Artificial Intelligence in the Life Sciences 1 (2021) 100015

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Methods & Protocols

Iterative DeepSARM modeling for compound optimization

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

[a] Institute for Theoretical Medicine, Inc., 26-1 Muraoka-Higashi 2-chome, Fujisawa, Kanagawa 251-0012, Japan
[b] Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 5/6, D-53115 Bonn, Germany

ARTICLE INFO

Keywords:
Structure-activity relationships
Molecular design
SAR matrix
Deep learning
Generative modeling
Compound optimization

ABSTRACT

The Structure-Activity Relationship (SAR) Matrix (SARM) method systematically extracts structurally related compound series from any source and organizes these series in a unique data structure formed by matrices similar to R-group tables from medicinal chemistry. In addition, the SARM method generates virtual analogues for structurally organized series that consist of new combinations of existing core structures and R-groups. For active compounds, SARMs visualize SAR patterns and aid in compound design. The SARM methodology and data structure was integrated with a recurrent neural network architecture to further expand the compound design capacity with deep generative models, leading to the DeepSARM approach. Herein, we present an extension of the DeepSARM framework for compound optimization termed iterative DeepSARM (iDeepSARM), which involves multiple iterations of deep generative modeling and fine-tuning to obtain increasingly likely active compounds for targets of interest. Hence, iDeepSARM adds computational hit-to-lead and lead optimization capability to the DeepSARM framework. In addition to detailing methodological features and calculation protocols, an exemplary compound design application is reported to illustrate the iDeepSARM approach.

1. Introduction

Computational methods for the systematic analysis and visualization of structure-activity relationships (SARs) are of high interest in medicinal chemistry [1–5]. Equally relevant are approaches providing computational decision support for compound (hit-to-lead and lead) optimization [6–10]. Most of the latter approaches employ statistical data analysis, identify key compounds generated during optimization efforts, and/or monitor SAR progression. However, these methods typically do not suggest new compounds for synthesis. The same applies to approaches for SAR visualization, which are mostly descriptive in nature.

To our knowledge, there currently is only one method available that combines the systematic assessment of structural relationships between active compounds with SAR visualization and prospective compound design, i.e., the SAR Matrix (SARM) approach [11]. SARM analysis suggests new analogues by systematically comparing core structures and substituents in related analogue series (ASs) and identifying unexplored core-substituent combinations across these series. The compound design component of SARM has been extended through deep generative modeling, leading to DeepSARM [12], which utilizes additional information from selected targets and associated compound data sets to further expand the design space.

Herein, we report the extension of DeepSARM for iterative compound optimization. In the following, we first introduce the SARM/DeepSARM framework and then detail the iterative DeepSARM (iDeepSARM) approach. An exemplary compound design application is presented to illustrate methodological characteristics of iDeepSARM and its design capabilities.

2. Methodological framework

2.1. SAR matrix concept

SARMs were originally designed to systematically extract ASs from large compound collections, organize structurally related ASs in matrices reminiscent of R-group tables, visualize SARs, and generate virtual candidate compounds to further expand ASs. The identification and organization of structurally related ASs is facilitated by a dual-step compound decomposition scheme [11] akin to fragmentation of bonds for the generation of matched molecular pairs (MMPs) [13]. In the first step, exocyclic single bonds in compounds are systematically cleaved applying size limitations for the resulting fragments, which yields keys (core structures, scaffolds) and values (substituents, R-groups) that are stored in an index table. This procedure identifies all analogues sharing a particular core with R-group replacements at a single site, hence defining a matching molecular series (MMS) [5] for each structurally unique core.

* Corresponding author at: University of Bonn, Germany.
E-mail address: bajorath@bit.uni-bonn.de (J. Bajorath).

https://doi.org/10.1016/j.ailsci.2021.100015
Received 20 October 2021; Received in revised form 14 November 2021; Accepted 16 November 2021
Available online 20 November 2021

