Computational Assessment of Chemical Saturation of Analogue Series under Varying Conditions

Dimitar Yonchev, Martin Vogt, Dagmar Stumpfe, Ryo Kunimoto, Tomoyuki Miyao, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany

Supporting Information

ABSTRACT: Assessing the degree to which analogue series are chemically saturated is of major relevance in compound optimization. Decisions to continue or discontinue series are typically made on the basis of subjective judgment. Currently, only very few methods are available to aid in decision making. We further investigate and extend a computational concept to quantitatively assess the progression and chemical saturation of a series. To these ends, existing analogues and virtual candidates are compared in chemical space and compound neighborhoods are systematically analyzed. A large number of analogue series from different sources are studied, and alternative chemical space representations and virtual analogues of different designs are explored. Furthermore, evolving analogue series are distinguished computationally according to different saturation levels. Taken together, our findings provide a basis for practical applications of computational saturation analysis in compound optimization.

1. INTRODUCTION

In medicinal chemistry, compound optimization relies on the generation of analogues to explore structure–activity relationships (SARs) and improve molecular properties. Chemical optimization is largely driven by intuition and experience. The optimization process is difficult to rationalize and formalize, and consequently, subjective criteria dominate decision making. In particular, it is very difficult to determine when a sufficient number of analogues have been generated and no further progress can be expected. Series progression can be evaluated on the basis of SAR features and/or chemical saturation. Both criteria go hand in hand but provide somewhat different perspectives. From an SAR viewpoint, the central question is whether or not compound potency and other relevant properties can be further improved by generating additional compounds. Chemical saturation, on the other hand, primarily addresses the question whether the chemical space around active compounds has been sufficiently covered to ensure that no potential optimization pathways remain unexplored.

So far, only a few approaches have been introduced to evaluate the generation of analogue series. These approaches include multiparameter optimization (MPO) using desirability functions, attrition curves, or risk statistics, which balance multiple compound properties. However, MPO does not directly assess progression saturation of analogue series, but it prioritizes candidate compounds with desirable property profiles. MPO can hence be applied to indirectly evaluate series progression by prioritizing candidate compounds with preferred property combinations that can still be obtained. Regardless of the MPO approach that is applied, a pool of candidates for evaluation must be generated separately. In addition, numerical and graphical SAR analysis methods have been introduced to monitor SAR progression of evolving compound series and evaluate whether newly generated analogues yield further SAR information. These approaches may also suggest preferred candidate compounds or provide diagnostics for SAR landscapes of compound data sets.

Previously, we have introduced a computational concept to more directly and quantitatively assess progression saturation of analogue series using virtual analogue populations that are mapped into chemical reference space together with exiting analogues. The approach is focused on the assessment of chemical progression saturation, but not SAR progression, and addresses the following questions: Do existing analogues provide sufficient chemical space coverage? Are virtual candidate compounds available that have a high likelihood of activity? As discussed in detail below, the methodology requires the use of chemical reference spaces and virtual candidate compounds to quantify chemical saturation. The degree to which the results of computational saturation analysis depend on such parameters is yet to be explored.

Received: August 18, 2018
Accepted: November 8, 2018
Published: November 20, 2018
Therefore, we have further investigated and extended the methodology for medicinal chemistry applications by analyzing many analogue series of different compositions, exploring alternative chemical space representations, and virtual analogues resulting from different design strategies. Therefore, critical computational parameter settings underlying the analysis were assessed. We also show that large series with similar numbers of analogues have different saturation characteristics. Distinguishing between different saturation levels relies on computational analysis, as demonstrated herein.

2. MATERIALS AND METHODS

2.1. Methodology. 2.1.1. Concept. Progression saturation of the analogue series is evaluated by comparing distributions of existing analogues and virtual candidate compounds for series expansion in chemical space, applying a neighborhood concept. First, chemical space coverage of the analogue series is estimated by determining the proportion of virtual candidates falling into predefined neighborhoods of existing analogues. For this purpose, a global saturation score is calculated, as defined below. In this case, chemical neighborhood of analogues are defined on the basis of nearest-neighbor distances between virtual compounds to measure global coverage of the chemical space. Second, the population of neighborhoods of active analogues (active neighborhoods) is assessed by determining virtual candidates falling into active neighborhoods. Therefore, neighborhoods are defined differently on the basis of median distances between active analogues.

