Communication www.molinf.com molecular informatics

DOI: 10.1002/minf.202000046

From SAR Diagnostics to Compound Design: Development Chronology of the Compound Optimization Monitor (COMO) Method

Dimitar Yonchev,[a] Martin Vogt,[a] and Jürgen Bajorath*[a]

Abstract: In medicinal chemistry, compound optimization largely depends on chemical knowledge, experience, and intuition, and progress in hit-to-lead and lead optimization projects is difficult to estimate. Accordingly, approaches are sought after that aid in assessing the odds of success with an optimization project and making decisions whether to continue or discontinue work on an analog series at a given stage. However, currently there are only very few approaches available that are capable of providing decision support. We introduce a computational methodology designed to combine the assessment of chemical saturation of analog series and structure-activity relationship (SAR) progression. The current endpoint of these development efforts, the compound optimization monitor (COMO), further extends lead optimization diagnostics to compound design and activity prediction. Hence, COMO plays dual role in supporting lead optimization campaigns.

Keywords: optimization · analog series · virtual analogs · chemical space · compound neighborhoods · chemical saturation · structure-activity relationships (SARs) · SAR progression · analog design · activity prediction

In this contribution to the *Strasbourg Summer School in Chemoinformatics – 2020*, we discuss a computational approach designed to monitor progress in chemical optimization and review its methodological evolution.

In medicinal chemistry, hit-to-lead and lead optimization (LO) efforts are of critical importance. Establishing dose-response behavior for newly identified active compounds through the design and evaluation of structural analogs and improving the potency (and other LO-relevant properties) of prioritized actives are essential for compound development and the generation of (pre-)clinical candidates. During LO, medicinal chemists continuously face the challenge to decide which compounds to synthesize next. Optimization processes are generally complex and difficult to rationalize. Consequently, they are mostly driven by chemical knowledge and intuition. The subjective nature of optimization efforts comes at a cost because it is difficult to estimate the odds of reaching the final goals of an LO campaign. Moreover, it is equally challenging to make fundamental decisions whether or not sufficient numbers of compounds have been made and, in the presence of limited success, that it might be time to discontinue work on an individual analog series, despite significant investments. Given these challenges, any approaches that help to rationalize compound optimization efforts and provide (even limited) decision support are highly desirable. In principle, computational methods are attractive to diagnose LO and provide some guidance. However, so far only few approaches have been introduced that are applicable to LO, even if only remotely. These include multi-parameter optimization (MPO),[1] compound attrition analysis,[2] statistical risk assessment,[3] and numerical or graphical structure-activity relationship (SAR) diagnostics.[4,5] Compound attrition and risk assessment curves diagnose SAR or property progression in different ways. MPO can only prioritize candidate compounds with favorable properties, but not assess LO progress, and SAR analysis methods determine whether the addition of new compounds to an evolving series adds SAR information. Hence, there are only limited opportunities to computationally evaluate LO, or certain aspects of its progress. In light of this situation, the method development efforts discussed herein chart new territory.[6–10] They ultimately led to the compound optimization monitor (COMO) approach[9,10] that integrates chemical saturation and SAR diagnostics with the prediction of candidate compounds. The following discussion of the theory and scoring schemes is primarily tailored towards a computational audience. For medicinal chemistry applica-

[a] D. Yonchev, Dr. M. Vogt, Prof. Dr. J. Bajorath
Department of Life Science Informatics
Bonn-Aachen International Center for Information Technology
Rheinische Friedrich-Wilhelms-Universität Bonn
Endenicher Allee 19c, D-53115 Bonn (Germany)
Tel: +49-228-7369-100
Fax: +49-228-7369-101
E-mail: bajorath@bit.uni-bonn.de

tions, considering the scoring ranges and combinations should be sufficient.

What criteria might be considered for assessing progress in compound optimization? During optimization, multiple molecular properties must ultimately be balanced such as, among others, compound potency, toxicity, solubility, or metabolic stability. In this context, at least three basic questions can be addressed: Has one sufficiently sampled chemical space for a compound series? Is there detectable SAR progression? Are physicochemical molecular properties favorable? Answering the first two questions in combination would provide immediate decision support. If more chemical modifications can be considered and significant potency variations are observed an analog series (AS) should have potential for further chemical exploration. On the contrary, if a series approaches chemical saturation and there are no significant SAR responses it should better be terminated. Hence, combining the assessment of *chemical saturation* and *SAR progression*, augmented by molecular property controls, provides a potential strategy for computational assessment of compound optimization. As discussed in the following, the computational approach we have designed and implemented makes use of virtual compounds as diagnostic tools; a unique feature. Naturally, *virtual analogs* (VAs) generated for a given AS might also be considered as potential candidate compounds for optimization efforts. If one would like to do so, then a fourth basic question must be taken into consideration: Is the generated VA space synthetically accessible? Although the primary focal point of the methodology is (chemical saturation and SAR) diagnostics, this question is also addressed below.

