Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

Monitoring the Progression of Structure−Activity Relationship Information during Lead Optimization

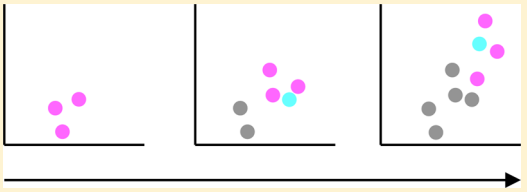
Veerabahu Shanmugasundaram,[†] Liying Zhang,[‡] Shilva Kayastha,[§,∥] Antonio de la Vega de León,[§,∥] Dilyana Dimova,[§] and Jürgen Bajorath*[,§]

[†]Center of Chemistry Innovation & Excellence, WorldWide Medicinal Chemistry, Pfizer PharmaTherapeutics Research & Development, Eastern Point Road, Groton, Connecticut 06340, United States

[‡]Computational Sciences CoE, WorldWide Medicinal Chemistry, Pfizer PharmaTherapeutics Research & Development, 610 Main Street, Cambridge, Massachusetts 06340, United States

[§]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Dahlmannstr. 2, Rheinische Friedrich-Wilhelms-Universität, D-53113 Bonn, Germany

ABSTRACT: Lead optimization (LO) in medicinal chemistry is largely driven by hypotheses and depends on the ingenuity, experience, and intuition of medicinal chemists, focusing on the key question of which compound should be made next. It is essentially impossible to predict whether an LO project might ultimately be successful, and it is also very difficult to estimate when a sufficient number of compounds has been evaluated to judge the odds of a project. Given the subjective nature of LO decisions and the inherent optimism of project teams, very few attempts have been made to systematically evaluate project progression. Herein, we introduce a computational framework to follow the evolution of structure−activity relationship (SAR) information over a time course. The approach is based on the use of SAR matrix data structures as a diagnostic tool and enables graphical analysis of SAR redundancy and project progression. This framework should help the process of making decisions in close-in analogue work.



INTRODUCTION

Lead optimization (LO) aims to transform selected active compounds into clinical candidates through iterative close-in analogue evaluation and is one of the most important challenges in the practice of medicinal chemistry.[1] To date, the multiparametric LO process[1] has been largely driven by a combination of hypotheses and empirical rules that vary based on chemical intuition and experience. The key question faced by medicinal chemists during LO is which compound(s) should be made next, and educated guesses about suitable chemical modifications typically provide the basis for generating analogues and advancing LO projects.

In addition to improving compound potency and selectivity, other properties that are also considered during optimization include solubility, permeability, metabolic stability, and bioavailability. Balancing multiple compound properties in the course of lead optimization is a significant challenge that strongly depends on the specifics of the therapeutic applications and compound classes under study.

Given the multiparametric nature of LO, computational approaches focusing on multiobjective optimization have been developed to aid compound design.[2,3] These methods often employ desirability functions or probability estimates to model and balance multiple drug-relevant properties and select computationally designed candidate compounds with preferred property profiles.[3] However, it is probably fair to say that advanced multiobjective optimization is more popular in library design efforts or in limiting an area of property space on which to focus rather than practical LO, where the pivotal *which compound should be made next* question rules day-to-day efforts.

LO projects often require long periods of time and a large amount of resources. It is not uncommon for hundreds or thousands of compounds to be generated over the course of several years by project teams pursuing multiple lead series, often while facing many roadblocks along the way. In light of this situation, it is difficult to objectively assess LO progression. If a project faces roadblocks, then there is always hope that the next compound(s) might present a breakthrough. This optimism might carry a LO project for a long period of time, and the more time and effort that are expended on it, the more difficult it typically becomes to let go and terminate a project due to limited success. It is therefore not surprising that medicinal chemistry leaders are equally concerned about positive, neutral, or negative project progression and that questions such as how many more compounds do we need to make in close-in analogue space until we reach a go/no-go decision are common place in industry. Accordingly, metrics to assess and quantify LO project progression in a more objective manner are highly desirable. However, only small advances have thus far been made to conceptualize and implement such metrics for the practice of medicinal chemistry.

Special Issue: Computational Methods for Medicinal Chemistry

Received: September 15, 2015

DOI: 10.1021/acs.jmedchem.5b01428
J. Med. Chem. XXXX, XXX, XXX−XXX

Although many computational methods for compound design and activity prediction are available, only very few attempts have been reported to computationally evaluate LO progression, a task that principally differs from compound design. For example, structure−activity relationships (SARs) contained in evolving compound data sets have been monitored in molecular network representations annotated with activity information as well as using three-dimensional activity landscape models.[4] In similarity-based compound networks, positive SAR progression over time is reflected by the formation of compound communities rich in SAR information, whereas lack of progression is indicated by increasing numbers of compounds populating flat SAR regions.[4] Comparison of networks generated at different time points of a project provides a qualitative view of SAR progression. However, the interpretation of SAR networks is not trivial for non-experts.

Furthermore, in a recent investigation, a statistical framework for assessing LO progress has been introduced.[5] For multiple LO parameters, the risk associated with a compound set is quantified from value distributions as the deviation from desired threshold values, and the global risk is obtained by combining all parameter contributions. During the LO process, the risk is expected to be minimized. Risk as a function of (temporal) project progression can be graphically analyzed in different ways, and key compounds making the largest contributions to risk minimization can be identified.[5] Pros of this statistical approach include the ability to monitor multiple properties, individually or in concert, and that it quantifies risk; cons include the requirement of the approach to define property thresholds and that it does not take structural information or relationships as parameters into account (for similarity or diversity assessment, additional computational methods must be employed). Therefore, it is not designed for systematic SAR exploration. In another recent investigation, LO attrition analysis has been introduced[6] to classify compounds according to the number of LO criteria they meet. For this purpose, (project-specific) preferred ranges of numerical properties must be defined and expressed as binary yes/no queries, and the number of compounds meeting an increasing number of queries is determined. Attrition curves are generated by plotting compound count vs parameter count (i.e., x compounds meet y parameters) and used to evaluate LO success.[6] As presented, the approach does not include a temporal component to monitor progress. For a given LO set, the attrition curves are suitable to provide a global view of compound quality. Further analyses performed thus far do not capture the totality of SAR information content for available analogues but, rather, debate the merits of each compound individually.

In this study, we introduce a conceptually different method for the evaluation of SAR progression during LO. The SAR matrix (SARM) data structure[7,8] originally developed for elucidation of SAR patterns in analogue series[7] has been adapted as an indicator of SAR information content for temporal analysis of LO data sets. SARM ensembles are calculated for evolving data sets and scored to quantify their SAR information content. In addition, matrices are classified according to the structural information they capture, which makes it possible to monitor the expansion of existing compound series as well as the introduction of structural novelty during LO in close-in analogue space. SARM distributions are graphically analyzed, and changes in distributions over time reveal SAR progression or a lack of progression. Indicator SARMs can also be annotated with multiple properties, and changes in property profiles can be monitored. Since SARMs exhaustively dissect compound sets in a systematic manner, it is envisioned that the wealth of SAR information during LO might be revealed through an analysis of SARM ensembles over a time course.