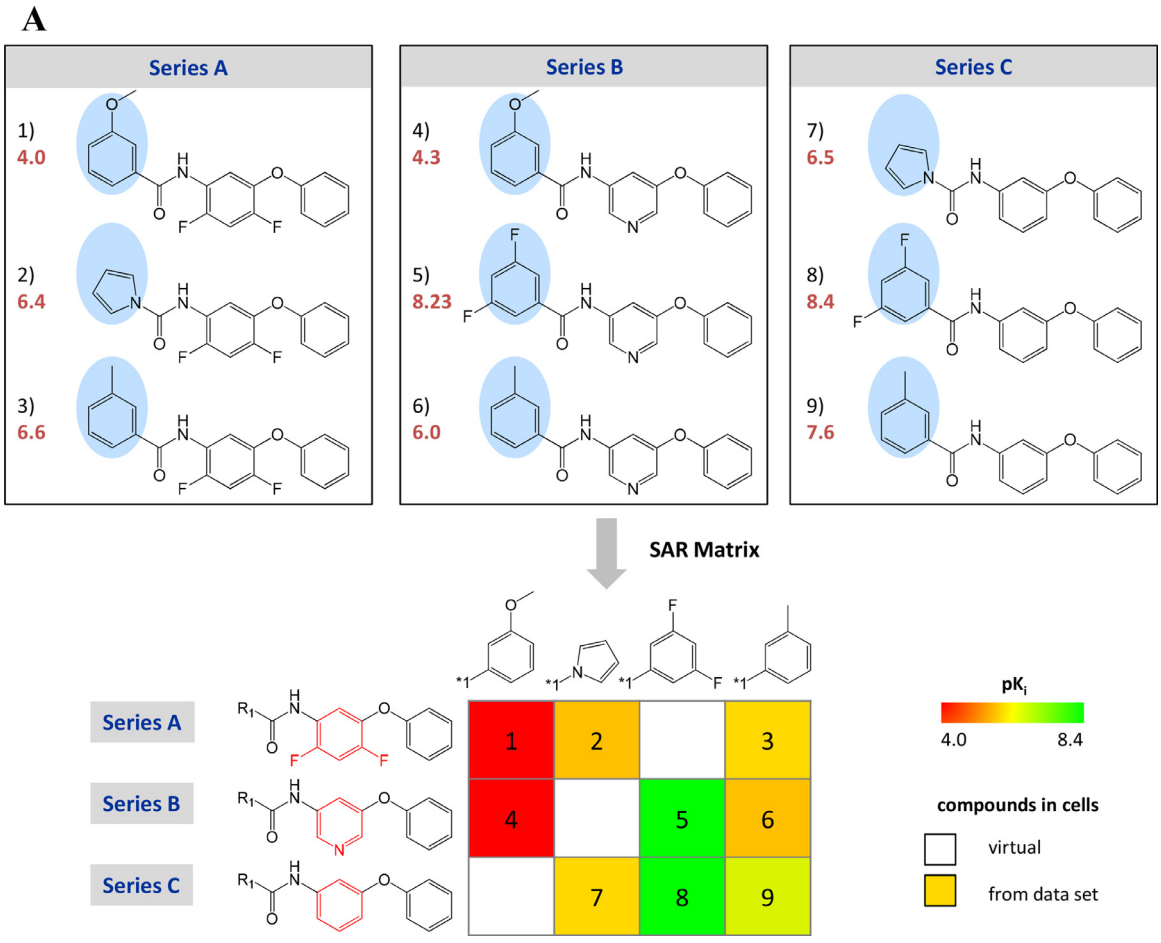

Fig. 1. SAR matrix generation using DeepSARM. (A) The design of the SARM data structure is illustrated using three small structurally related compound series (A, B, and C). Analogues from the different series are consecutively numbered and their pK_i values are reported in red. Distinguishing substituents are shown on a blue background and substructures differentiating scaffolds (keys) are colored red. In the SARM, each row contains an MMS and each column compounds from different series with the same substituent (value). Existing analogues are represented by cells that are color-coded by activity. In addition, empty cells represent virtual analogues. The figure was adopted from [15]. **(B)** The architecture of DeepSARM is illustrated, which consists of three LSTM encoder-decoder units. The figure was adopted from [12]. **(C)** The construction of a SARM with DeepSARM is illustrated using Seq2Seq models for value 2 and value 1. Key 2 is a scaffold, which represents the input of the Seq2Seq model for generating value 2, the R_2 substituent of the key 2 scaffold. Key 1 is constructed from key 2 and value 2. Key 1 is the input of the Seq2Seq model for value 1, the R_1 substituent of key 1 or key 2. The SARM is constructed from the resulting key 1 and value 1 fragments and color-coded by log-likelihood scores. Value 1 and value 2 filters represent structural screens to remove chemically questionable substituents (SMARTS patterns of in-house collections of chemically undesired substructures).

In the second step, all keys obtained in the first round are re-submitted to fragmentation, which then identifies all structurally analogous cores with a chemical change at a single site and the corresponding MMSs. Each subset of MMSs with unique structurally related cores is organized in an individual SARM such that each row contains an MMS and each column compounds from different series sharing the same R-group, as illustrated in Fig. 1**A**. Depending on the ASs contained in a given compound collection and their structural relationships, varying numbers of SARMs are obtained. Each cell in a SARM represents a unique compound. SAR information is visualized by coloring cells according to compound potency. Hence, structural relationships and associated activity patterns can be traced within SARMs. Empty cells represent virtual analogues consisting of non-existing key-value (core-substituent) combinations, as also illustrated in Fig. 1**A.** Accordingly, virtual candidate compounds from SARMs further extend chemical space of related ASs and can be envisioned to form an envelope in chemical space around these series. The SARM methodology and resulting SARM data structure bridge between SAR visualization and compound design. The approach has been further extended through the integration of activity prediction methods [14] and molecular grid maps providing a consensus view of existing and virtual compounds organized across different SARMs of a compound data set [15]. SARM analysis and selection of virtual candidates have led to the identification of new active compounds for different targets [16–18].

2.2. Deep learning extension

DeepSARM was based on the idea to further expand analogue design by taking information from compounds with activity against different targets into account that are related to the primary target of interest [12]. For example, a DeepSARM model can initially be trained with compounds active against the target family to which the target of interest belongs, followed by fine-tuning of the model for the primary target. This procedure increases the close-in analogue design capacity of the SARM approach. SARMs resulting from original two-step fragmentation can then be further expanded with novel key and value fragments and additional SARMs entirely consisting of novel fragments and compounds can be obtained.

For generative design on the basis of key and value fragments encoded as canonical SMILES strings [19], an encoder-decoder frame-

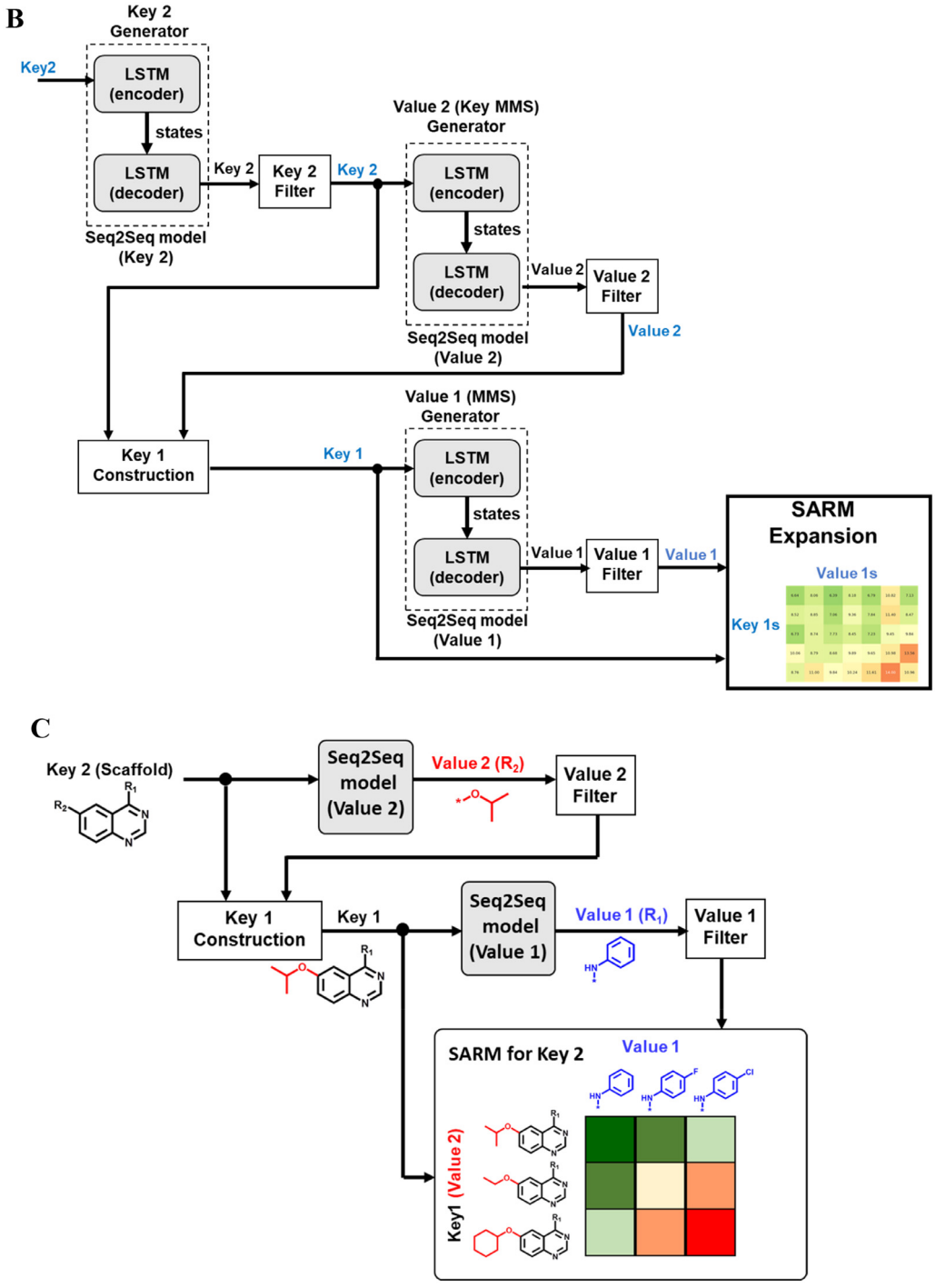

Fig. 1. Continued

work consisting of long short-term memory (LSTM) units [20] represents a preferred recurrent neural network (RNN) architecture [21]. The encoder-decoder framework is used to derive sequence-to-sequence (Seq2Seq) models that translate one sequence of one-hot encoded SMILES strings into another [22]. The encoder LSTM transforms input sequences into two-state vectors in latent space and the decoder LSTM is trained to return the same sequences as transformed SMILES. For DeepSARM, encoder-decoder units were built using keras [23] (with 256-dimensional latent LSTM encoding space).