Figure 1. Saturation scores and categorized combinations. (a, b) Calculation of the raw global and local saturation scores is illustrated, respectively. The coordinate system represents a chemical reference space containing an analogue series and virtual candidates. "D" stands for descriptor. Each descriptor adds a dimension to the space. (c) Combinations of global and local saturation scores are categorized as indicators of different progression saturation stages. Figure panels were adapted from ref 6.
compounds in the chemical space and a local saturation score is calculated, as also defined below. Relating global and local compound distributions and resulting scores to each other makes it possible to evaluate progression saturation of the analogue series. The assessment is primarily focused on chemical progression saturation (rather than SAR progression), addressing the questions whether chemical space coverage by existing analogues is extensive and, in addition, whether a significant number of virtual candidates exist that are likely to be active.

2.1.2. Scoring Scheme. A quantitative measure of progression saturation is obtained by calculating two scores and relating them to each other.6 The raw global saturation score is defined as the ratio of the number of virtual candidate compounds that fall into neighborhoods of experimental analogues relative to the total number of virtual compounds

\[ S = \frac{|V_{\text{Exp}}|}{|V|} \]

S denotes the set of experimental analogues, V denotes the set of virtual analogues, and \( V_{\text{Exp}} \) denotes the set of virtual compounds falling into neighborhoods of experimental analogues. The neighborhood radius is determined as follows. For each virtual analogue, mean Euclidian distances to the top 1% of its nearest virtual neighbors are calculated. The median of these distances is used as the neighborhood radius. Accordingly, only closely related virtual candidates map to the same neighborhood. The top 1% of nearest virtual neighbors were selected on the basis of initial test calculations in which the percentage of nearest virtual neighbors was systematically varied. For different percentages, comparable scores were obtained and the top 1% were selected to control the number of distance calculations.

The so-defined raw global saturation score, the calculation of which is illustrated in Figure 1a, measures the chemical space coverage of existing analogues and virtual candidates. Hence, the larger the score, the more virtual analogues map to active neighborhoods. A Laplace-like correction is applied to the denominator to avoid numerical instability when \( V_{\text{active}} \) is small or 0. For calculating the raw local saturation score, the neighborhood radius of each active analogue is set to the median value of pairwise distances.

\[ A = \frac{|A|}{|V_{\text{active}}| + 1} \]

A is a set of active compounds and \( V_{\text{active}} \) is a set of virtual candidates in active neighborhoods. A Laplace-like correction by adding 1 is applied to the denominator to avoid numerical instabilities when \( V_{\text{active}} \) is small or 0. For calculating the raw local saturation score, the neighborhood radius of each active analogue is set to the median value of pairwise distances between active analogues. Hence, the size of the so-defined neighborhood accounts for typically observed distances between active analogues in chemical space. The raw local saturation score, the calculation of which is illustrated in Figure 1b, measures the distribution of virtual analogues around active analogues. The larger the score, the less populated are the neighborhoods of active analogues with virtual candidates.

The raw global scores and logarithmically transformed local saturation scores are converted into conventional Z-scores on the basis of the mean and standard deviation of the score population of large sets of analogue series. Accordingly, the mean of the score population was subtracted from each raw score and the difference was divided by the standard deviation of the distribution, yielding the Z-score. Accordingly, the resulting Z-scores have a mean of 0 and standard deviation of 1. Global and local scores generally display low correlation.