EXPERIMENTAL SECTION

SARM Generation. SARMs are generated after subjecting compound sets to two-stage matched molecular pair (MMP) generation.[7,8] A MMP is defined as a pair of compounds that differ only by a structural modification at a single site.[9] MMPs are efficiently generated by systematic fragmentation of exocyclic single bonds in compounds (permitting single, double, and triple cuts) and collection of core structures and associated substituents in index tables.[10]

In the first step, MMPs are generated for all compounds. In the second step, which is uniquely applied for SARMs, all core structures resulting from the first round of fragmentation are again subjected to MMP generation. Compounds forming MMPs from the first step are organized as matching molecular series (MMSs). A MMS is defined as a series of compounds that share the same core and have different substituents at a single site (representing an extension of the MMP concept).[11] It follows that compounds comprising an MMS must form all possible pairwise MMPs. Each MMS is represented as the shared core plus the set of distinguishing substituents. Core MMPs from the second round of fragmentation then identify all structurally analogous cores (differing only by a change at a single site). Each SARM contains a unique subset of MMSs with structurally analogous cores. In the matrix, each row represents an MMS with a unique core (and each column represents a substituent). As a consequence of systematic MMP fragmentation, compounds typically participate in multiple MMSs and occur in multiple SARMs. The ensemble of SARMs generated from a compound set captures all possible analogue relationships. As shown in Figure 1A, SARMs are reminiscent of conventional R-group tables. Each cell represents a unique combination of a core and substituent resulting from the fragmentation (including virtual compounds that have not yet been generated). Cells can be annotated with property information, for example, they can be color-coded according to compound potency, as also illustrated in Figure 1A.

Following the protocol outlined above, SARMs were generated with a Java program utilizing the OEChem toolkit.[12]

SAR Evaluation. The SAR information contained in a SARM was quantified by calculating two different values: the median potency of all compounds comprising the SARM and a matrix-based SAR discontinuity score (Figure 1B). SAR discontinuity is high when structurally similar or analogous compounds have significant potency variations.[13] Such compounds typically reveal SAR information. A SAR discontinuity score quantifying this information was first introduced by systematically accounting for pairwise potency differences between compounds meeting a predefined similarity criterion.[13] For SAR monitoring, we defined a SARM-based discontinuity score (SARM_Disc)

$$\text{SARM\_Disc} = \frac{\sum_i^m \sum_{j,j>i}^m |\text{pot}_i - \text{pot}_j|}{N} \quad \forall\, i, j \rightarrow \text{MMP}$$

where i and j are compounds in a SARM that form an MMP, m is the total number of SARM compounds, N is the total number of MMPs contained in the SARM, pot_i is the potency of compound i, and pot_j is the potency of compound j. For each SARM, the SARM_Disc value was calculated.

Graphical Analysis. SARM distributions were analyzed in scatterplots of median potency vs SARM_Disc scores. In addition, trend plots were generated from SARM distributions to separately monitor the progression of potency and SARM_Disc scores over time. Trend plots were obtained by fitting potency and SARM_Disc values averaged at different time intervals to a linear function.

DOI: 10.1021/acs.jmedchem.5b01428
J. Med. Chem. XXXX, XXX, XXX−XXX

Table 1. ChEMBL Compound Data Sets and SAR Matrices[a]

ID	target name	years	first year		last year	
			no. cpds	no. SARMs	no. cpds	no. SARMs
1908	cytochrome P450 11B1	2006−2013	68	7	464	206
4015	C−C chemokine receptor type 2	2006−2011	124	182	836	1365
344	melanin concentrating hormone receptor 1	2005−2010	259	329	990	1086
3468	caspase-7	2005−2014	61	13	232	125

[a]For each data set, the ChEMBL ID and target name are reported as well as the time period (years) over which the growth of the data set was monitored using SARM ensembles. In addition, the compound composition (no. cpds) and corresponding SARM statistics (no. SARMs) are provided for the first and last years of each time period.

Table 2. Pfizer LO Data Sets and SAR Matrices[a]

LO targets and sets		years	first year		last year	
			no. cpds	no. SARMs	no. cpds	no. SARMs
neurodegenerative	series 1	2010−2014	10	1	431	672
	series 2	2010−2015	46	49	125	128
inflammation	series 1	2011Q1−2012Q3	20	5	88	93
	series 2	2010Q2−2010Q4	18	9	78	43

[a]For each LO set, the time period (years) is reported over which the growth of the corresponding compound series was monitored using SARM ensembles. Q means quarter. In addition, the compound composition (no. cpds) and corresponding SARM statistics (no. SARMs) are provided for the first and last intervals of each time period.

Public Domain Data Sets. Compounds and activity data were taken from ChEMBL[14] (version 20). To assemble data sets evolving over time, compounds for proof-of-concept studies active against human targets at the highest confidence level (ChEMBL confidence score 9) with reported direct binding interactions (ChEMBL relationship type D) and IC_{50} values as potency measurements were considered. For all preselected compounds, publication dates were recorded. A qualifying target-based data set was required to contain compounds reported in increments over a period of at least 5 subsequent years (for each year, a new compound subset had to be available), with a minimum of 50 compounds available in the first year. Four data sets meeting these criteria were assembled, as reported in Table 1.

LO Data Sets. In addition to ChEMBL sets, two LO data sets originating from two different drug discovery projects at Pfizer were studied. Each project team pursued two different chemical series. In each case, one of the series was deemed to be a successful chemical series because the project team was able to identify and nominate preclinical candidate(s), and the second was an unsuccessful series from which no candidate compound was nominated. The first target protein was an enzyme, which was pursued as a biological target for a neurodegenerative indication. The end point for potency in this project was inhibitory activity assessed in a direct enzymatic assay. Although the project team also evaluated other properties during LO, for the purposes of this study, the primary potency end point was used to monitor SAR progression. The second target was also an enzyme, and downregulation of the activity of this enzyme was targeted for an inflammation indication. Also in this case, the end point for potency was inhibitory activity in an enzymatic assay. A series definition used by the project team was added to each compound. IC_{50} values for both projects were converted to logarithmic units. For temporal analysis, dates when compounds were first registered internally were determined and used for monitoring SAR progression. Details of the LO data sets are reported in Table 2.