DeepSARM includes three encoder-decoder units for the generation of Seq2Seq models, as illustrated in Fig. 1**B**. The Seq2Seq model for key 2 (i.e., the key 2 generator) is trained using input key 2 / output key 2 pairs, the model for value 2 using key 2 / value 2 pairs, and the model for value 1 using key 1 / value 1 pairs. Compounds with newly generated key fragments are added to an original SARM containing structurally analogous keys (meeting the step-2 fragmentation criterion) and hence further expand the SARM.

The three Seq2Seq models are derived as follows:

(I) Model (key 2) using key 2 (input) / key 2 (target) pairs;
(II) Model (value 2) using key 2 (input) / value 2 (target) pairs;
(III) Model (value 1) using key 1 (input) / value 1 (target) pairs.

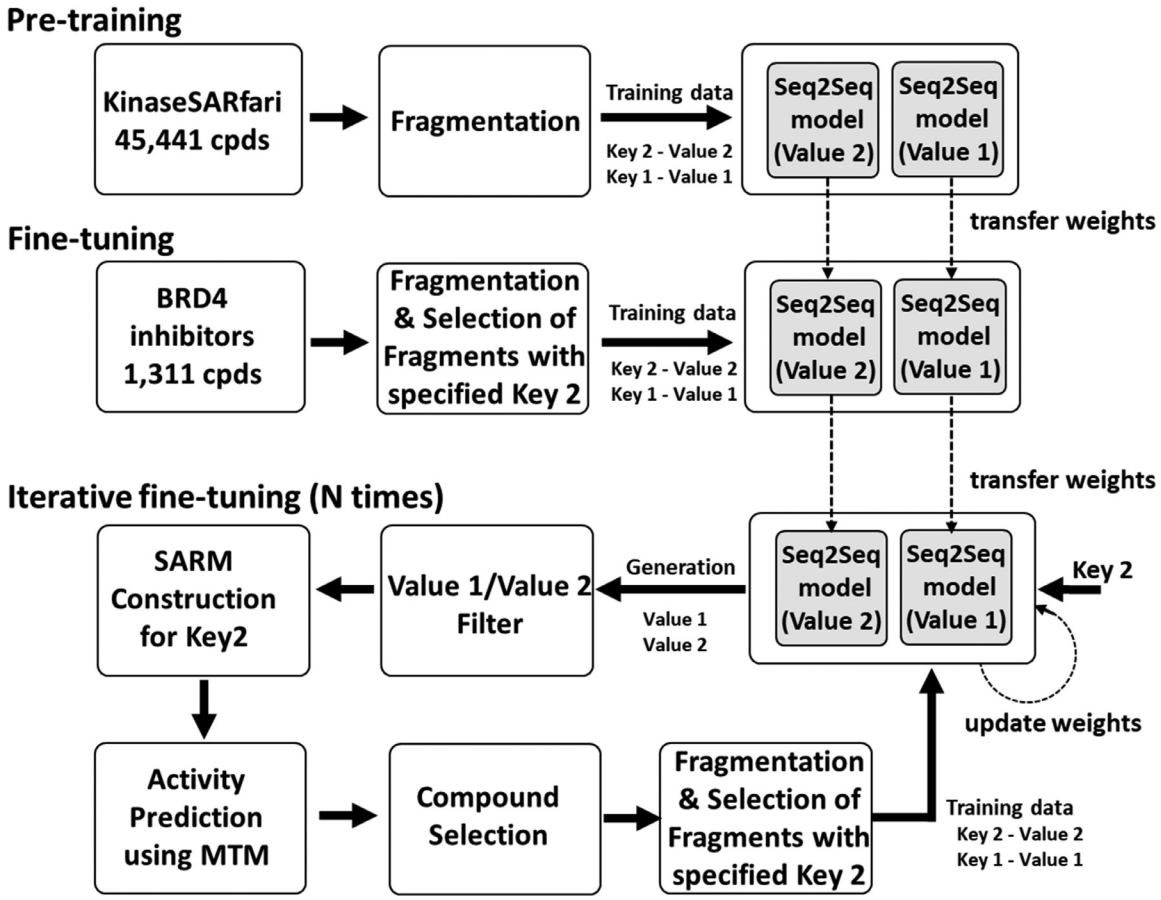

Fig. 2. Components of iterative DeepSARM. The three components (steps) comprising the iDeepSARM approach are illustrated including pre-training, fine-tuning, and iterative fine-tuning, as discussed in the text.

Pre-training is carried out using large numbers of compounds with activity against a target family and fine-tuning using a comparably small set of compounds active against the primary target. During fine-tuning, internal model weights are adjusted.

The Seq2Seq models generate key 2, value 2, and value 1 fragments that are evaluated on the basis of a log-likelihood score:

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

P represents the probability distribution of the decoder and T the number of SMILES tokens for a given fragment. The minus sign ensures that high probabilities result in small scores for fragment prioritization.

A log-likelihood threshold can be applied to filter fragments. Compounds are then obtained by combining newly generated key 1 and value 1 fragments. Further DeepSARM calculation details are provided in [12].

Fig. 1**C** illustrates the generation of a SARM using the Seq2Seq models for value 2 and value 1. Key 2 fragments provide the input for the Seq2Seq model generating value 2 fragments. Key 1 is then constructed from key 2 and value 2 fragments. The resulting key 1 fragments serve as is input for the Seq2Seq model of value 1. New compounds comprising the SARM are then obtained by combining the resulting key 1 and value 1 fragments. Each cell of the SARM represents a new compound and is color-coded by the combined log-likelihood score. Invalid SMILES strings are removed and value 1 and value 2 filters are SMARTS screens to remove chemically undesired substituents (such as unstable fragments).