2.1.3. Score Combinations. Combinations of global and local saturation scores can be divided into four categories that characterize different levels of saturation progression,6 as schematically shown in Figure 1c: (1) low global and high local scores (category low/high, upper left quadrant) characterize the series that have low analogue coverage of chemical reference space and only few virtual candidates in active neighborhoods. Hence, these series are only little explored and thus rationalized as early-stage series. (2) Category low/low (lower left quadrant) describes the series with low chemical space coverage by experimental analogues but with many virtual candidates located in active neighborhoods. Such series are more advanced chemically and rationalized as mid-stage series. (3) Category high/low (lower right) identifies the series with more extensive coverage of experimental analogues and many virtual candidates that are present in active neighborhoods. Such series are better characterized as late-stage series, which approach saturation. (4) Category high/high (upper right) characterizes series with extensive analogue coverage but only few remaining virtual candidates in active neighborhoods, thus indicating a high level of chemical saturation (saturated series).

Threshold values for categorization of Z-score combinations are set to 1, i.e., one standard deviation (\( \sigma \)) above the mean of the fitted normal distribution of Z-scores for sets of analogue series.

2.2. Analogue Series. Analogue series were assembled using a computational method that identifies analogue series in compound data sets of any composition.9 This method makes use of the matched molecular pair (MMP) concept.10 An MMP is defined as a pair of compounds that are distinguished by a chemical modification at only a single site.16 This modification involves the exchange of a pair of substructures, which is termed as transformation.11 For our study, MMPs were generated by single-cut fragmentation of exocyclic single bonds11 on the basis of retrosynthetic rules,12 yielding RECAP-MMPs.13

Comounds sharing the same RECAP-MMP core form a matching molecular series (MMS),14 which represents an analogue series with a single substitution site.9 By contrast, different MMSs sharing analogues form a series with multiple substitution sites.9 In this case, substitution sites of the corresponding RECAP-MMP cores are transferred to shared analogues from which an analogue-series-based (ASB) scaffold15,16 with multiple substitution sites16 is extracted. This ASB scaffold then represents an analogue series with multiple substitution sites.

Two sets of analogue series were extracted from screening data and medicinal chemistry sources, respectively.

2.2.1. Series of Screening Compounds. A set of 80 series containing active and inactive analogues was extracted from the PubChem Bioassay database.17 The series were required to have single substitution sites, consist of at least 30 analogues tested in the same assay, and include at least three active compounds. They contained a total of 1618 compounds and covered 25 biochemical assays and 23 unique targets. These series were used in the proof-of-concept study introducing
computational progression saturation analysis. They are termed PubChem series in the following.

### 2.2.2. Series from Medicinal Chemistry

A set of 64 analogue series yielding ASB scaffolds with multiple substitution sites were extracted from ChEMBL (release 23) on the basis of high-confidence activity data. Accordingly, only compounds with direct interactions (type “D”) with human targets at the highest assay confidence level (confidence score 9) were selected. As potency measurements, only specified equilibrium constants ($K_i$ values) or $IC_{50}$ values were

![Figure 2. Representative analogue series. Exemplary compounds of series from (a) PubChem and (b) ChEMBL are shown with associated target and potency information and corresponding virtual analogues (red). In addition, the composition of each series is reported and its ASB scaffold is shown.](image-url)
considered. These analogue series, termed ChEMBL series in the following, were required to consist of 10–30 active analogues. They contained a total of 1422 compounds and covered 62 unique targets.

In addition, three large series with 100 or more analogues were extracted from ChEMBL to model evolving series and analyze their saturation characteristics.

Compositions of all analogue series, their structures, and targets are reported in Table S1 in the Supporting Information. Figure 2 shows exemplary compounds from representative series with associated activity and target information together with virtual analogues. Because our approach is designed to assess chemical saturation of analogue series, rather than SAR progression, potency and target information for analogues is not of primary relevance. As discussed, compound potency is not a computational parameter here.