On the basis of pre-selected active compounds, ASs are generated by varying substituents (R-groups) at one or more synthetically accessible substitution sites in the compound core structure. Hence, ASs typically evolve around a core structure that is shared by *existing analogs* (EAs). For computational analysis, a key question is how to assess chemical saturation of a series. However, answering this question is a non-trivial task.

A first computational concept for the assessment of chemical saturation of an AS was formulated that considered sampling of chemical space around the series as a pivotal criterion.[6] For this purpose, an AS and a corresponding population of VAs might be projected into chemical space and their distance relationships determined.[6] Central to this concept is the definition of distance-based *chemical neighborhoods* (NBHs) of each EA, as illustrated in Figure 1. The EA and VA content of NBHs can then be determined to assess sampling of chemical space around an AS as an indicator of chemical saturation. Therefore, a *global* and a *local saturation score* were defined.[6]

The global saturation score measures the degree of chemical space coverage by (active or inactive) EAs:

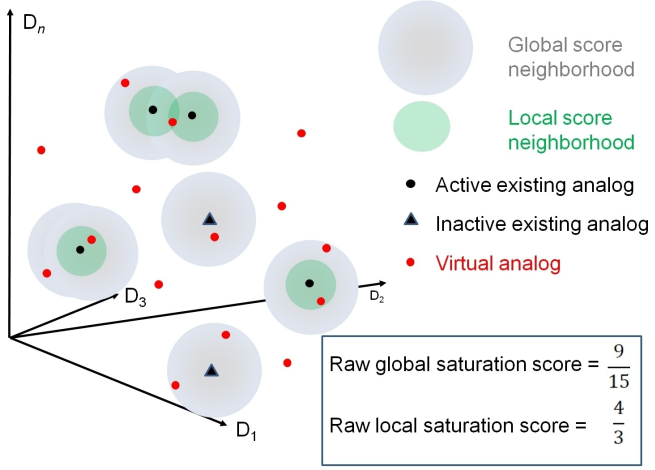
D$_n$
D$_3$
D$_2$
D$_1$
Global score neighborhood
Local score neighborhood
● Active existing analog
▲ Inactive existing analog
● Virtual analog
Raw global saturation score = $\frac{9}{15}$
Raw local saturation score = $\frac{4}{3}$

Figure 1. Global and local saturation scores. Shown is a schematic representation of chemical space populated with an AS and corresponding VAs. Neighborhoods of existing (active or inactive analogs) and, in addition, of active analogs for score calculations specified in the text are represented as large gray and small blue circles, respectively. Exemplary score calculations are given. Nine out of all 15 VAs fall into NBHs of active or inactive analogs. In addition, three VAs exclusively map to NBHs of four active analogs. This compound distribution over NBHs of the AS gives rise to the reported global and local saturation scores. The format of this figure was adapted from reference [6] and modified.

$$\text{Global saturation score} = \frac{|v_{NBH}|}{|V|} \qquad (1)$$

where V denotes the set of all VAs and v_{NBH} the set of VAs falling into NBHs of EAs. High global score values indicate extensive chemical space coverage by EAs of a series. If only active analogs are available, the global score is calculated in the same way.

In addition, the local saturation score specifically emphasizes the NBH coverage for (optimization-relevant) active EAs and quantifies the VA content:

$$\text{Local saturation score} = \frac{|A|}{|V_{NBH_A}| + 1} \qquad (2)$$

Here, A is the subset of all active EAs and V_{NBH_A} the set of VAs located in NBHs of only active EAs. Figure 1 illustrates the calculation of both scores. Raw global and local scores are subsequently converted into conventional z-scores.

Of critical relevance for the analysis is the way in which NBH radii are determined. For calculating the global score, the NBH radius for each EA is set to the median value of the distribution of mean Euclidian distances between each VA and the top 1% of its nearest virtual neighbors. Hence, in this case, distance relationships between VAs whose numbers are much larger than EAs -and their chemical space coverage thus much more extensive- determine the NBH radius. For calculating the local score, the NBH radius for each active EA is set to the median value of the pairwise distances between active analogs. Thus, distance relationships between active analogs determine the NBH radius here.

This scoring scheme enables the consideration of active and inactive analogs. Therefore, ASs were extracted from publicly available biological screening data,[11] yielding active and inactive compounds, and also from medicinal chemistry sources,[12] yielding only active analogs. For a given AS, VA populations were generated by decorating substitution sites in the common AS core with R-groups randomly selected from a large substituent library extracted from ChEMBL[12] compounds. As a chemical reference space for computational analysis, multi-dimensional chemical feature (descriptor) spaces were used in which the position of a compound is determined by an N-dimensional feature vector.