3. Iterative DeepSARM

3.1. Methodological concept and workflow

The iDeepSARM approach was designed for the optimization of newly identified active compounds. It consists of three components (steps) including pre-training, fine-tuning and iterative fine-tuning, as summarized in Fig. 2. Following SARM fragmentation of compounds, key 1 / value 1 and key 2 / value 2 pairs are used to train Seq2Seq model (value 1) and Seq2Seq model (value 2), respectively. Iterative fine-tuning focuses on key 2 fragments (scaffolds) from DeepSARM. Each scaffold represents an MMS. Value 1 and value 2 fragments serve as R-groups for two different substitution sites. Newly generated value 1 and 2 fragments are filtered for structural alerts (SMARTS of in-house collections of chemically undesired substructures) and accepted fragments are used to construct a SARM variant with pre-defined dimensionality for each key 2. This SARM variant consists of value 1 / value 2 combinations from which corresponding compounds (key 2 / value 1 / value 2 combinations) are enumerated and subjected to activity prediction using the molecular topographic map (MTM) approach [24] further described below. Structural elitism selection focuses on compounds with n top-ranked predicted activities, which are then subjected to re-fragmentation and selection of key 1 / value 1 and key 2 / value 2 pairs related to the key 2 fragment (scaffold). Newly generated key 1 / value 1 and key 2 / value 2 pairs are then used as input data for the next fine-tuning cycle. For the initial iteration, Seq2Seq model weights are transferred from DeepSARM fine-tuning. During iterative fine-tuning, model weights are updated following each cycle. Iterations are carried out un-

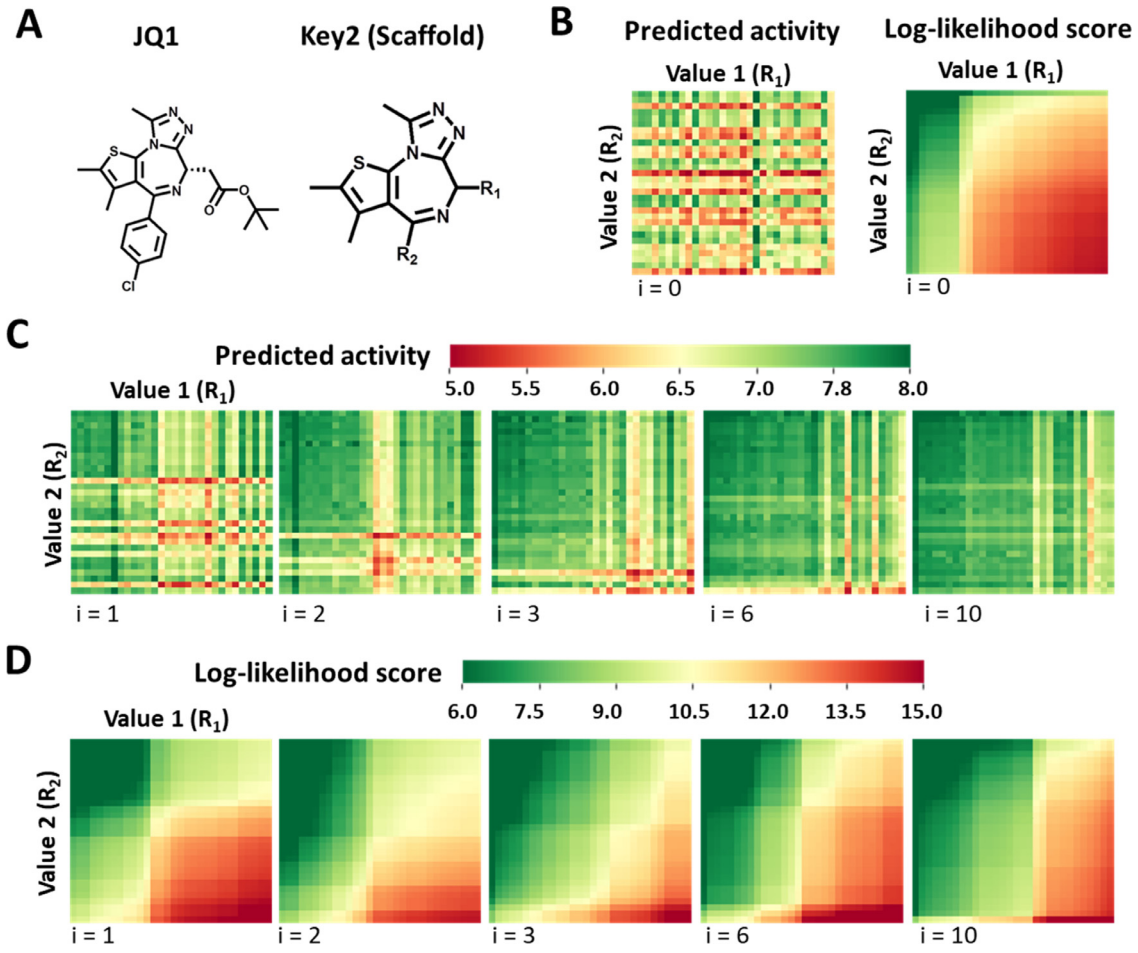

Fig. 3. Compound optimization using iDeepSARM. The optimization of key 2-containing compounds using iDeepSARM is illustrated. (**A**) The structure of BRD4 inhibitor JQ1 and its key 2 are shown for which the optimization is performed. (**B**) A SARM with JQ1 key 2-containing compounds generated with DeepSARM is displayed, representing the starting point of the optimization (iteration $i = 0$). The SARM consists of 30 value 1 (R_1) and 30 value 2 (R_2) fragments yielding 900 compounds. It is color-coded by predicted compound activity (pIC_{50}) or log-likelihood score for value 1 / value 2 combinations. (**C**) The SARM is updated with new value 1 and 2 fragments obtained during iterative fine-tuning. Results are shown for five iterations ($i = 1, 2, 3, 6, 10$) color-coded by compound activity. (**D**) The five iteratively updated SARMs are color-coded by log-likelihood scores. (**E**) Value 2 and (**F**) value 1 structures from the SARM in (B) are shown (sorted in ascending order of likelihood). (**G**) Value 2 and (**H**) value 1 structures from the SARM for $i = 10$ are shown.

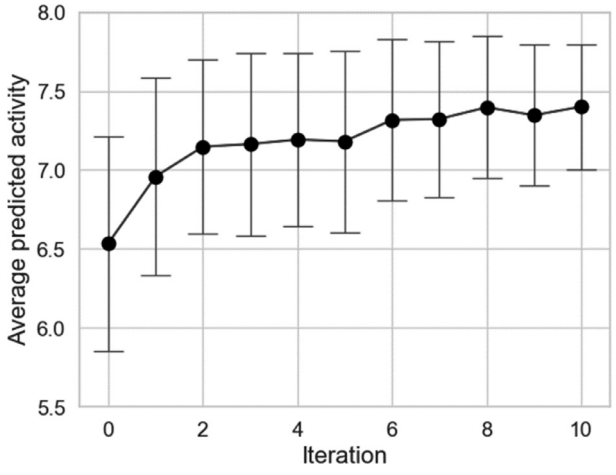

Fig. 4. Iterative compound activity prediction. Mean predicted compound activities are reported for the SARM in Fig. 3 during iterative optimization. Error bars indicate standard deviations.

til a pre-defined activity threshold or number of iterations is reached. Success of iterative fine-tuning intrinsically depends on the reliability of compounds activity predictions and log-likelihood scores. Preferably, comparable trends in the progression of activity and log-likelihood estimates should be observed during the optimization procedure.