2.3. Chemical Reference Spaces. Two overlapping chemical reference spaces were generated using descriptor subsets of different designs and information contents. Because chemical reference space is a variable for saturation analysis, investigating the influence of alternative chemical space representations is an important aspect of our study. The first space representation was 7-dimensional and formed by intuitive molecular descriptors including molecular weight, the number of hydrogen-bond donor and acceptor atoms, the number of rotatable bonds, the logarithm of the octanol/water partition coefficient \((\log P)\), aqueous solubility, and topological polar surface area. These descriptors accounted for chemical features known to be relevant to ligand—target interactions. In addition, a 14-dimensional reference space was generated by adding seven more abstract two-dimensional descriptors with little pairwise correlation to the initial set. Selected descriptors (Table 1) were calculated with the Molecular Operating Environment (MOE)\(^{19}\) and scaled to zero mean and unit variance.

2.4. Virtual Analogues. Two conceptually different strategies were applied to generate virtual analogues for series including a transformation- and a matrix-based approach. Virtual analogues served as candidates for series progression.

2.4.1. Transformation-Based Virtual Candidates. Systematic RECAP-MMP fragmentation, as described above, was applied to ChEMBL compounds with high-confidence activity data to sample chemical transformations from which individual substituent fragments were extracted. A total of 13 203 unique substituents were obtained. These substituents were systematically recombined with the RECAP-MMP core of each screening analogue series, yielding a constant number of 13 203 virtual analogues per series. This strategy was applied in our initial study.\(^{6}\) In addition, the substituent pool was recombined with ASB scaffolds representing medicinal chemistry series with multiple substitution sites to randomly sample the same number of virtual candidates per series. Transformation-based virtual analogues were generated with Python scripts aided by the OpenEye toolkit.\(^{20}\)

2.4.2. Matrix-Based Virtual Candidates. Furthermore, the SAR matrix (SARM) data structure\(^ {21,22}\) was used as a source of virtual analogues for medicinal chemistry series. SARMs are obtained from compound sets through systematic two-step MMP fragmentation and identify all subsets that have structurally analogous cores, i.e., core structures that are distinguished by a structural modification only at a single site.\(^ {25}\) Each subset of analogue series with structurally related cores is represented in a single SARM that is reminiscent of a standard R-group table. A matrix cell represents a unique combination of a core and substituent. Analogue series in SARMs typically contain different substituents. Hence, the systematic recombination of structurally related cores obtained from the second round of fragmentation and substituents from the first round reproduces all existing analogues and, in addition, generates virtual analogues representing as of yet unexplored core—substituent combinations.\(^ {21}\)

By design, structurally related cores and SARMs can be obtained only from analogue series with multiple substitution sites. Such series typically yield multiple SARMs. Depending on the number of structurally related cores and substituents resulting from two-step fragmentation, a series-specific number of virtual candidates is obtained.

A major difference between the transformation- and matrix-based virtual analogue generation approaches is that matrix-based candidates are generally more closely related to existing analogues than transformation-based candidates, for which a diverse array of possible substituents is available. The distribution of virtual analogues from SARMs can be rationalized as an envelope in chemical space formed around an existing series.\(^ {22}\)

The 64 ChEMBL series with 10–30 analogues yielded between 101 and 501 matrix-based virtual candidates per series. In each case, the same number of transformation-based virtual analogues were generated through random sampling. In addition, for direct comparison with PubChem series, a constant number of 13 203 transformation-based virtual analogues per ChEMBL series was also generated through random sampling. The same number of transformation-based virtual analogues was generated for three large ChEMBL series with more than 100 analogues.

3. RESULTS AND DISCUSSION

3.1. Progression Saturation Analysis. The scoring scheme underlying progression saturation assessment is illustrated in Figure 1a,b. Global and local saturation scores
quantitatively account for global chemical space coverage of analogues and the distribution of virtual candidates across chemical neighborhoods of active analogues, respectively. These scores yield characteristic combinations that reflect different levels of saturation and make it possible to assign analogue series to different progression stages according to Figure 1c. Relationships among score magnitudes, analogue distributions, and saturation states are detailed in Section 2. We note that there is no a priori preferred saturation level for analogue series. The methodology is designed as a diagnostics to characterize and differentiate between different levels of chemical saturation, which is important for practical applications.

