series during later stages of LO campaigns. In such cases, all structural relationships between series can be systematically investigated and analog design via DeepSARM can explore potential opportunities to bridge between series and combine different structural features. Furthermore, given DeepSARM's two-level fragment generation approach, it can be elegantly used to complement analog design

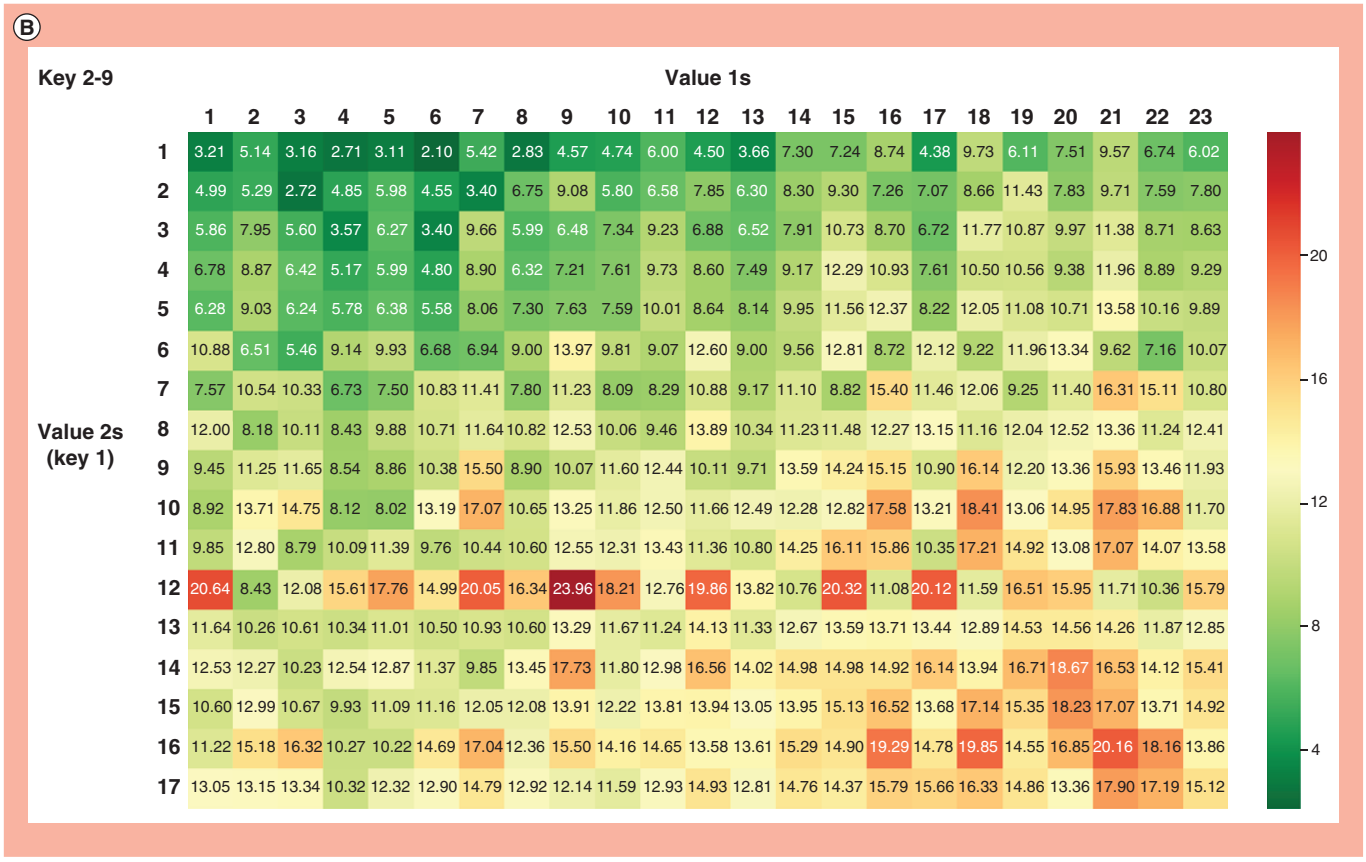



**Figure 6. Novel structure–activity relationship matrix (cont.). (A)** Novel key 2, value 2 and value 1 structures are shown used for structure–activity relationship matrix (SARM) expansion. **(B)** The resulting SARM exclusively contains virtual analogs not present in original SARMs. The representation is according to Figure 5C.

for evolving series with information from external compound sources. In this case, the target family compound pool contains active compounds of interest from external sources. The implemented learning approach is capable of producing structural novelty through generating and combining fragments of different origins. Moreover, DeepSARM will also be applicable to design tasks outside the LO context. For example, for target families with single or multiple high-priority targets, DeepSARM models can generate focused virtual libraries on the basis of existing compound datasets. Hence, there is a variety of attractive opportunities for future applications of the DeepSARM approach. Finally, from a computational viewpoint, one might also consider adding other filter functions to DeepSARM for further prioritizing virtual candidates on the basis of multiple optimization-relevant molecular properties.

## Author contributions

A Yoshimori and J Bajorath conceived the study, implemented the methods, analyzed the results and prepared the manuscript.

## Acknowledgments

We thank H Kouji for insightful comments and suggestions.