3.2. Activity prediction

For iDeepSARM, a new image-based activity prediction method is employed, i.e. MTM [24]. The approach is summarized in **Supplementary Figure S1A**. The underlying idea is to use feature sets of a compound to generate a generative topographic mapping (GTM) model [25] that projects features onto a two-dimensional matrix and produces an image representation. The GTM model is calculated with the runGTM function of ugtm [26]. A convolutional neural network (CNN) [27] is then trained on molecular images of known active compounds and their activity data and used for activity prediction. The architecture of the MTM CNN is illustrated in **Supplementary Figure S1B**. Conv2D_1 and Conv2D_2 are convolution layers and MaxPooling2D_1 and MaxPooling2D_2 pooling layers. In the Dropout layer, randomly selected neurons are ignored during training. In the Flatten layer, a matrix is converted

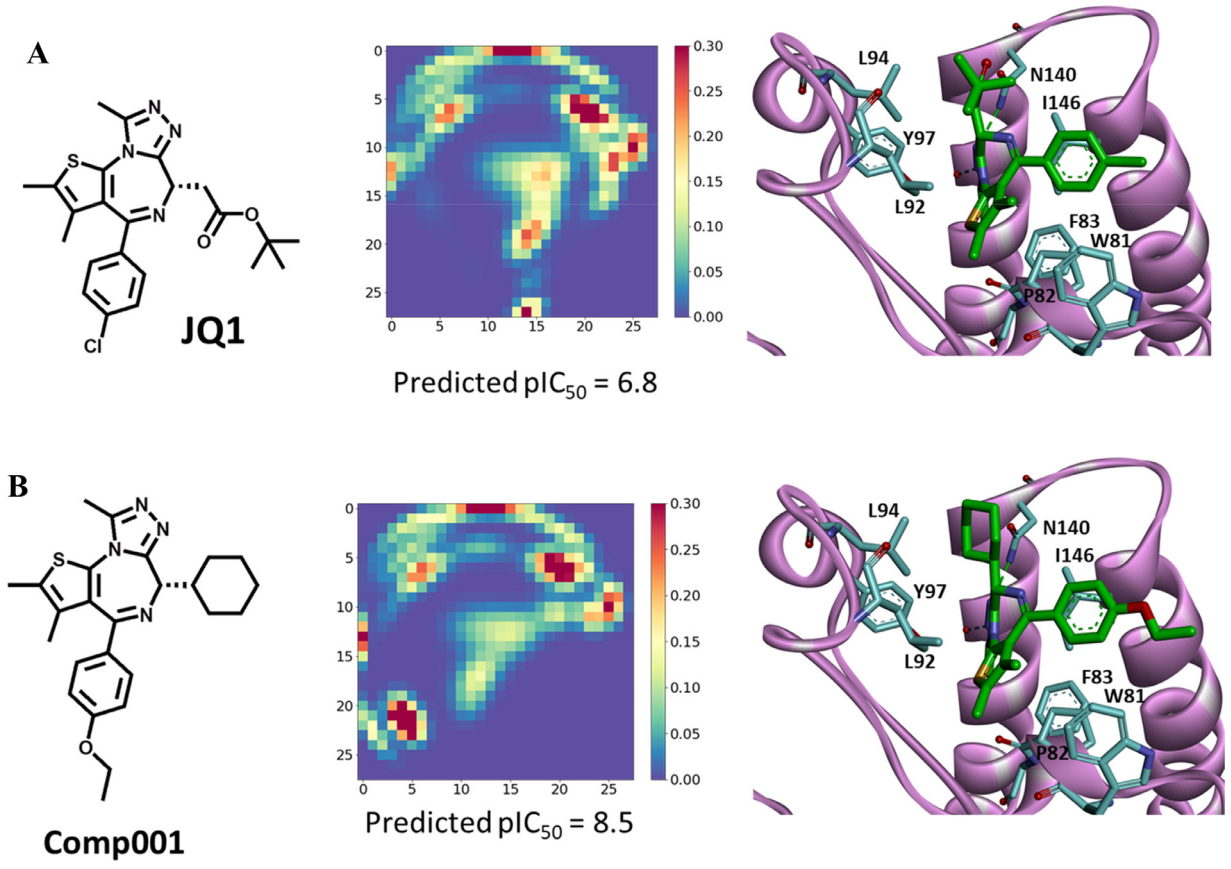

Fig. 5. Ligand-target interactions. (**A**) Compound JQ1 is shown together with its MTM and the X-ray structure of its complex with BRD4 (PDB ID: 3MXF). MTM-based activity prediction for JQ1 is reported (experimental IC_{50} value: 77 nM). (**B**) The corresponding representation is shown for a docking model of a computational candidate compound from iDeepSARM predicted to have higher activity (Comp001). In the complexes, carbon atoms of the ligands and selected binding site residues of BRD4 are colored green and cyan, respectively (otherwise, standard atom coloring is used). Remaining parts of BRD4 are depicted as a ribbon representation.

into a single array. Dense_1 and _2 represent fully connected layers. The CNN structure was implemented using keras [23]. An attractive feature of the MTM approach is the ability to visualize feature set (dis)similarity of test compounds to complement activity prediction. However, a major motivation for using MTM for iDeepSARM is data augmentation, which is required when only small compound sets are available for fine-tuning. For images, a robust data augmentation approach is available that generates hybrid images from pairs of images on the basis of continuously valued mixing coefficients [28], which also enables precise calculation of associated hypothetical activity values.

For the calculations reported below, atomic features sets were calculated using an algorithm [24] similar to the one for generating the extended connectivity fingerprint with bond diameter 4 (ECFP4) [29]. For runGTM, the following parameters were used: $k = 28$, $m = 2$; k is the square root of the number of GTM nodes and m the square root of the number of radial basis function centers. For other parameters, default settings were used. After constructing the GTM model, the responsibilities of atomic features were calculated using the transform function of ugtm and the MTM was generated based upon the summed responsibilities.

4. Exemplary compound design and optimization

4.1. Design strategy

To illustrate iDeepSARM modeling, optimization of putative multi-target compounds is attempted. The application was inspired by a recent study [30] where DeepSARM was used to identify candidates for dual-target inhibitors of a PLK1 kinase and bromodomain-containing protein 4 (BRD4), a chromatin-associated protein functioning as an epigenetic regular [31]. Both PLK1 and BRD4 are cancer targets. In this compound design effort, DeepSARM was pre-trained using compounds with activity against PLK1 or BRD4 and fine-tuned with PLK1/BRD4 dual-target inhibitors [30]. Here, we have modified the design protocol by pre-training with inhibitors of the human kinome, followed by fine-tuning with BRD4 inhibitors and iDeepSARM optimization of BRD4 inhibitory activity. Hence, generated candidate compounds are expected to have the potential to be multi-kinase and BRD4 inhibitors and are optimized for BRD4. Although we currently do not have the option to experimentally test preferred candidates, this design effort is well suited to illustrate the computational iDeepSARM approach.