Herein, we have systematically analyzed sets of series from different sources. The major difference between PubChem and ChEMBL series is that the former contained both active and inactive analogues, whereas the latter exclusively consisted of active analogues from medicinal chemistry publications. Different parameters were evaluated that were expected to impact the computational assessment of progression saturation.

3.2. Z-Scores. Global and local saturation scores must be separately calculated for different sets of analogue series and varying parameter settings including chemical space representations and populations of virtual candidates. For example, for the 80 screening series with 13 203 transformation-based virtual candidates projected into the seven-dimensional reference space, global and local saturation Z-scores covered the intervals [−2.1, 2.3] and [−1.3, 4.6], respectively. Other parameter settings yielded only slightly different score distributions. The threshold value for high global and local saturation Z-scores was set to 1σ above the mean in all cases and was thus constantly 1.0. On the basis of this threshold value, the four different score combinations for classifying analogue series according to Figure 1c were calculated.

3.3. Alternative Chemical Reference Spaces. For the assessment of progression saturation, analogue series must be projected into chemical reference spaces. We first investigated the influence of alternative chemical space representations on score and series categorization. Figure 3 shows the comparison of 7-dimensional (Figure 3a) and 14-dimensional reference spaces (Figure 3b) for 80 PubChem series in the presence of 13 203 transformation-based virtual analogues. In both reference spaces, similar score combinations were obtained for analogue series, leading to a closely corresponding assignment of series to different progression stages (Figure 3c). In both reference spaces of different dimensionalities, the majority of series (60 vs 58) fell into the low/low global/local saturation score category, 13 series belonged to the high/low category, and only three series were assigned to the high/high category. As shown in Figure 4, equivalent observations were made for the 64 ChEMBL series in the presence of 13 203 transformation-based virtual analogues. In both chemical reference spaces, most series (43 vs 49) belonged to the low/low category. In this case, no series were assigned to the high/high category. Hence, both sets of PubChem and ChEMBL series were dominated by analogue series with mid-stage character. Only 17 PubChem and 12 ChEMBL series were found to belong to different categories in the 7- and 14-dimensional reference spaces, consistent with the closely corresponding distributions observed in Figures 3 and 4. Thus, scoring was stable in both reference spaces and very similar assignments were obtained, indicating that progression saturation assessment was not sensitive to chemical reference space variation. Furthermore, the score distributions in Figures 3 and 4 also reveal a significant spread of series across the scoring range, indicating the capacity of global and local scores to distinguish between different series. At least for analogue series from chemical optimization projects published in the medicinal chemistry literature, the observed prevalence of series with a mid-stage character would be expected.
3.4. Virtual Analogues of Different Designs. Next, we compared different ensembles of virtual candidates for progression saturation assessment, which represented another parameter of the analysis. For ChEMBL series having multiple substitution sites, matrix-based virtual analogues were generated, yielding varying numbers of 101−501 candidates per series. For each series, the corresponding number of transformation-based virtual analogues were generated and saturation scores were calculated on the basis of alternative sets of virtual analogues in 14-dimensional reference space. Figure 5a,b shows the score distributions obtained for matrix- and transformation-based virtual candidates, respectively. The resulting assignment of series to score combination categories is shown in Figure 5c. Here, moderate changes were observed, predominantly for the low/high and low/low categories.

![Figure 4](image1.png)  
**Figure 4.** Analysis of ChEMBL series in different chemical reference spaces. Progression saturation of analogue series from ChEMBL with 13,203 virtual candidates was assessed in (a) 7-dimensional and (b) 14-dimensional reference spaces. (c) Assignment of series to different score combination categories. The representation is according to Figure 3.