RESULTS AND DISCUSSION

Concept of Indicator SARMs. SARMs were originally developed for a completely different purpose than for monitoring SAR progression during LO, i.e., to systematically organize analogue series, elucidate SAR patterns for structurally related series, suggest virtual compounds, and predict their activity.[8] In Figure 1A, a small model SARM formed by six compounds (two MMSs) is shown on the left, and a slightly larger SARM (seven compounds, two MMSs) is shown on the right, which also contains a virtual compound (non-colored cell). We reasoned that several characteristics of SARMs might render them suitable for monitoring SAR progression:

(1) SARMs systematically extract all analogue relationships from compound sets. If LO sets contain multiple series, then SARMs not only organize these series as MMSs but also detect all structural relationships among them. Each SARM contains a unique subset of MMSs with related core structures, regardless of the origin of these structural relationships.
(2) SARMs can be easily annotated with compound properties that can then be analyzed based upon the structural organization provided by SARMs.
(3) Depending on the structural relationships contained in a compound data set, varying numbers of SARMs are obtained. This is illustrated in Table 1, which reports compound and SARM statistics for the public domain data sets. Since LO sets are typically centered on single or multiple lead series, they tend to produce large SARM ensembles, thus enabling statistical analysis of SARMs and SARM-associated properties. As a rule-of-thumb, the number of SARMs obtained for structurally homogeneous data sets is often roughly comparable to the number of data set compounds (Table 1).

Given these characteristics, we introduced three modifications to SARMs specifically for the purpose of SAR progression analysis:

(1) SARMs were iteratively calculated for evolving compound data sets at different time points. Thereby, SARM ensembles were obtained that systematically captured all structural relationships between existing and new compounds.
(2) For the analysis of these ensembles, SARMs were classified into three categories including *existing*, *expanded*, and *new* SARMs. Existing SARMs were not modified through the addition of new compounds,

DOI: 10.1021/acs.jmedchem.5b01428
J. Med. Chem. XXXX, XXX, XXX−XXX

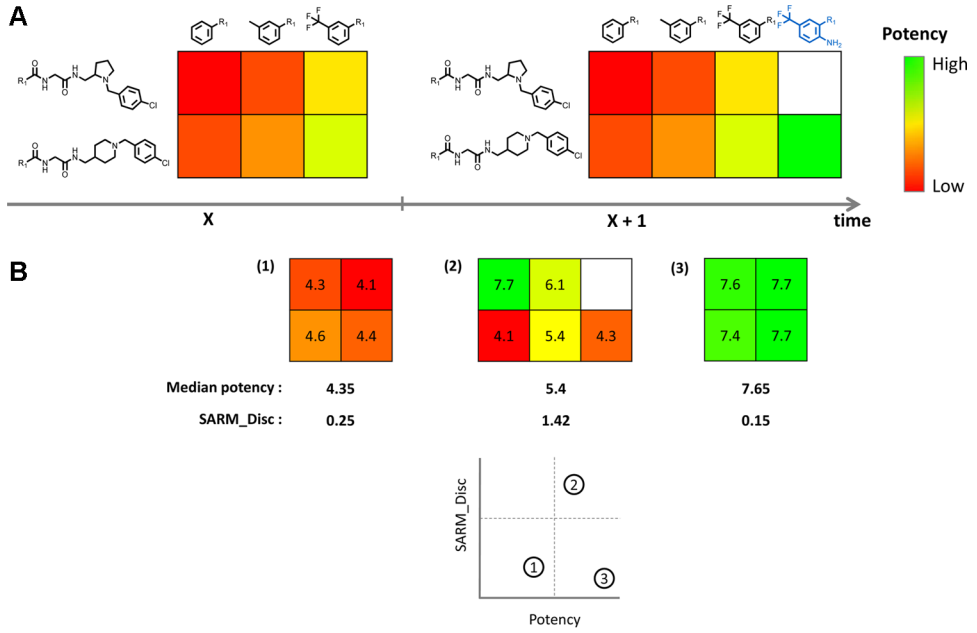


Figure 1. SARM, expansion, and characterization. (A) In the SARM, each row represents a matching molecular series (MMS), i.e., a series of compounds that have a common core (shown left from the row) and are distinguished only by a substituent at a single site (top of each column). Each cell represents an individual compound (unique combination of a core and substituent), either a known data set compound (colored by potency using a continuous spectrum from (lowest) red to (highest) green) or a virtual compound (an as of yet unexplored combination of a core and substituent; non-colored cell). All MMSs contained in a given SARM have related cores that are distinguished only by a structural change at a single site. The matrix on the left was expanded through the addition of a new compound that was detected to match the core of one of the MMSs contained in this matrix. The resulting expanded matrix is shown on the right (the substituent of the new compound is highlighted in blue). (B) Exemplary SARMs with varying SAR information content. SARMs were characterized by calculating their median compound potency and the SARM_Disc score (see text). Accordingly, the SARM_Disc score of a SARM is high if the structurally related compounds comprising the SARM have large potency variations. Therefore, SARM_Disc scores serve as an indicator of SAR information content. As can be seen (and easily rationalized), median potency does not per se correlate with SARM_Disc. The three exemplary SARMs are shown in a scatterplot of median potency vs SARM_Disc. The scatterplot is divided into four quadrants. SARMs with high information, such as matrix 2 in this example, map to the upper right quadrant.

whereas expanded SARMs were obtained when new compounds form structural relationships with already available compounds (as is the case when new analogues are generated for an existing series). Figure 1A illustrates the process of SARM expansion. A new compound complements one of the two MMSs contained in the matrix on the left, leading to the generation of an expanded SARM on the right. Moreover, if newly added compounds introduced structural novelty, i.e., if they formed novel MMSs, then new SARMs were obtained.

(3) For SAR monitoring, SARMs were annotated with two properties, including compound potency and the newly introduced SARM-based SAR discontinuity score (SARM_Disc), as illustrated in Figure 1B. For each SARM, the median potency and the SARM_Disc score were calculated. A high SARM_Disc score indicated the presence of structural analogues with significant potency variations. This situation corresponded to high SAR information content of a SARM because it encoded structural changes that significantly affected potency (different from SARMs that exclusively consisted of weakly or highly potent analogues). Taken together, median potency and SARM_Disc made it possible to prioritize matrices for SAR monitoring. From a SAR information perspective, progress during LO is generally made when SAR-sensitive analogues are obtained including increasingly potent compounds during the course of the project. Following our analysis concept, this is reflected by the generation of SARMs with high median potency and high SARM_Disc scores (as an inflection point during the course of the project), as revealed by time-dependent analysis of matrix distributions.

SARM distributions were recorded in scatterplots of median potency vs SARM_Disc, as schematically represented in Figure 1B (bottom). Preferred SARMs with high median potency and high discontinuity scores mapped to the upper right quadrant of these plots.

The original SARM approach was focused on exploring individual matrices and the compound information that they contained, as discussed above. Because we did not consider the content of individual SARMs for monitoring SAR progression but studied SARM distributions with respect to property values over time, matrix ensembles generated for our current analysis were termed *indicator SARMs*.