4.2. Compound data and calculation parameters

From the Kinase SARfari collection of ChEMBL [32], 45,441 unique human kinase inhibitors with reliable activity measurements were selected for pre-training (step 1 in Fig. 2). In addition, 2280 unique BRD4 inhibitors were extracted from ChEMBL (ID: 1,163,125). For initial fine-tuning, 1311 BRD4 inhibitors with $pIC_{50} \geq 6$ were selected (step 2 in Fig. 2). For deriving the MTM CNN activity prediction model, all 2280 inhibitors were used.

For training of iDeepSARM's iterative fine-tuning Seq2Seq models (step 3 in Fig. 2), the number of epochs was set to 10 and batch size to 64. Key 1 / value 1 and key 2 / value 2 pairs were divided into training and validation sets (9:1), respectively. For iterative fine-tuning, temperature factors for value 1 and 2 generation were set to 1.0. Fine-tuning was carried out over 10 iterations. For each iteration, 100 computed compounds with highest predicted activity were selected as input. For

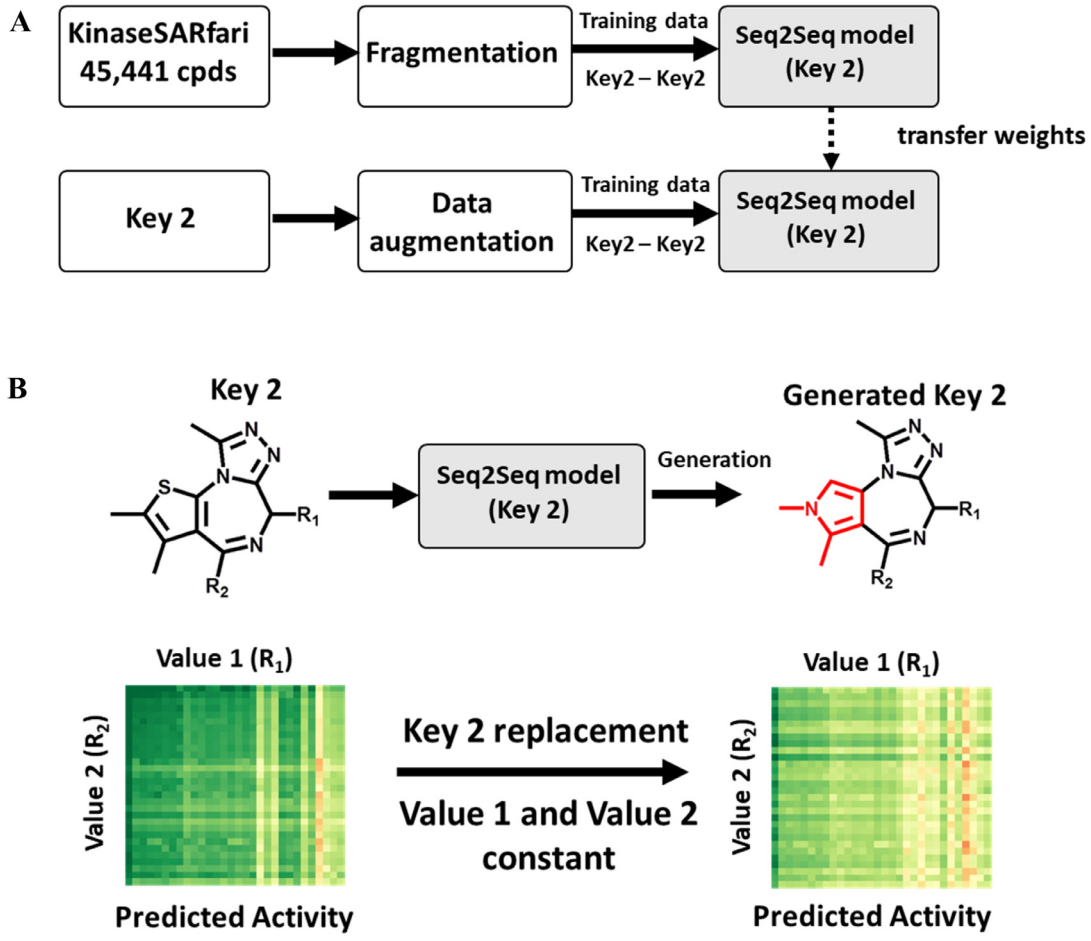

Fig. 6. Generation of new key 2 structures. (**A**) The protocol for generating new key 2 structures via iDeepSARM with the aid of data augmentation is shown. (**B**) At the top, the original and an exemplary predicted new key 2 are shown. The replaced ring moiety is colored red. At the bottom, SARMs comprising compounds containing the original (left) and new key 2 (right) with constant value 1 and 2 fragments are color-coded by predicted activity according to Fig. 3C.

a specific key 2, a value 1 x value 2 SARM variant with matrix dimensionality of 30×30 was generated.

For deriving the MTM CNN model, the compound/MTM data set was divided into training, validation, and test sets (8:1:1). The model is illustrated in **Supplementary Figure S1B.** Hyperparameters of the model (see legend of Supplementary Fig. 1) were determined using the optuna hyperparameter optimization software framework [33].

Supplementary Figure S2A and **S2B** show the correlation between predicted and experimental pIC_{50} values for the training and test set, respectively. The results indicate reasonable accuracy of the MTM CNN model, with only a limited number of prediction errors greater than one order of magnitude for the test set.

4.3. Iterative optimization

In Fig. 3**A**, a known BRD4 inhibitor termed JQ1 is shown together with its key 2 fragment that contains two substitution sites, R_1 and R_2, for value 1 and value 2 fragments, respectively. For JQ1, value 1 and 2 were predicted using DeepSARM. Fig. 3**B** shows the resulting (value 1 x value 2) SARM variant color-coded by predicted compound activity (left) or combined log-likelihood score from Seq2Seq models. As can be seen, the initially generated JQ1 analogues had significantly varying predicted activity and the distribution of log-likelihood scores was dominated by relatively low values. On the basis of these input data, iDeepSARM optimization was carried out over 10 iterations.

Fig. 3**C** and 3**D** show the corresponding SARM representations for five iterations color-coded by predicted compound activity and log-likelihood scores, respectively. The SARMs displayed a clearly evident parallel progression towards increasing compound activity as well as decreasing log-likelihood and hence indicated successful compound optimization.

Supplementary Figure S3A and **S3B** show value 2 and 1 fragments, respectively, from the initial SARM ($i = 0$), the starting point of the optimization, and Supplementary **Figure S3C** and **S3D** value 2 and value 1 fragments, respectively, after the final ($i = 10$) iDeepSARM iteration. The starting structures were more diverse and the optimization increasingly prioritized substituted phenyl rings for value 2 (R_2 position) and aliphatic substituents or, to a lesser extent, cyclohexyl derivatives for value 1 (R_1). Fig. 4 shows the concomitant evolution of predicted compound activity over the 10 iterations. Taken together, the results provide a representative example of a successful iDeepSARM optimization effort.