## Financial & competing interests disclosure

A Yoshimori leads the Institute for Theoretical Medicine, Inc. (ITM), which develops commercial scientific software. J Bajorath is a consultant to ITM. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

## Summary points

- The generation of analogs plays a central role in medicinal chemistry.
- Computational methods for analog design are of high interest.
- The structure–activity relationship matrix (SARM) methodology is unique among analog design approaches.
- SARM is further extended for exploring new chemical space.

**Methodology**

- The DeepSARM concept is introduced.
- Sequence-to-sequence models are built for generating new structural fragments.
- The DeepSARM architecture and workflow for analog design are detailed.
- Expansion of SARMs with novel fragments and compounds is illustrated.

**Exemplary application**

- The generation of new analogs of Aurora A kinase inhibitors is described.
- Kinome-wide inhibitors are taken into account.
- SARM expansion is detailed using representative examples.
- Coverage of chemical space around Aurora inhibitors is extended.

**Future perspective**

- The DeepSARM methodology is extensively tested.
- Practical applications of DeepSARM will be a major focal point.
- Examples include late-stage lead optimization efforts and focused library design.
- Functionalities will be added to further prioritize virtual candidates.

## References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Cannon JG. Analog design. In: *Burger's Medicinal Chemistry and Drug Discovery.* Abraham DJ (Ed.). John Wiley & Sons, Inc, NJ, USA, 687–714 (2003).
2. Wermuth CG. Similarity in drugs: reflections on analogue design. *Drug Discov. Today* 11(7–8), 348–354 (2006).
3. Bon RS, Waldmann H. Bioactivity-guided navigation of chemical space. *Acc. Chem. Res.* 43(8), 1103–1114 (2010).
4. Ertl P. Intuitive ordering of scaffolds and scaffold similarity searching using scaffold keys. *J. Chem. Inf. Model.* 54(6), 1617–1622 (2014).
5. Dimova D, Stumpfe D, Hu Y, Bajorath J. Analog series-based scaffolds: computational design and exploration of a new type of molecular scaffolds for medicinal chemistry. *Future Sci. OA* 2(1), FSO149 (2016).
6. Griffen E, Leach AG, Robb GR, Warner DJ. Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* 54(22), 7739–7750 (2011).
7. Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* 50(3), 339–348 (2010).

•• **Effective algorithm for matched molecular pair fragmentation.**

8. Walters WP. Virtual chemical libraries. *J. Med. Chem.* 62(3), 1116–1124 (2019).

• **Instructive discussion of large virtual libraries and their potential utility.**

9. Stewart KD, Shiroda M, James CA. Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.* 14(20), 7011–7022 (2006).

• **Early and popular program for analog design.**

10. Wassermann AM, Haebel P, Weskamp N, Bajorath J. SAR matrices: automated extraction of information-rich SAR tables from large compound data sets. *J. Chem. Inf. Model.* 52(7), 1769–1776 (2012).

•• **Introduction of the structure–activity relationship matrix concept.**

11. Gupta-Ostermann D, Bajorath J. The 'SAR matrix' method and its extensions for applications in medicinal chemistry and chemogenomics. *F1000Res.* 3(1), e113 (2014).
12. 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. *F1000Res.* 4(1), e75 (2015).
13. Gupta-Ostermann D, Shanmugasundaram V, Bajorath J. Neighborhood-based prediction of novel active compounds from SAR matrices. *J. Chem. Inf. Model.* 54(3), 801–809 (2014).

14. Yoshimori A, Tanoue T, Bajorath J. Integrating the structure–activity relationship matrix method with molecular grid maps and activity landscape models for medicinal chemistry applications. *ACS Omega* 4(4), 7061–7069 (2019).

15. Yoshimori A, Horita Y, Tanoue T, Bajorath J. Method for systematic analogue search using the mega SAR matrix database. *J. Chem. Inf. Model.* 59(9), 3727–3734 (2019).

• **Approach for large-scale structure–activity relationships matrix application.**

16. Fleming N. How artificial intelligence is changing drug discovery. *Nature* 567(7706), S55–S55 (2018).

17. Mater AC, Coote ML. Deep learning in chemistry. *J. Chem. Inf. Model.* 59(6), 2545–2559 (2019).

•• **Comprehensive review of deep learning approaches in chemistry.**

18. Wawer M, Bajorath J. Local structural changes, global data views: graphical substructure-activity relationship trailing. *J. Med. Chem.* 54(8), 2944–2951 (2011).

19. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28(1), 31–36 (1988).

20. RDKit: cheminformatics and machine learning software (2013). www.rdkit.org

21. Zheng S, Yan X, Gu Q *et al.* QBMG: quasi-biogenic molecule generator with deep recurrent neural network. *J. Cheminform.* 11(1), e5 (2019).

• **Recurrent neural network architecture for generative molecular design.**

22. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: *Advances in Neural Information Processing Systems 27*. Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (Eds). Massachusetts Institute of Technology Press, MA, USA, 3104–3112 (2014).

• **Introduction of sequence-to-sequence models.**

23. Hochreiter S, Schmidhuber J. Long short-term memory. *Neur. Comput.* 9(8), 1735–1780 (1997).

24. Keras. https://github.com/keras-team/keras

25. Bento AP, Gaulton A, Hersey A *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42(Database issue), D1083–D1090 (2014).

26. Kinase SARfari (2019). www.ebi.ac.uk/chembl/sarfari/kinasesarfari