Graphical SARM Distribution Analysis. Figure 2A summarizes the principles of time-dependent indicator SARM analysis. SARM ensembles were calculated for an evolving data set following each addition of a compound subset and classified according to the compounds and structural relationships that they captured. The resulting SARM distributions were monitored over time in scatterplots reflecting their SAR information content. Figure 2B shows exemplary progression trends. At the top, positive SAR progression is illustrated. In this case, matrix populations grew over time through the addition of new SARMs and, to a lesser extent, expanded

DOI: 10.1021/acs.jmedchem.5b01428
J. Med. Chem. XXXX, XXX, XXX−XXX

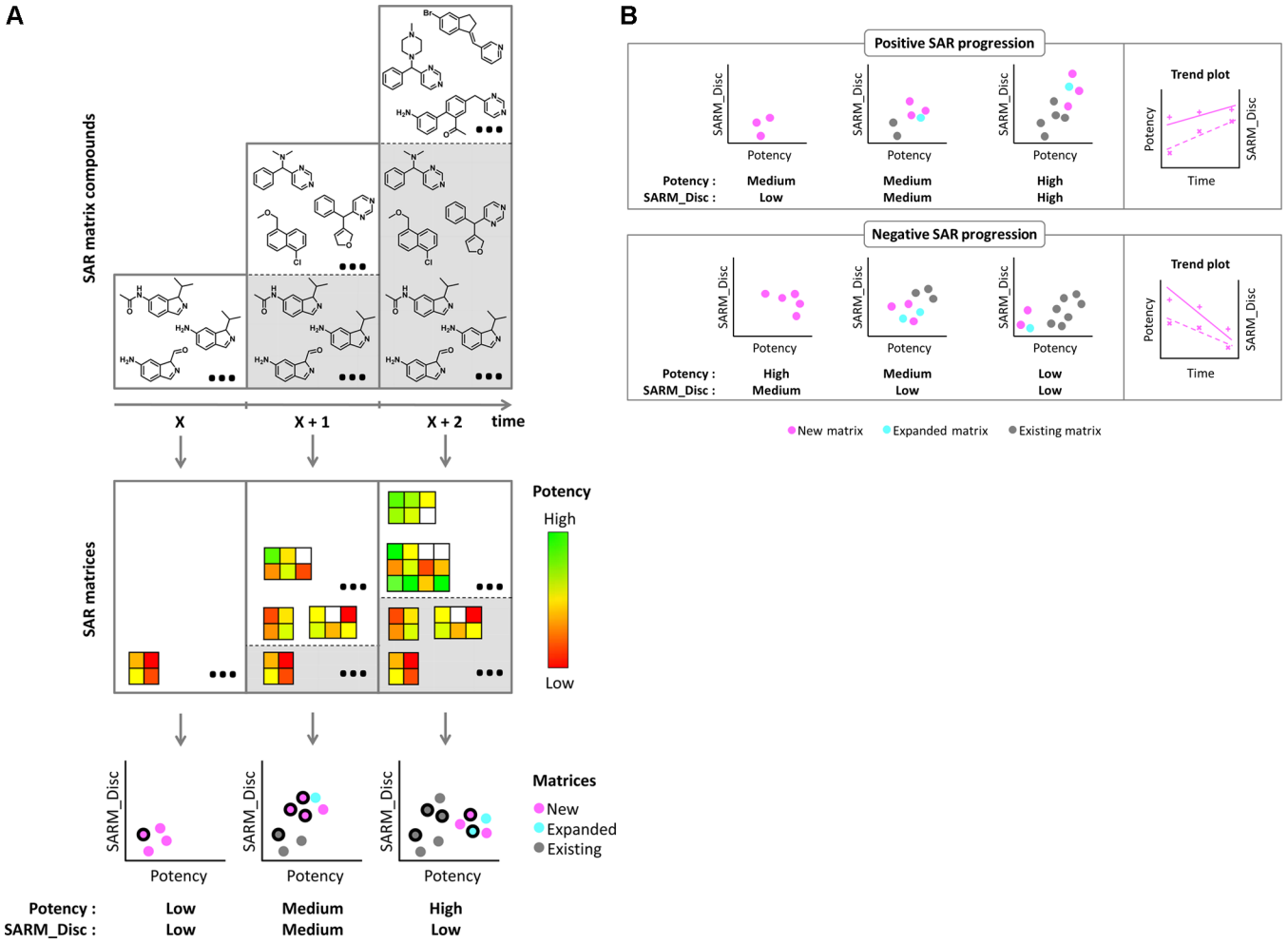


Figure 2. Monitoring SAR progression. (A) Schematic representation illustrating the concept of monitoring SAR progression over time using SARMs. Newly synthesized compounds (shown on a white background) are added in time intervals to evolving lead optimization sets (gray background), and SARMs are systematically calculated at each time point. Matrix representation is according to Figure 1. SARMs calculated at each time point are retained and compared to newly derived matrices. For visualization purposes, not all compounds and SARMs are shown. Distributions of SARMs are monitored in scatterplots of median potency vs SARM_Disc in which each SARM is represented as a color-coded dot. Dots with black border correspond to SARMs shown above the scatterplots. For temporal analysis, three categories of SARMs are distinguished: *existing* (colored gray), *expanded* (cyan), and *new* SARMs (magenta). Existing (old) matrices are not modified through the addition of newly synthesized compounds. Expanded SARMs evolve from existing matrices through the addition of analogues that further extend currently available MMSs. New SARMs contain new MMSs and capture previously unobserved structural relationships due to the addition of novel structures. (B) Two sets of SARM scatterplots are shown and color-coded as in panel (A). Comparison of SARM scatterplots makes it possible to follow SAR progression on a time course and judge the success of lead optimization (LO) efforts. For example, a desirable LO profile (top; positive SAR progression) would display a shift of matrix distributions over time toward the upper right quadrant of the scatterplot (characterized by the presence of high median potency and high SARM_Disc), with an enrichment of new SARMs. By contrast, the scatterplots at the bottom display negative progression of SAR over time because the matrix distribution shifts toward the bottom left quadrant (characterized by the presence of low median potency and low SARM_Disc). On the right, trend plots are shown obtained from indicator SARM distributions by fitting average potency and SARM_Disc scores of new matrices (magenta) for each year to linear functions. Trend lines monitor the development of SARM_Disc and potency for an indicator SARM category over time.

SARMs. A gradual shift of SARM distributions toward the upper right quadrants of the scatterplots was observed, revealing a steady increase in SAR information and the generation of increasingly potent compounds. By contrast, the example at the bottom illustrates (undesired) negative SAR progression characterized by the occurrence of expanded and new SARMs with low median potency and low discontinuity scores and the absence of an upward shift of SARM distributions over time. Positive and negative SAR progressions can also be visualized in trend plots (shown on the right of Figure 2B) that are derived from the SARM distributions by fitting linear models and separately monitoring potency and SARM_Disc progression over time. The trend lines were fitted to data averaged over time intervals. Ideally, in the case of positive SAR progression, these trend lines should have positive slopes.

Monitoring SAR Progression. Applying the approach summarized in Figure 2, SAR progression was monitored for different types of compound sets.