Comparison of the global and local saturation scores makes it possible to define characteristic score combinations that are indicative of different LO development stages. This is illustrated in Figure 2a. For example, the high/high global/local score combination (upper right quadrant) is indicative of extensive chemical space coverage with only limited numbers of remaining virtual candidates contained in NBHs of active compounds, which would characterize chemically saturated series. By contrast, the high/low score combination reflects extensive chemical space coverage with many virtual candidates in NBHs of active EAs that might still be explored. Hence, this combination would characterize late stage series.

Figure 2a shows score combinations of 80 relatively small ASs from screening data with 30 to 65 analogs per series.[6] The majority of these ASs fall into the lower left quadrant (low/low global/local scores) indicating mid-stage character, with still limited chemical space coverage but NBHs of active EAs already containing many virtual candidates. Figure 2b shows three ASs from the medicinal chemistry literature[12] for which growth was modeled by considering randomly selected compound increments. Accordingly, saturation scores were calculated for differently sized subsets.[7] All three ASs displayed distinct patterns of increasing chemical saturation. A systematic investigation of analysis and scoring parameters revealed that saturation scores were essentially stable for feature spaces of different composition and VA populations covering a wide range.[7]

Chemical saturation scoring was further refined and combined with another numerical scoring scheme accounting for SAR progression.[8]

The basic principles underlying chemical saturation scoring discussed above were further extended as follows. If VA populations are used to uniformly sample series-centric chemical space (i.e., a section of chemical space where EAs are located), the saturation of an AS can be assessed by considering two factors including the *(i) extensiveness* of chemical space coverage and *(ii) density* of coverage. The extensiveness of chemical space coverage is quantified by the *coverage score C*:

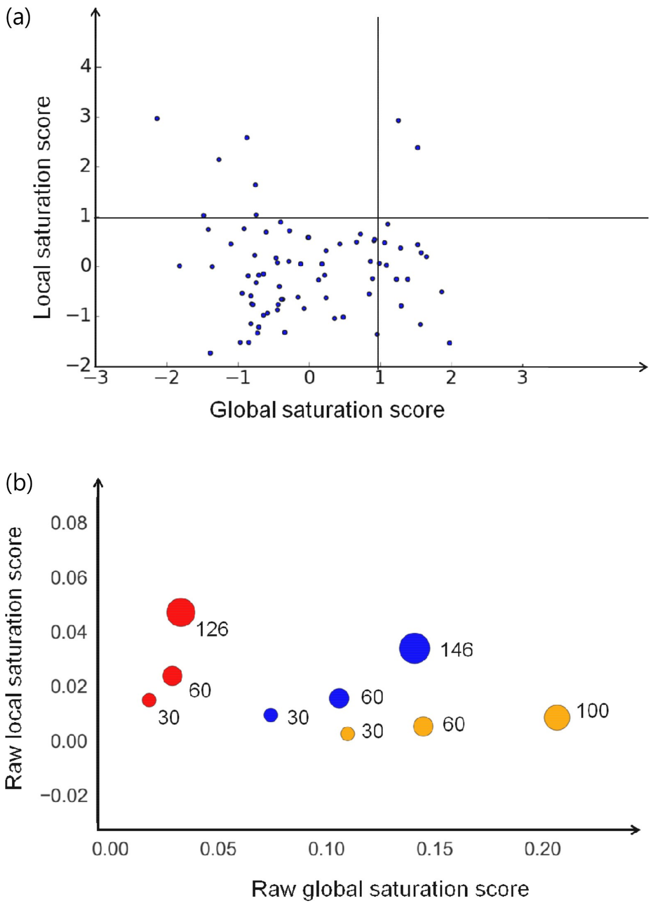

Figure 2. Saturation scores of analog series. (**a**) A scatter plot shows the distributions for global (x-axis) and local (y-axis) saturation z-scores for a set of ASs (black dots). The quadrants indicate regions characterized by different score combinations that are indicative of different LO stages, as described in the text. The figure was adapted from ref. [6] and modified. (**b**) A scatter plot shows raw saturation scores for three different ASs (inhibitors of acetyl-CoA carboxylase 2a, red; phosphodiesterase 10, blue; and S-lipooxygenase activating protein, orange) for which growth over time was modeled. Dots are scaled in size according to compound numbers, which are reported next to each dot. The figure was adapted from ref. [7] and modified.

$$C = \frac{n_{NBH}}{n_V} \qquad (3)$$

Here, n_{NBH} and n_V refer to the number of VAs in NBHs of EAs and the total number of VAs, respectively. The C score is identical to the initially defined global saturation score and has the range [0,1]. As further extension, NBH radii are determined here as the *n*-th quantile of the shortest distances between VAs, which can be easily adjusted.

In addition, the *density of coverage score D* determines how closely EAs map chemical space by quantifying the overlap of their NBHs:

$$D = 1 - \frac{1}{d_{mean}} \tag{4}$$

The term d_{mean} is defined as the number of overlapping NBHs containing VAs (NBH_{O_VA}) relative to the total number of VAs (n_{NBH}) falling into NBHs of EAs:

$$d_{mean} = \frac{NBH_{O\_VA}}{n_{NBH}} \tag{5}$$

The D score is normalized to the range [0,1] and is complementary to the C score.