For inhibitor JQ1, an X-ray structure in complex with BRD4 was available in the Protein Data Bank (PDB) [34]. Fig. 5**A** shows this complex together with the MTM calculated for JQ1. For comparison, Fig. 5**B** shows a candidate compound from the iDeepSARM optimization (designated Comp001) that was predicted to have ~50-fold higher activity than JQ1. The MTM of this candidate compound displayed global similarities to the one of JQ1, but also distinct differences in projected feature sets, for example, an additional feature center in the lower left quadrant of the map. A docking model of the candidate compound was generated with AutoDock Vina [35]. When the candidate compound was docked to BRD4, the same enantiomer as observed for JQ1 in the X-ray structure was used. In the model, a hypothetical binding mode similar to JQ1 was observed with conserved interactions, but also mod-

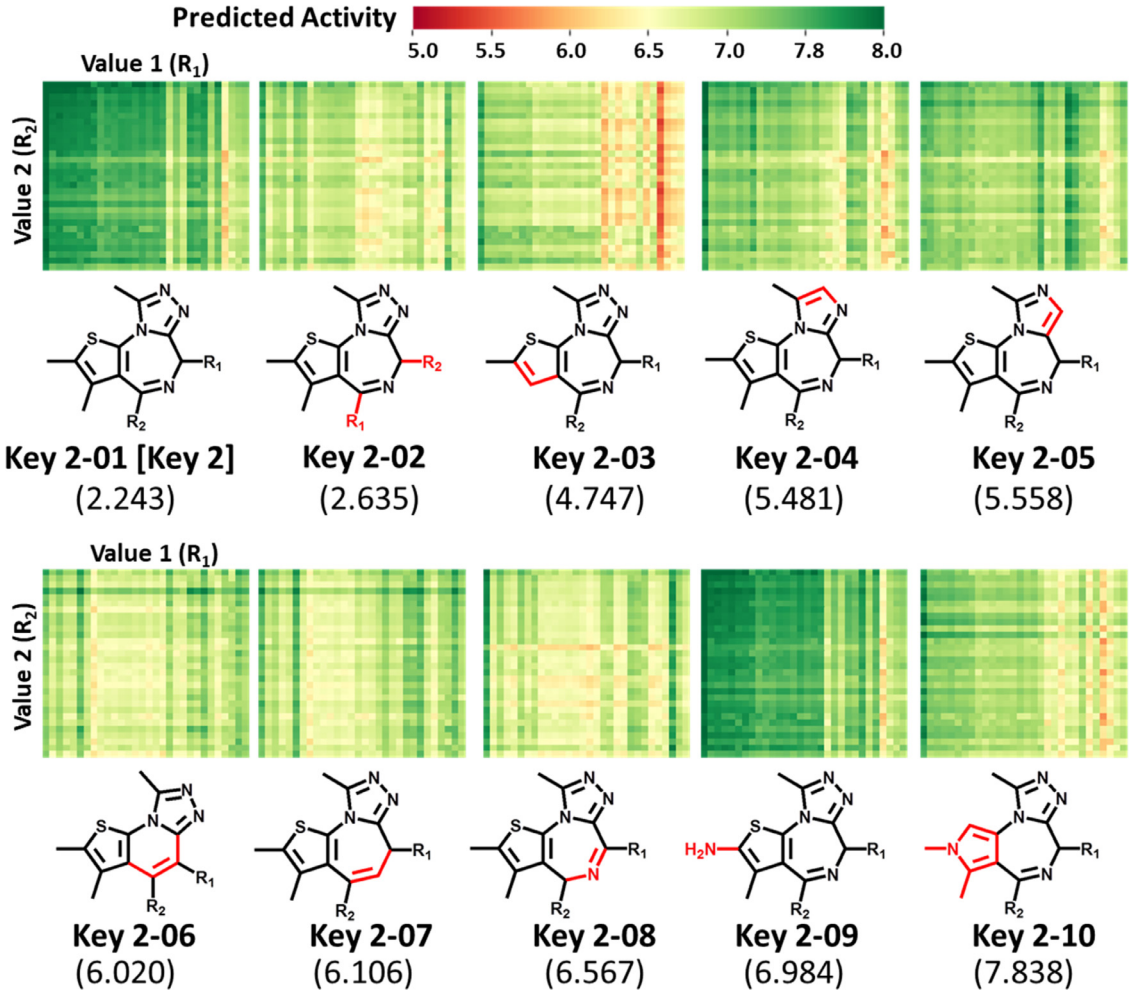

Fig. 7. Iterative scaffold replacements. SARMs with constant value 1 and 2 fragments and new key 2 structures (scaffolds) resulting from iDeepSARM optimization are shown color-coded by predicted activity. Key 2 structures and corresponding scaffold-replacement SARMs are arranged in the order of the optimization and key 2 log-likelihood scores are reported in parentheses. Structural modifications compared to the original key 2–01 are highlighted in red.

ulated interaction patterns for the new substituents at the R_1 and R_2 positions.

4.4. iDeepSARM variant for scaffold replacement

We also designed an alternative iDeepSARM architecture for key 2 (scaffold) replacement. This variant further extends the design capacity of iDeepSARM because in addition to selecting preferred R-group combinations (value 1 / value 2) for given scaffolds, the core structure of an AS can also be optimized. For this purpose, the Seq2Seq model for key 2 (Fig. 1**B**) was used for pre-training and fine-tuning, as shown in Fig. 6**A**. In the example presented herein, KinaseSARfari compounds were used for pre-training and the JQ1 key 2 for fine-tuning.

Following compound fragmentation for pre-training, key 2 fragments pairs were generated to train the Seq2Seq (key 2) model. For a selected key 2, the model was then fine-tuned using pairs of this key 2 and other key 2 fragments obtained by data augmentation based on enumeration of non-canonical SMILES [36]. Using the selected key 2 as input, the fine-tuned model was then applied to predict modified key 2 fragments, as illustrated in Fig. 6**B**. For alternative keys, SARM variants with constant value 1 and value 2 fragments were generated and the resulting compounds subjected to activity predictions to select preferred candidates. The constantly used value 1 and value 2 fragments were the ones shown in **Supplementary Figure S3D** and **S3C,** respectively. Using the fine-tuned model, 300 key 2 fragments were generated for iDeepSARM scaffold replacement. Fig. 7 shows the top 10 key 2 fragments selected by log-likelihood score.

Compared to the value 1 and 2 (R-group) optimization for the JQ1 scaffold discussed above, there was no continuous improvement in compound activity, as shown in **Supplementary Figure S4**, indicating that the JQ1 scaffold was difficult to replace. Compounds containing key 2–09 in Fig. 7 in which a methyl group in JQ1 was replaced with an amino group displayed similar activity to JQ1-containing compounds. Importantly, the key 2 fragments obtained during optimization displayed a variety of structural modifications across the JQ1 scaffold, hence providing proof-of-principle for iDeepSARM's ability to optimize complex ring structures by generating many new structural variants.