Public Domain Compound Sets. The four compound data sets from ChEMBL represented prototypic compound sets evolving over time and were generated to mimic LO sets by combining compounds active against different targets taken from the scientific literature (only high-confidence activity data

DOI: 10.1021/acs.jmedchem.5b01428
J. Med. Chem. XXXX, XXX, XXX−XXX

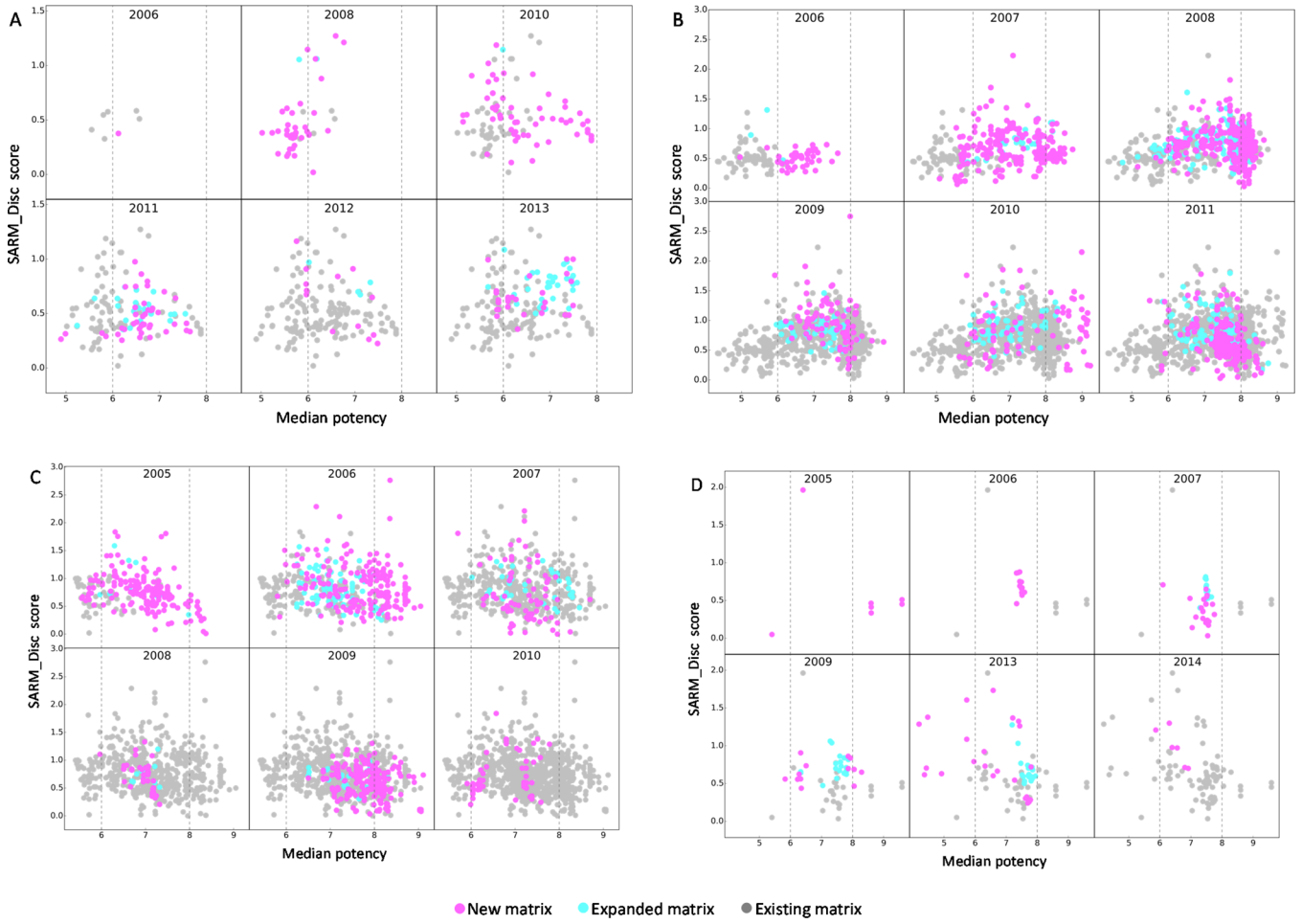


Figure 3. Indicator SARM distributions over a time course. Scatterplots are shown for four public domain data sets that were incrementally assembled over different years on the basis of compound publication dates. The SARM representation is according to Figure 2. In addition, dotted lines at potency values of six and eight log units differentiate SARMs with high, intermediate, or low median potency. (A) Cytochrome P450 11B1 inhibitors, (B) C−C chemokine receptor type 2 ligands, (C) melanin-concentrating hormone receptor 1 ligands, and (D) caspase-7 inhibitors. Compound and SARM statistics for the monitored time periods are provided in Table 1. We note that active compounds were available in each case prior to the first year monitored in a scatterplot. For compounds available in the preceding year, SARMs were calculated and used as a reference ensemble to generate classified SARMs for the first year of the monitored period.

were taken into consideration for compound selection). Because selected compounds originated from a variety of literature sources, these sets were structurally more heterogeneous than typical LO sets, thus presenting a challenge for a proof-of-concept assessment of indicator SARM analysis. These four data sets are made freely available as an open-access deposition.[15]

Figure 3 shows the distribution of indicator SARMs obtained from the data sets over a period of six subsequent years. The median potency and SARM_Disc scores of SARMs were plotted and colored according to their matrix category.

Figure 3A reports the temporal analysis of inhibitors of cytochrome P450 11B1. This set contained 464 compounds but yielded only 206 SARMs (Table 1), indicating structural heterogeneity. Nonetheless, interesting SAR trends were detected. From 2006 to 2011, added inhibitors often represented new analogue series (MMSs), resulting in a gradual increase in the number of new SARMs (magenta) during this period. In 2011 and especially 2013, a larger number of expanded SARMs (blue) was observed, indicating follow-up investigations on existing series. Between 2011 and 2013, a shift of expanded and new SARMs toward the upper right quadrant of the plots was observed, revealing overall promising SAR progression.

The set of C−C chemokine receptor type 2 ligands in Figure 3B was much larger (836 compounds) than the cytochrome P450 11B1 inhibitor set and ultimately yielded 1365 SARMs (resulting in high-density scatterplots). Between 2006 and 2008, a shift of the SARM distributions toward the right of the plots was observed. During subsequent years, the distributions became increasingly dominated by a large number of new SARMs with high median potency (in addition, SARM expansion was also observed). Thus, many novel series containing highly potent compounds became available, reflecting successful compound design efforts. A different picture emerged for ligands of melanin-concentrating hormone receptor 1 in Figure 3C, the largest data set (990 compounds) producing 1086 SARMs. In 2005, the distribution was dominated by new SARMs (resulting from structurally novel compounds not available during the preceding year). In 2006, many SARMs were expanded, reflecting follow-up chemistry efforts, and the distribution shifted toward higher potency and discontinuity scores, indicating SAR progression. However, during 2007 and 2008, the number of new and expanded