Both scores are combined by calculating their harmonic mean yielding the second generation *chemical saturation score S*:

$$S = \frac{2CD}{C + D} \tag{6}$$

Hence, the original global and local saturation scores were replaced by a generalized measure of AS-centric chemical space coverage.

In the next step, scores are derived to account for SAR progression of ASs by applying the NBH principle, consistent with saturation scoring. As a measure of SAR progression, SAR discontinuity can be considered, which results from potency variations among structural analogs[13] and thus mirrors the SAR response to chemical modifications. To account for SAR discontinuity, the following scoring scheme was developed.[8] For VAs located in overlapping NBHs, the mean pairwise potency range among EAs associated with those NBHs is quantified as the NBH-specific term $\bar{\Delta}_i$:

$$\bar{\Delta}_i = \frac{2}{m_i(m_i - 1)} \sum_{\substack{j,k=1 \\ j<k}}^{m_i} |\mathrm{potj} - \mathrm{potk}| \tag{7}$$

Here, m_i is the number of EAs with overlapping NBHs into which the VA falls, potj and potk represent the logarithmic (log) potency of EA *j* and *k*, respectively. The *SAR progression score P* for the entire AS represents the mean over $\bar{\Delta}_i$ for all n_N VAs in overlapping NBHs of EAs applying a weighting scheme $w_i = \frac{1}{m_i}$ if $m_i > 1$ and $w_i = 0$ if $m_i = 1$:

$$P = \frac{1}{\sum_{i=1}^{n_N} w_i} \sum_{i=1}^{n_N} w_i \bar{\Delta}_i \tag{8}$$

According to this definition, increasing SAR discontinuity in densely populated regions of chemical space translates into high SAR progression, yielding large P scores. Figure 3a illustrates principles of C, D, S, and P score calculations.

Figure 3b compares S and P scores for different ASs from medicinal chemistry sources.[8] Importantly, S and P scores are essentially uncorrelated and do not scale with the size of ASs. Characteristic score combinations assign ASs to different optimization stages. For example, ASs with low/low S/P scores represent early-stage series that are subject to further chemical exploration. By contrast, ASs with high/low S/P scores are highly saturated and lack SAR progression, which represents a termination criterion, regardless of whether these ASs have reached LO milestones or not. Only if potency optimization was successful, generation of further analogs might be considered in order to balance multiple compound properties. Moreover, ASs with high/high scores are chemically saturated, but still display a high degree of SAR discontinuity/progression, which indicates the presence of steep SARs. These SARs are prone to producing activity cliffs that are often undesired at late stages of LO when multiple properties must be balanced and high potency retained. Finally, ASs with intermediate S and P scores generally represent promising candidates that merit further consideration. These conclusions regarding LO characteristics of investigated ASs are confined to the context of the dual S/P scoring scheme, but provide unbiased optimization diagnostics complementing subjective chemical assessment.

Using the dual saturation and progression scoring scheme as a foundation, the compound optimization monitor (COMO) framework was introduced including additional score components.[9]

The presence of SAR heterogeneity in a compound series or data set results from the coexistence or combination of continuous and discontinuous SAR components.[13] To complement the VA-dependent SAR progression score discussed above with a VA-independent score, an *SAR heterogeneity score H* was developed.[9] This score is calculated on the basis of all EAs. For a given AS, it assesses the magnitude and directionality of potency progression.[9] By design, the H score was defined without taking VA populations into consideration to generate a VA-independent measure:

$$H = \frac{\sum_{i=1}^{n} w_{NBH_i} pot_i}{\sum_{i=1}^{n} w_{NBH_i}} - \overline{pot_{AS}} \tag{9}$$

It represents the weighted mean potency deviation from the average compound potency $\overline{pot_{AS}}$ within an AS where pot_i denotes the individual logarithmic potency of each EA *i* and the weighting factor w_{NBH_i} is equal to the number of overlapping NBHs formed between analog *i* and other EAs. Increasing positive or negative H scores indicate that the majority of NBHs of EAs contribute to an overall positive or negative potency gradient within a series.