5. Concluding discussion

Herein, we have introduced a computational methodology for compound optimization that builds upon the SARM/DeepSARM framework and further extends its molecular design capacity. The core of the iDeepSARM approach is an iterative optimization scheme based upon an RNN architecture that uses DeepSARM results as input. Iterative fine-tuning of iDeepSARM models is supported by a new CNN-based activity prediction method as well as matrix variants for visualization. Importantly, iDeepSARM is applicable to optimize both R-group combinations and individual scaffolds, which represents a novel feature. For scaffolds consisting of complex ring systems, a variety of structural modifications are

obtained. R-group and scaffold optimization can be carried out in a complementary manner. Instead of activity, other molecular properties can also be optimized including various physico-chemical properties such as hydrophobicity or solubility and ADMET-relevant properties such as membrane permeability or metabolic stability. SARM-based visualization of log-likelihood and property value distributions over multiple iterations makes it possible to monitor progress during optimization in an intuitive manner, as demonstrated for the exemplary design application reported herein. iDeepSARM optimization is possible for compounds with single- or multi-target activity. In addition, when only limited numbers of compounds with activity against a target of interest are available, they can be exclusively used for fine-tuning while pre-training can be carried out with compounds active against related or distantly related targets. Furthermore, iDeepSARM modeling focuses generative compound design on given ASs and their structural features, consistent with requirements of hit-to-lead and lead optimization. Given this focus on existing ASs, synthetic accessibility of newly designed candidates compounds is typically high. Currently, there are only few computational methods available that provide support for lead optimization [6–10]. However, these methods are statistical in nature and/or facilitate SAR visualization, but do not include compound design. From this point of view, iDeepSARM is currently unique in its ability to generate new candidate compounds in an iterative manner and monitor progress in computational optimization, which should be of interest for practical applications in medicinal chemistry and drug design.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2021.100015.

References

[1] Agrafiotis DK, Shemanarev M, Connolly PJ, Farnum M, Lobanov VSSAR Maps. A new SAR visualization technique for medicinal chemists. J Med Chem 2007;50:5926–37.
[2] Peltason L, Bajorath JSAR Index. Quantifying the nature of structure-activity relationships. J Med Chem 2007;50:5571–8.
[3] Guha R, Van Drie JH. Structure-activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 2008;48:646–58.
[4] Renner S, van Otterlo WAL, Seoane MD, Möcklinghoff S, Hofmann B, Wetzel S, et al. Bioactivity-guided mapping and navigation of chemical space. Nat Chem Biol 2009;5:585–92.
[5] Wawer M, Bajorath J. Local structural changes, global data views: graphical substructure-activity relationship trailing. J Med Chem 2011;54:2944–51.
[6] Nicolaou CA, Brown N, Pattichis CS. Molecular optimization using computational multi-objective methods. Curr Opin Drug Discov Develop 2007;10:316–24.
[7] Munson M, Lieberman H, Tserlin E, Rocnik J, Ge J, Fitzgerald M, et al. Lead optimization attrition analysis (LOAA): a novel and general methodology for medicinal chemistry. Drug Discov Today 2015;20:978–87.
[8] 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 2015;59:4235–44.
[9] Maynard AT, Roberts CD. Quantifying, Visualizing, and Monitoring Lead Optimization. J Med Chem 2015;59:4189–201.
[10] Vogt M, Yonchev D, Bajorath J. Computational method to evaluate progress in lead optimization. J Med Chem 2018;61:10895–900.
[11] Wassermann AM, Haebel P, Weskamp N, Bajorath JSAR Matrices. Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. J Chem Inf Model 2012;52:1769–76.
[12] Yoshimori A, Bajorath J. Deep SAR Matrix: SAR Matrix Expansion for Advanced Analog Design Using Deep Learning Architectures. Future Drug Discov 2020;2:FDD36.
[13] Hussain J, Rea C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. J Chem Inf Model 2010;50:339–48.
[14] Gupta-Ostermann D, Shanmugasundaram V, Bajorath J. Neighborhood-Based Prediction of Novel Active Compounds from SAR Matrices. J Chem Inf Model 2014;54:801–9.
[15] 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 2019;4:7061–9.
[16] 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 2015;4:75.
[17] Asawa Y, Yoshimori A, Bajorath J, Nakamura H. Prediction of an MMP-1 Inhibitor Activity Cliff Using the SAR Matrix Approach and its Experimental Validation. Sci Rep 2020;10:14710.
[18] Utomo RY, Asawa Y, Okada S, Yoshimori A, Bajorath J, Nakamura H. Development of Curcumin-Based Amyloid β Aggregation Inhibitors for Alzheimer's Disease Using the SAR Matrix Approach. Bioorg Med Chem 2021;46:116357.
[19] Weininger D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J Chem Inf Comput Sci 1988;28 31-36.
[20] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neur Comput 1997;9:1735–80.
[21] Zheng S, Yan X, Gu Q, Yang Y, Du Y, Lu Y, et al. QBMG: quasi-Biogenic Molecule Generator with Deep Recurrent Neural Network. J Cheminf 2019;11:5.
[22] Sutskever I, Vinyals O, Le QV. Sequence to Sequence Learning with Neural Networks. Adv Neur Inf Proc Sys 2014;1 3104-3112.
[23] Ketkar N. Introduction to keras. In: Deep learning with python. Berkeley, CA: Apress; 2017. p. 97–111.
[24] Yoshimori A. Prediction of Molecular Properties Using Molecular Topographic Map. Molecules 2021;26:4475.
[25] Bishop CM, Svensén M, Williams CKGTM. The Generative Topographic Mapping. Neur Comput 1998;10:215–34.
[26] Gaspar HA. ugtm: a Python Package for Data Modeling and Visualization Using Generative Topographic Mapping. J Open Res Softw 2018;6:26.
[27] Zhong S, Hu J, Yu X, Zhang H. Molecular Image-Convolutional Neural Network (CNN) Assisted QSAR Models for Predicting Contaminant Reactivity toward OH Radicals: transfer Learning, Data Augmentation and Model Interpretation. Chem Eng J 2021;408:127998.
[28] Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: beyond empirical risk minimization. Preprint arXiv:1710.09412 (2017).
[29] Rogers D, Hahn M. Extended-Connectivity Fingerprints. J Chem Inf Model 2010;50:742–54.
[30] Yoshimori A, Bajorath J. Adapting the DeepSARM Approach for Dual-Target Ligand Design. J Comput-Aided Mol Des 2021;35:587–600.
[31] Liu Z, Wang P, Chen H, Wold EA, Tian B, Brasier AR, et al. Drug Discovery Targeting Bromodomain-Containing Protein 4. J Med Chem 2017;60:4533–58.
[32] Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP. The ChEMBL Bioactivity Database: an Update. Nucleic Acids Res 2014;42:D1083–90.
[33] Optuna. A Hyperparameter Optimization Framework. https://github.com/optuna/optuna.
[34] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.
[35] Trott O, Olson AJ. AutoDock Vina: improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. J Comput Chem 2000;31:455–61.
[36] Bjerrum, E.J. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. Preprint arXiv:1703.07076v2 (2017).