DOI: 10.1021/acs.jmedchem.5b01428
J. Med. Chem. XXXX, XXX, XXX−XXX

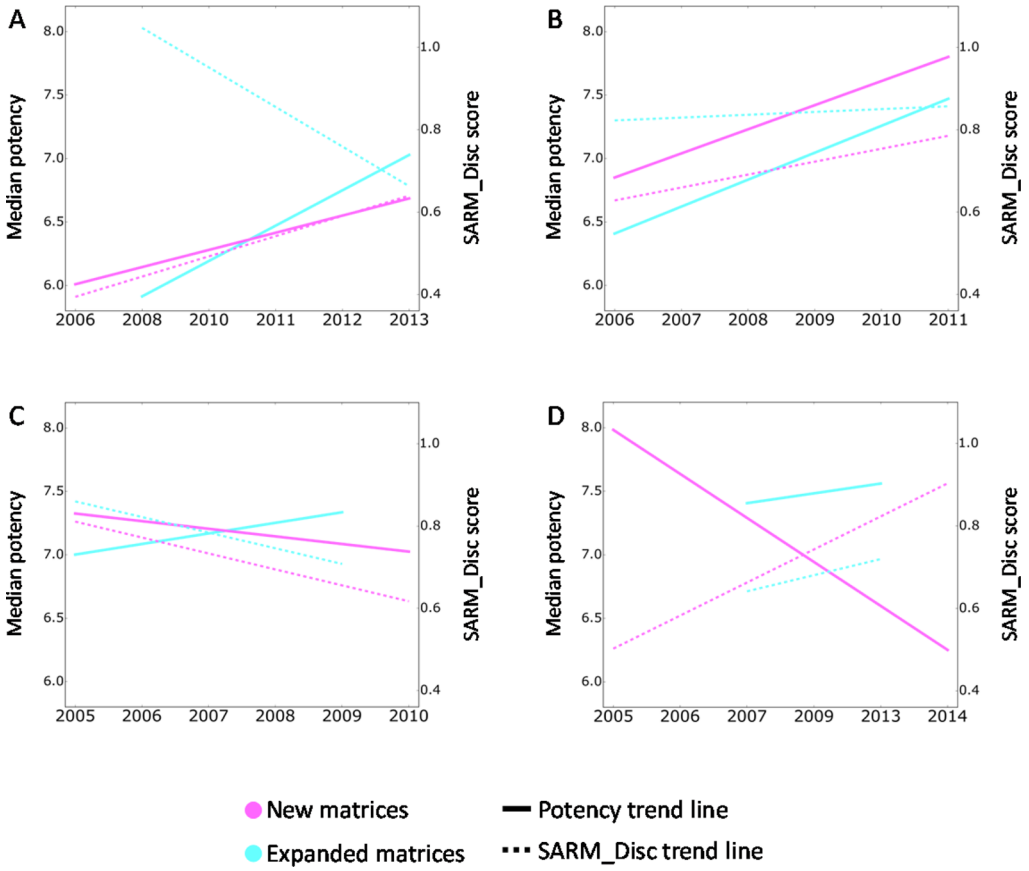


Figure 4. Trend plots of expanded and new indicator SARMs according to Figure 2B derived from the data distributions in Figure 3. (A) Cytochrome P450 11B1 inhibitors, (B) C−C chemokine receptor type 2 ligands, (C) melanin-concentrating hormone receptor 1 ligands, and (D) caspase-7 inhibitors. Trend lines separately monitor the development of median potency and SARM_Disc scores over time for a given category of indicator SARMs.

SARMs declined, indicating reduced chemistry efforts. Another boost in novel active compounds was detected in 2009, which further increased median potency. However, there was essentially no matrix expansion in 2010, and the number of new SARMs also declined again. Hence, in this case, different intervals of strong and weak SAR progression were detected. Figure 3D monitors the smallest of the four data sets, consisting of 232 inhibitors of caspase-7, that yielded a total of only 125 SARMs. Although the number of SARMs was small in this case, their temporal distributions revealed an obvious trend. During 2005 and 2006, a limited number of inhibitors and SARMs became available, and expanded SARMs were first detected in 2007. However, between 2009 and 2014, an increasing number of SARMs was found to map to the upper left quadrant of the plots, characterized by the presence of low median potency and high discontinuity, resulting from the addition of more and more weakly potent compounds to a small number of highly potent ones. Thus, in this case, negative SAR progression was observed.

Figure 4 reports trend plots for new and expanded SARMs generated from the distributions in Figure 3. Especially for very large SARM ensembles, trend lines that separately monitor potency and discontinuity help to better understand characteristics of SAR progression, although they are only approximate. Figure 4A confirms the conclusions drawn from SARM distribution analysis for the cytochrome P450 11B1 inhibitor set. The median potency and discontinuity score of new SARMs were increasing, and potency of expanded SARMs also increased. The only exception to overall positive SAR progression was the observed decrease in discontinuity of expanded SARMs, which likely resulted from the increasing number of analogues of existing series having comparable potency. Furthermore, Figure 4B also reveals a clear example of positive SAR progression, consistent with SARM distribution analysis, for the large set of C−C chemokine receptor type 2 ligands. In this case, median potency and discontinuity increased for all SARMs or remained essentially constant at a high level (i.e., discontinuity of expanded SARMs). Figure 4C reflects overall limited SAR progression for the set of melanin-concentrating hormone receptor 1 ligands, as discussed, and Figure 4D displays negative trends for caspase-7 inhibitors. Here, a strong decline of median potency was detected for new SARMs, which was accompanied by an increase in discontinuity. Although this observation might be puzzling at a first glance, it can be easily rationalized as resulting from the presence of analogues with decreasing potency in SARMs also containing highly potent compounds. Furthermore, for a small number of expanded SARMs, potency increased only slightly and discontinuity remained at a low level.

Taken together, temporal distribution analysis of indicator SARMs from exemplary target-based compound sets evolving over time detected clear differences in SAR progression, hence providing support for the underlying methodological concept. Next, actual LO data sets originating from drug discovery were investigated. Such data sets are currently not available in the public domain.