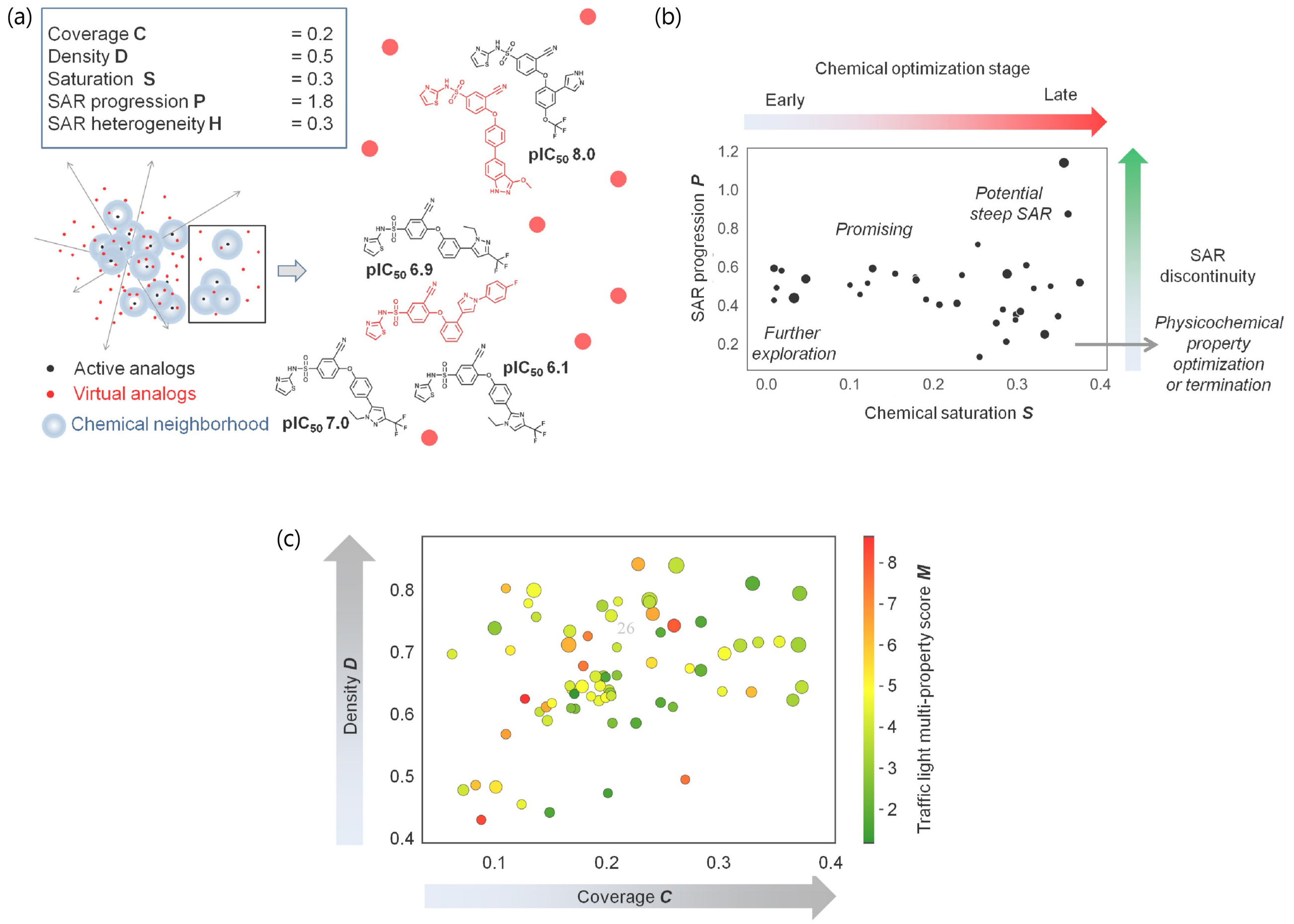

Figure 3. Chemical saturation and SAR progression scores. In (**a**), score calculations are illustrated based on the distributions of active and virtual analogs in chemical space. The enlarged section on the right contains four active and 10 virtual analogs. Three of four active analogs have overlapping NBHs that share a virtual analog. In addition, another virtual analog maps to the NBH of an isolated active analog. This compound distribution and the potency values of active analogs give rise to the exemplary scores that are reported. The figure was adapted from ref. [9] and modified. (**b**) The scatter plot compares S and P scores for a set of ASs. Each dot represents an AS and is scaled in size according to compound numbers per series. Red and green arrows represent LO stages of ASs (ranging from early-stage to late-stage series) and increasing SAR discontinuity, respectively. Different regions in the plot are labeled with AS characteristics on the basis of specific score combinations. For example, ASs with low S and P scores represent early-stage series that merit further chemical exploration. The figure was adapted from ref. [8] and modified. (**c**) The scatter plot compares D and C scores accounting for chemical saturation for another set of ASs (dots). The dots are color-coded according to an independently calculated multiple-property score applying the “traffic light” spectrum shown on the right. Low scores are favorable and high scores unfavorable. The figure was adapted from ref. [9] and modified.

Hence, the score serves as an additional SAR progression diagnostic. An exemplary H score calculation is reported in Figure 3a.

Independent of the saturation and progression scoring scheme, the *in silico ADME traffic light score*[14] was implemented in COMO as a *multi-property score M*. This score combines contributions from several *in vivo*-relevant compound properties[14] and can be calculated at any LO stage to assess whether mean property characteristics of ASs are favorable or not, as illustrated in Figure 3c. As such, the M score complements the COMO scoring formalism.