![Figure 5](image2.png)  
**Figure 5.** Progression saturation analysis using virtual analogues of different designs. Progression saturation of ChEMBL series was assessed on the basis of corresponding numbers of (a) matrix- and (b) transformation-based virtual candidates. Scatter plots report local and global saturation scores obtained for all series (each dot represents a series). (c) Comparison of the assignment of series to different categories. Here, the raw global saturation score calculation was modified using the top 10% of nearest virtual neighbors for determining the neighborhood radius to account for the smaller number of virtual candidates.
Progression saturation assessment in the presence of transformation-based virtual analogues assigned 13 series to the high/low and 38 series to the low/low category. By contrast, assessment in the presence of matrix-based virtual candidates yielded eight high/low and 44 low/low series. In addition, 13 and 12 low/high series were obtained using transformation- and matrix-based virtual analogues, respectively. Thus, there was a shift from late-stage series toward series with the mid-stage character for matrix-based virtual analogues compared to that of transformation-based candidates. By design, matrix-based virtual analogues were structurally closer to existing analogues than transformation-based virtual candidates, which contained structurally more diverse substituents. Hence, matrix-based candidates should be more likely to map to neighborhoods of active analogues than transformation-based virtual candidates, which contained structurally more diverse substituents. However, lower raw saturation scores do not necessarily translate into significant differences in category assignments because the category of a series is determined on the basis of the magnitude of its global and local scores relative to the scores of the other analogue series. Mapping of transformation- or matrix-based virtual analogues delineates the chemical space of an analogue series. One might argue that the space covered by matrix-based virtual compounds, which are closely related to existing analogues, might more accurately reflect the space that is relevant for a given series.

3.5. Increasing Numbers of Virtual Analogues. In addition to alternative analogue design strategies, the use of varying numbers of virtual analogues to sample chemical space was also investigated. Figure 6 shows the results of progression saturation analysis of ChEMBL series in the presence of stepwise increasing numbers of transformation-based virtual analogues. Only small variations in series assignments were observed over a wide range of virtual candidates, indicating that the number of virtual analogues was not a critical parameter for saturation progression assessment. This can be rationalized by taking into account that, independent of the size of the set, virtual candidates are used to sample the chemical space centered on an analogue series. This implies that larger sets of virtual analogues do not necessarily cover a larger section of chemical space. Rather, they more densely sample chemical space around an analogue series. Saturation scores assess the chemical space coverage by the series on the basis of these virtual analogues. Because a small set of samples is less likely to be evenly distributed, one would expect larger variations for smaller sets of virtual candidates and more stable category assignments for larger sets. Over different set sizes, only small fluctuations in distributions were observed, indicating that chemical space coverage by a few hundreds to a few thousands virtual analogues was sufficient for ensuring stable category distributions. For practical applications, the results in Figure 6 suggest that on the order of 1000 virtual candidates are sufficient for the analysis of moderately sized analogue series.

3.6. Evolving Analogue Series. Although parameter evaluation depends on the analysis of large ensembles of similar analogue series from which statistically sound scores can be derived, the assessment of individual series is of particular interest in medicinal chemistry. In practice, the analysis of progression saturation would primarily be on the agenda for evolving series for which a significant number of analogues have already been generated. However, such series are rarely disclosed and only few examples are available in the public domain, which prohibits Z-score calculations. From ChEMBL, we have obtained three series with 100 or more analogues and analyzed them on the basis of raw scores instead. For consistency with calculations reported above, 13 203 transformation-based virtual analogues were generated for each series and projected into the 14-dimensional reference space. To model the evolution of series, subsets of 30 and 60 analogues were taken from each series and compared to those of the complete series. The results are shown in Figure 7a, and representative analogues from each series are depicted in Figure 7b. The analogue series include 126 inhibitors of acetyl-CoA carboxylase 2; 146 inhibitors of phosphodiesterase 10A, the largest available series; and 100 inhibitors of the 5-lipoxygenase activating protein. For compounds from each series, exemplary transformation-based virtual analogues are shown. Figure 7a reveals for each series an increase in saturation scores for subsets of increasing size, consistent with the expectation that increasing numbers of analogues should generally result in increasing levels of chemical saturation. Interestingly, the progression saturation characteristics of all three series differed. The series on the left in Figure 7a yielded the lowest global and highest local saturation scores, the series in the middle had intermediate scores, and the series on the right yielded the highest global and lowest local scores. Thus, on a relative scale, the series on the left had lower analogue coverage and fewer virtual candidates in active neighborhoods than those in the others. By contrast, the series on the right had more extensive analogue coverage than the others and more virtual candidates in active neighborhoods. In qualitative terms, progression saturation of the series in Figure 7a increased from the left to the right. The analogue series on
4. CONCLUSIONS