DOI: 10.1021/acs.jmedchem.5b01428
J. Med. Chem. XXXX, XXX, XXX−XXX

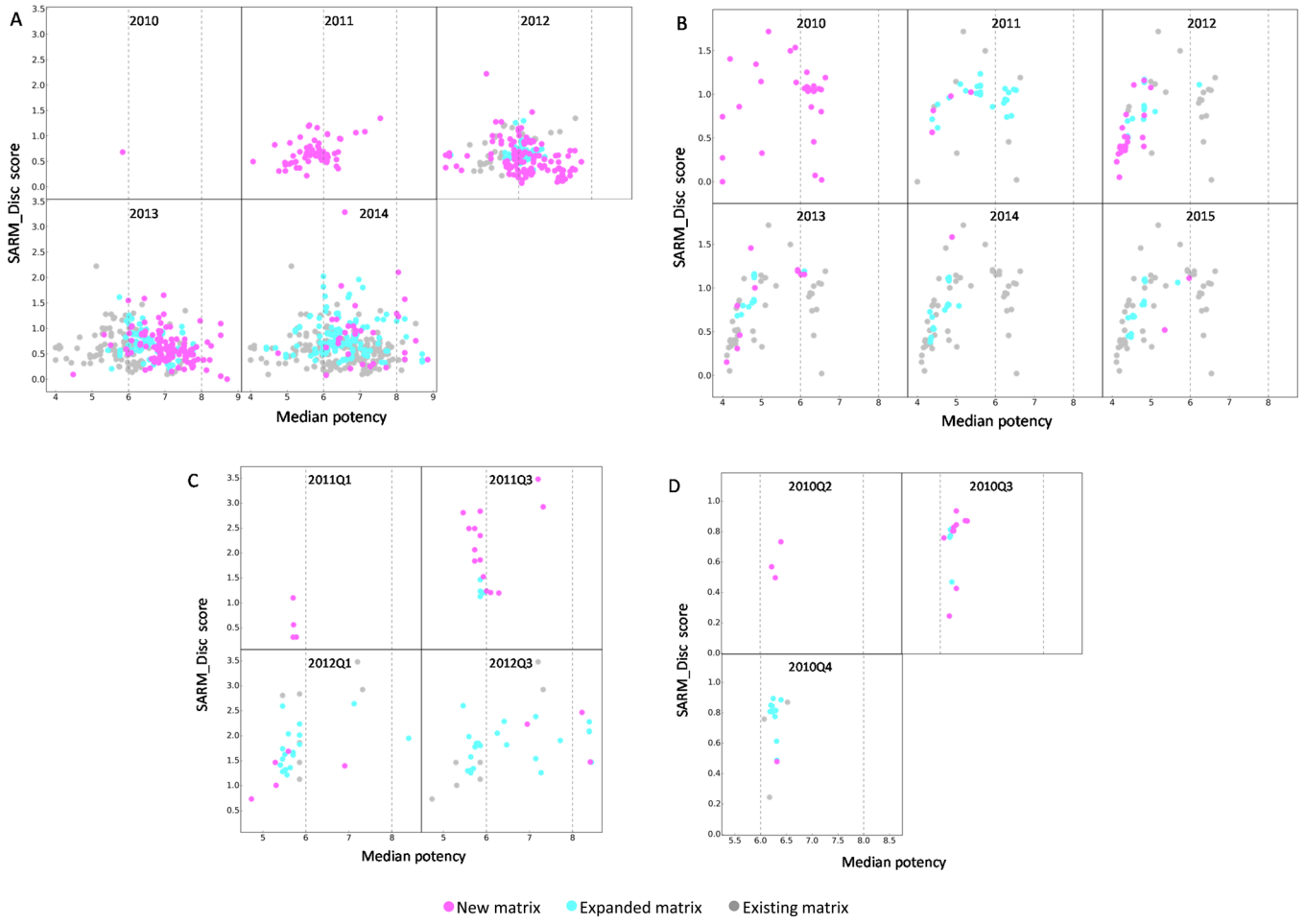


Figure 5. Indicator SARM distributions over a time course for LO sets. Scatterplots are shown for two LO data sets that were assembled from Pfizer project team data on the basis of project progression information. (A) Neurodegenerative target, series 1, (B) neurodegenerative target, series 2, (C) inflammation target, series 1, and (D) inflammation target, series 2. Compound and SARM statistics for the monitored time periods are provided in Table 2. Series 1 in (A) and (C) represented successful project progressions from which compounds were nominated as candidates for preclinical studies. By contrast, series 2 in (B) and (D) represented unsuccessful project progressions from which no compounds were nominated.

LO Data Sets. Two LO sets from different Pfizer drug discovery projects were investigated. Each project team pursued two different chemical series per target. In each case, one of the series was considered to be successful because the project team was able to nominate preclinical candidate(s) from this series, and the other series was unsuccessful, yielding no candidate compounds. Table 2 provides a description of these LO sets.

Figure 5 shows the distribution of indicator SARMs obtained over a period of 4 to 5 years for the neurodegenerative target and 3 to 7 quarters for the inflammation target. Figure 5A monitors the SAR progression of series 1 of the neurodegenerative project. This set ultimately yielded 672 SARMs for 431 compounds (Table 2), indicating structural homogeneity. In 2010, LO efforts on this series started with 10 analogues active in the micromolar range contained in a single SARM. Figure 5A reveals that there was consistent positive SAR progression for series 1. Starting in 2012, new and expanded SARMs were detected, and there were clear breakthroughs in 2013 and 2014, yielding highly potent compounds in increasingly informative SAR environments. On the basis of SAR monitoring, LO on series 1 was a highly promising project, consistent with its ultimate success. Similar trends were not observed for series 2 in Figure 5B, although there was much more compound and SAR information available initially than that for series 1. LO efforts on series 2 started with 46 compounds, and a total of 125 inhibitors were evaluated over a period of 6 years. However, the project team was unable to break a potency barrier with this chemical series. Although matrix expansion occurred during the first 3 years, no notable SAR progression was detected, and in 2014, it was evident that the LO project faced a roadblock.

The comparably small series of inflammation inhibitors in Figure 5C,D with, ultimately, 88 and 78 compounds, respectively, also exhibited rather different SAR progression. Series 1 in Figure 5C displayed very positive SAR trends with significantly increasing SAR information content and compound potency already detectable during the first two time intervals. By contrast, very little SAR progression was observed for series 2 in Figure 5D from the second to the third quarter of 2010, but no further progression was observed during the fourth quarter. Thus, SAR monitoring contrasts these two series of inflammation inhibitors, and it is easy to reconcile why series 1 was ultimately successful and series 2 was not.

The trend plots for these LO sets in Figure 6 strongly support conclusions drawn from indicator SARM distribution analysis. The successful series 1 of neurodegeneration inhibitors in Figure 6A and inflammation inhibitors in Figure 6C displayed an increase in all trend lines for new and expanded