AS-specific VA populations were primarily generated as diagnostic tools. However, they can also be considered as a pool of potential virtual candidate compounds for further optimization. For achieving this dual purpose, balancing diagnostic chemical space coverage and synthetic accessibility is essential. Diagnostic VA populations were initially generated by systematically combining AS cores with substituents from a large ChEMBL-based library,[6] as mentioned above. To further increase synthetic accessibility of candidate compounds, a refined VA generation protocol was implemented in COMO[9] to enumerate compounds on the basis of retrosynthetic rules.[15] For diagnostic purposes, original and refined VA populations were found to be equivalent. For example, the C and D score calculations in Figure 3c were carried out on the basis of retrosynthetic VAs, confirming that these scores were uncorrelated and did not scale with AS size.

If one would like to consider VAs in the practice of medicinal chemistry as candidate compounds depends on individual preferences and project requirements. For a project team, there are many synthesis constraints to balance and computed VAs might not meet specific requirements. However, it is possible to prioritize VAs as potential candidates, which one may take into consideration.

For considering COMO's VA populations as a potential source of candidate compounds, a key question was how to best prioritize and select them. First and foremost, potency predictions of VAs were considered to narrow down the choice to candidates.

For this purpose, machine learning (ML) regression models were extensively tested on a set of 24 ASs with at least 100 compounds per series.[10] ML approaches included support vector regression (SVR)[16] representing a generally preferred approach for potency predictions in the presence of non-linear SARs. However, for half of the investigated ASs, no reliable ML models including SVR could be derived, likely owing to the still limited number of active compounds for training, even in the case of the larger ASs investigated here.[10] Predicting highly potent compounds, which are most interesting for prospective applications, was particularly difficult via ML regression. These findings precluded potency predictions using SVR or other regression models for VA populations of many ASs. Therefore, as a simple alternative, Free-Wilson (FW) additivity analysis[17] was considered, as illustrated in Figure 4a. FW analysis predicts compound potency on the basis of contributions of individual substituents that are derived from corresponding analogs and assumed to be independent and additive, representing approximations.[17] As such, FW analysis is tailor-made for quantitative structure-activity relationship analysis (QSAR) of ASs.[18] However, for any given compound, the applicability of FW predictions strictly depends on the availability of suitable analogs that permit estimating contributions from individual substituents, as illustrated in Figure 4a. Such analogs are termed FW analogs herein. For the majority of the investigated ASs, at least limited numbers of FW predictions for model validation were possible. In these cases, potency prediction accuracy was clearly improved compared to ML models.[10] On the basis of these findings, FW potency predictions were prioritized for VA populations of ASs.

Despite promising results obtained for original ASs in model validation, FW predictions were intrinsically limited for VA populations. This was the case because only few qualifying virtual FW analogs (FW VAs) were found in most diagnostic VA populations including retrosynthetic VAs, hence prohibiting systematic predictions on these populations. To circumvent these shortcomings, a new algorithm employing the matched molecular pair (MMP) formalism[19] and MMP networks was devised to generate FW VAs for given ASs.[10] For all but one of the investigated 24 ASs, hundreds to thousands of FW VAs were obtained, hence enabling systematic predictions of candidate compounds on the basis of FW VAs instead of diagnostic VA populations.

Figure 4b compares potency predictions on FW VA populations using FW analysis and SVR models for different ASs. For each AS, the experimental potency value distributions for EAs is also shown. The comparison of FW and SVR predictions reveals some trends. The FW potency distribution is often broader than the corresponding SVR distribution, has a lower median value, but contains a larger number of statistical "outliers" having higher (or lower) predicted potency than EAs. Figure 4c shows a high-resolution view of experimental and predicted potency distributions for an individual AS, further illustrating such trends. Albeit theoretical in nature, these findings render FW predictions attractive for practical applications because they consistently yield predictions of FW VAs having higher potency than EAs, as illustrated in Figure 4b and 4c, which would represent prime candidates for experimental evaluation. Such predictions can be particularly useful during later optimization stages when larger numbers of EAs become available that enable the generation of and potency predictions on large pools of corresponding FW VAs.[10]

Progress in compound optimization is notoriously difficult to estimate. This especially applies to the questions if sufficient numbers of analogs of a given AS might have been made and when it might be time to discontinue a series. Accordingly, computational methods that aid in evaluating hit-to-lead or LO efforts are highly desirable. However, currently only few approaches are available that are capable of providing decision support. To these ends, the COMO methodology was devised, as discussed herein.