Herein, we have investigated computational progression saturation analysis of different sets of analogue series, originating from biological screening or medicinal chemistry, under varying parameter settings. The computational approach depends on relating chemical space distributions of existing analogues and virtual candidates to each other and on analyzing chemical neighborhoods of existing analogues. Accordingly, it is important to explore how different analysis conditions might influence the results of progression saturation analysis. Relevant parameters include alternative chemical space representations, varying compound numbers, and different virtual analogue design strategies. Moreover, it is of critical relevance to evaluate the influence of varying computational analysis settings on the categorization of series on the basis of characteristic score combinations; a prerequisite for meaningful practical applications. Therefore, different chemical reference spaces and populations of virtual compounds were explored. Essentially, scoring remained stable under these conditions and only minor to moderate alterations in series categorization were observed. Furthermore, we have analyzed exemplary large analogue series, which are rare in the public domain, and used these analogue series to model series progression. Analysis of evolving series revealed an intuitive increase in chemical saturation when series grew in size, lending further credence to the methodological concept. Moreover, computational comparison of large series comprising similar numbers of analogues revealed different saturation characteristics, which is of high relevance for medicinal chemistry applications. Distinguishing between these characteristics relied on our computational analysis scheme. Herein, we provide all details required for computational progression saturation analysis, which should make it straightforward for interested investigators to implement the methodology.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.8b02087.

Composition of analogue series (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de. Phone: +49-228-7369-100.

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Present Addresses

1Data Science Center and Graduate School of Science and Technology, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan (T.M.).

2Medicinal Chemistry Management Group, Research Function, R&D Division Daichi Sankyo Company, Ltd. Shinagawa R&D Center, 1-2-58, Hiromachi, Shinagawa-ku, Tokyo 140-8710, Japan (R.K.).

Author Contributions

The study was carried out and the manuscript written with contributions of all authors. All authors have approved the final version of the manuscript.
Dimitar Yonchev is supported by a fellowship from the Jürgen Manchot Foundation, Düsseldorf, Germany. The authors thank Chemical Computing Group and OpenEye Scientific Software for providing academic software licenses.

**REFERENCES**

1. Segall, M. Advances in Multiparameter Optimization Methods for De Novo Drug Design. *Expert Opin. Drug Discovery* 2014, 9, 803–817.

2. 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.

3. Maynard, A. T.; Roberts, C. D. Quantifying, Visualizing, And Monitoring Lead Optimization. *J. Med. Chem.* 2016, 59, 4189–4201.

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

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

6. Kunimoto, R.; Miyao, T.; Bajorath, J. Computational Method for Estimating Progression Saturation of Analog Series. *RSC Adv.* 2018, 8, 5484–5492.

7. Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of “Molecular Diversity” Descriptors. *J. Med. Chem.* 1996, 39, 3049–3059.

8. Maggiora, G.; Vogt, M.; Stumpf, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* 2014, 57, 3186–3204.

9. Stumpf, D.; Dimova, D.; Bajorath, J. Computational Method for Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. *J. Med. Chem.* 2016, 59, 7667–7676.

10. Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*, Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271–285.

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

12. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—Retroflexive Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* 1998, 38, 511–522.

13. de la Vega de León, A.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. *Med. Chem. Commun.* 2014, 5, 64–67.

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

15. Dimova, D.; Stumpf, 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* 2016, 2, No. FSO149.

16. Dimova, D.; Stumpf, D.; Bajorath, J. Computational Design of New Molecular Scaffolds for Medicinal Chemistry, Part II: Generalization of Analog Series-Based Scaffolds. *Future Sci. OA* 2018, 4, No. FSO267.

17. Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A. B.; Shoemaker, A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* 2017, 45, D955–D963.