DOI: 10.1021/acs.jmedchem.5b01428
J. Med. Chem. XXXX, XXX, XXX−XXX

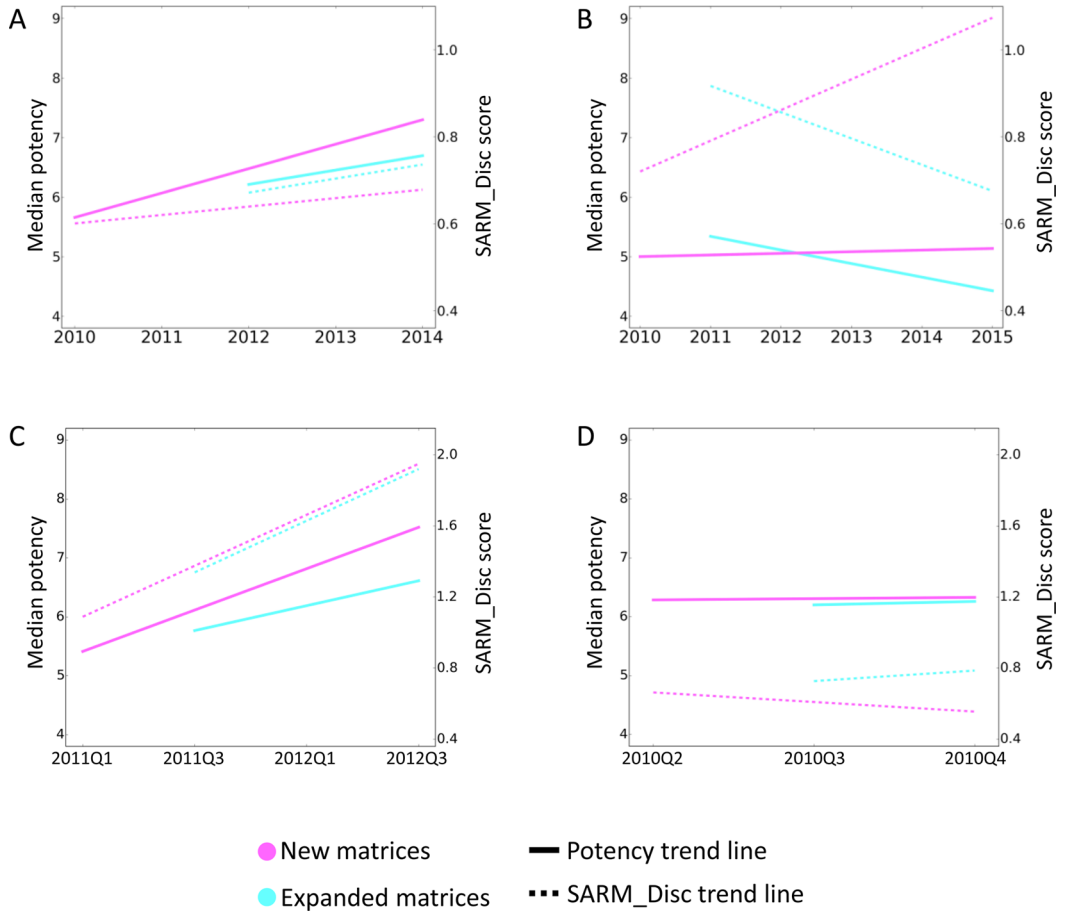


Figure 6. Trend plots for LO sets showing expanded and new indicator SARMs derived from the data distributions in Figure 5. (A) Neurodegenerative target, series 1, (B) neurodegenerative target, series 2, (C) inflammation target, series 1, and (D) inflammation target, series 2. Trend lines separately monitor the development of median potency and SARM_Disc scores over time for a given category of indicator SARMs. Series 1 in (A) and (C) represented successful chemical series and displayed positive SAR progression with an increase in both median potency and SARM_Disc scores. Series 2 in (B) and (D) represented unsuccessful chemical series, which displayed negative SAR progression for expanded SARMs with a decrease in median potency and SARM_Disc scores and essentially flat SARs for new SARMs.

matrices. By contrast, the unsuccessful series 2 of neurodegeneration inhibitors in Figure 6B was characterized by decreasing trend lines for expanded matrices, reflecting negative SAR progression of close-in analoging attempts and diverging trend lines for new matrices, with an increase in SAR information content resulting from the addition of new but only weakly potent compounds that could not be further optimized. Moreover, the series 2 of inflammation inhibitors in Figure 6D displayed essentially flat SAR characteristics throughout.

On the basis of the comparisons reported in Figures 5 and 6, successful LO series of neurodegeneration and inflammation inhibitors were clearly distinguished from unsuccessful series. Analysis of indicator SARM distributions would have made it possible to predict the lack of SAR progression for the latter series during the course of LO.

CONCLUSIONS

Lead optimization is a largely hypothesis-driven process that depends mainly on medicinal chemistry experience and intuition. Only few efforts have thus far been made to rationalize this process and assess LO progress. Efforts in this direction are highly desirable to support decision making because it is very difficult to predict the ultimate outcome of LO campaigns and control the number of compounds to be evaluated before meaningful conclusions can be reached. In this study, we have introduced a computational framework to monitor the progression of SAR information content during LO over a time course. The SAR matrix data structure, which was originally developed for a completely different purpose, i.e., the elucidation of SAR patterns in related analogue series and compound prediction, was adapted as a diagnostic tool to evaluate SAR progression. This was accomplished by generation of SARM ensembles for compound sets evolving over time, classification of SARMs based on the compounds they contain, and characterization of their SAR information content. SAR information contained in individual SARMs was quantified on the basis of a newly introduced matrix discontinuity score combined with median potency calculations. Characteristic shifts of SARM ensembles in scatter plots were found to indicate positive, neutral, or negative SAR progression and revealed significant differences between target-based compound sets. Analysis of SARM distributions was complemented by trend plots designed to summarize SAR progression over time. Our proof-of-concept investigations show that SARM ensembles are capable of detecting differences in SAR progression in compound sets of distinct composition. As a diagnostic tool, they can be used to distinguish SAR progression from redundancy, i.e., when increasing numbers of

DOI: 10.1021/acs.jmedchem.5b01428
J. Med. Chem. XXXX, XXX, XXX−XXX

compounds are made that do not add novel SAR information or further improve potency. Application of the approach to actual LO sets from drug discovery projects revealed very clear SAR trends over time for series that were ultimately successful or unsuccessful. Such insights are valuable in project decision making. Taken together, the results reported herein suggest that indicator SARMs should merit further investigation in LO assessment. Since the SARM data structure can be easily annotated with different molecular properties, multiple parameters can be monitored.

AUTHOR INFORMATION

Corresponding Author

*Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Author Contributions

‖S.K. and A.d.l.V.d.L. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The coauthors of the University of Bonn would like to thank OpenEye for providing an academic license. For the two Pfizer LO sets, a waiver on ACS data deposition requirements has been granted.

ABBREVIATIONS USED

LO, lead optimization; MMP, matched molecular pair; MMS, matching molecular series; SAR, structure−activity relationship; SARM, SAR matrix

REFERENCES

(1) *The Practice of Medicinal Chemistry*, 3rd ed.; Wermuth, C. G., Ed.; Academic Press: Boston, MA, 2008.

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

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

(4) Iyer, P.; Hu, Y.; Bajorath, J. SAR Monitoring of Evolving Compound Data Sets Using Activity Landscapes. *J. Chem. Inf. Model.* **2011**, *51*, 532−540.

(5) Maynard, A. T.; Roberts, C. D. Quantifying, Visualizing, and Monitoring Lead Optimization. *J. Med. Chem.* **2015**, DOI: 10.1021/acs.jmedchem.5b00948.

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

(7) Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. *J. Chem. Inf. Model.* **2012**, *52*, 1769−1776.

(8) Gupta-Ostermann, D.; Bajorath, J. The 'SAR Matrix' Method and its Extensions for Applications in Medicinal Chemistry and Chemogenomics. *F1000Research* **2014**, *3*, 113.

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

(10) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.

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

(12) *OEChem TK*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2012.

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

(14) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(15) Shanmugasundaram, V.; Zhang, L.; Kayastha, S.; de la Vega de León, A.; Dimova, D.; Bajorath, J. Data Sets for SAR Progression Analysis. *Zenodo* **2015**, DOI: 10.5281/zenodo.32794.

DOI: 10.1021/acs.jmedchem.5b01428
J. Med. Chem. XXXX, XXX, XXX−XXX