The development efforts began with formulating and implementing a new computational concept for the assessment of chemical saturation of ASs, which was based upon a compound NBH principle and involved diagnostic VA populations to quantify chemical space coverage. Saturation scoring was then further refined and combined with a new SAR progression scoring scheme, providing the foundation of COMO, which incorporated additional complementary score components. Furthermore, the use of diagnostic and synthetically accessible VA populations suggested the expansion of the approach to prioritize candidate compounds for further optimization, if appropriate. This dual role of VAs represents another characteristic feature of COMO. Therefore, different prediction strategies and VA sources were explored for prioritizing candidate compounds from VA populations. While ML regression models yielded only limited prediction performance on ASs of varying size, FW analysis was identified as a preferred approach. However, diagnostic VA populations were found to contain too few FW VAs for meaningful predictions. Therefore, augmenting diagnostic VA populations with specially designed FW VAs yielded a wealth of candidates for systematic FW predictions, providing a perspective for practical applications.

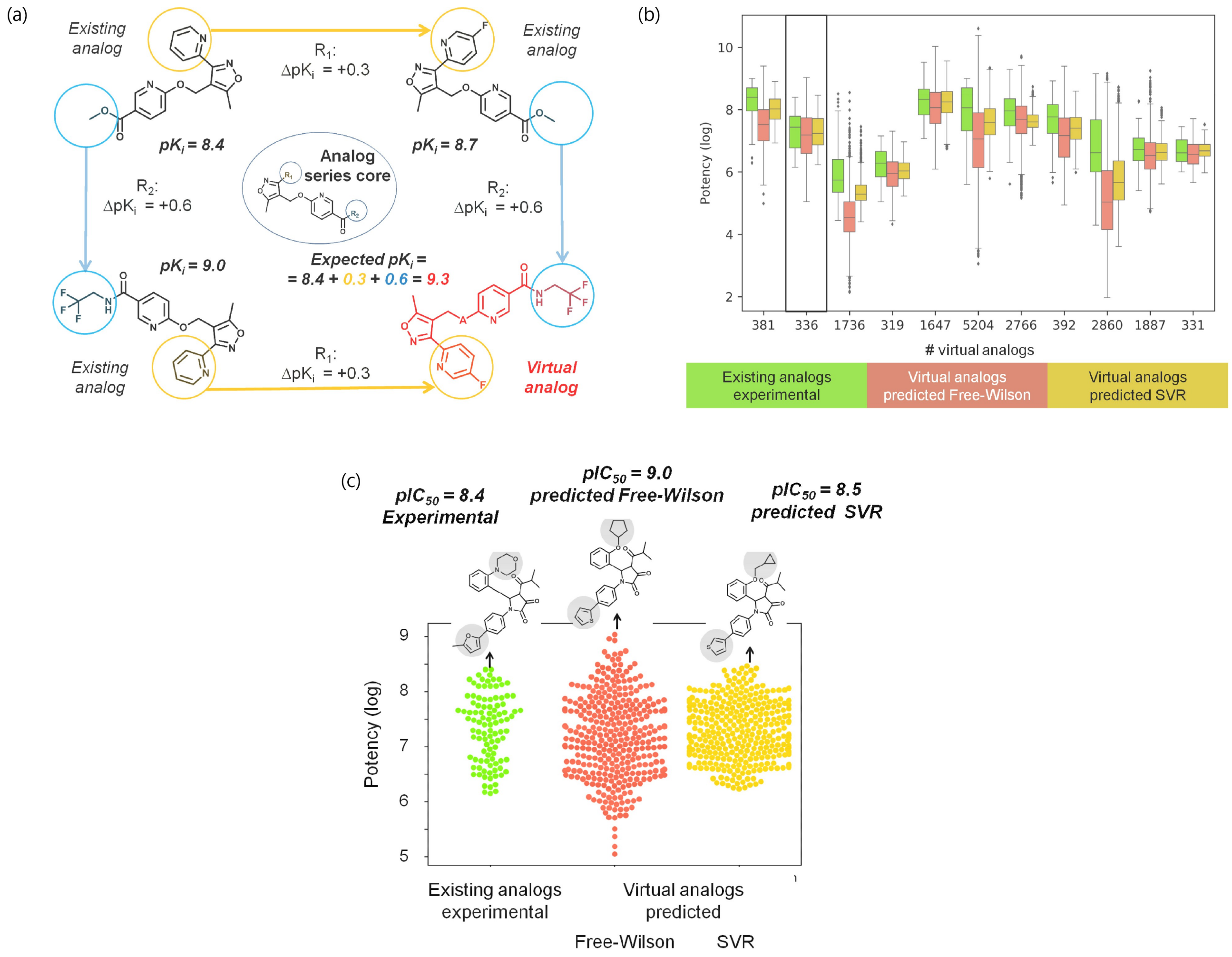

Figure 4. Potency predictions and candidate compounds. In (**a**), the Free-Wilson additivity principle is illustrated. Shown are four analogs active against the GABA receptor alpha-5 subunit. The potency values of three of these analogs are reported. The compound shown in red on the lower right was considered a VA whose potency was predicted following the FW approach. The figure format was adapted from ref. [10] and modified. In (**b**), box plots report experimental (green) and predicted (FW, red; SVR, yellow) potency value distributions for 11 ASs and their corresponding FW VA populations respectively. FW VA numbers are given on the x-axis. For FW VAs, mean potency values over multiple predictions are reported. For each AS, three box plots are given. The second AS from the left with 336 FW VAs is marked using a black frame. The figure was adapted from ref. [10] and modified. (**c**) shows a high-resolution view of the experimental and predicted potency value distributions for the AS marked in (b) consisting of P2X purinoreceptor 3 ligands. Each dot represents an individual compound. For each distribution, the structure of the most potent EA or VA is shown. Distinguishing substituents are displayed on a gray background and experimental or predicted potency values are reported. The figure was adapted from ref. [10] and modified.

The COMO scoring scheme is straightforward to implement, while the theory behind some score details might be more challenging to rationalize. However, for practical applications, an in-depth study of the underlying theory is not required. Once COMO scores are implemented and VA populations generated, which falls into the domain of computational chemists, the scores are easy to calculate. In the practice of medicinal chemistry, calculating S and P scores and evaluating them in combination, which is intuitive, is readily sufficient to provide some insights into AS saturation and progression characteristics. As far as further development of COMO is concerned, it is likely that more efforts will be made to complement current sourcing and prioritization of candidate compounds with additional design strategies. Furthermore, the COMO scoring scheme will be further extended with formalisms to balance multiple LO-relevant properties.

Taken together, the method development efforts leading to COMO reflect the evolution of a novel computational methodology to tackle a central problem in medicinal chemistry, the rationalization of LO efforts.

Conflict of Interest

None declared.

Acknowledgements

The authors gratefully acknowledge Ryo Kunimoto and Tomoyuki Miyao for their contributions to computational chemical saturation analysis. Furthermore, the authors thank OpenEye Scientific Software and Chemical Computing Group for academic licenses. D.Y. is a fellow of the Jürgen Manchot Stiftung, Düsseldorf, Germany. Open Access funding enabled and organized by Projekt DEAL.

References

[1] M. Segall, *Expert Opin. Drug Discovery* **2014**, *9*, 803–817.
[2] M. Munson, H. Lieberman, E. Tserlin, J. Rocnik, J. Ge, M. Fitzgerald, V. Patel, C. Garcia-Echeverria, *Drug Discovery Today* **2015**, *20*, 978–987.
[3] A. T. Maynard, C. G. Roberts, *J. Med. Chem.* **2015**, *59*, 4189–4201.
[4] V. Shanmugasundaram, L. Zhang, S. Kayastha, A. de la Vega de León, D. Dimova, J. Bajorath, *J. Med. Chem.* **2015**, *59*, 4235–4244.
[5] P. Iyer, Y. Hu, J. Bajorath, *J. Chem. Inf. Model.* **2011**, *51*, 532–540.
[6] R. Kunimoto, T. Miyao, J. Bajorath, *RSC Adv.* **2018**, *8*, 5484–5492.
[7] D. Yonchev, M. Vogt, D. Stumpfe, R. Kunimoto, T. Miyao, J. Bajorath, *ACS Omega* **2018**, *3*, 15799–15808.
[8] M. Vogt, D. Yonchev, J. Bajorath, *J. Med. Chem.* **2018**, *61*, 10895–10900.
[9] D. Yonchev, M. Vogt, J. Bajorath, *Future Drug Discov.* **2019**, *1*, FDD15.
[10] D. Yonchev, J. Bajorath, *Future Science OA* **2020**, *6*, FSO451.
[11] Y. Wang, S. H. Bryant, T. Sheng, J. Wang, A. B. Gindulyte, A. Shoemaker, P. A. Thiessen, S. He, J. Zhang, *Nucleic Acids Res.* **2017**, *45*, D955–D963.
[12] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J. P. Overington, *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.
[13] L. Peltason, J. Bajorath, *J. Med. Chem.* **2007**, *50*, 5571–5578.
[14] M. Lobell, M. Hendrix, B. Hinzen, J. Keldenich, H. Meier, C. Schmeck, R. Schohe-Loop, T. Wunberg, A. Hillisch, *ChemMedChem* **2006**, *1*, 1229–1236.
[15] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 511–522.
[16] H. Drucker, C. Burges, *Adv. Neural Inform. Process. Systems* **1997**, *9*, 155–161.
[17] S. M. Free, J. W. Wilson, *J. Med. Chem.* **1964**, *7*, 395–399.
[18] H. Kubinyi, *Quant. Struct.-Act. Relat.* **1988**, *7*, 121–133.
[19] J. Hussain, C. Rea, *J. Chem. Inf. Model.* **2010**, *50*, 339–348.

Received: March 16, 2020
Accepted: April 13, 2020
Published online on April 24, 2020