Future Drug Discovery

Methodology

Compound optimization monitor (COMO) method for computational evaluation of progress in medicinal chemistry projects

Dimitar Yonchev[1], Martin Vogt[1] & Jürgen Bajorath*,[1]

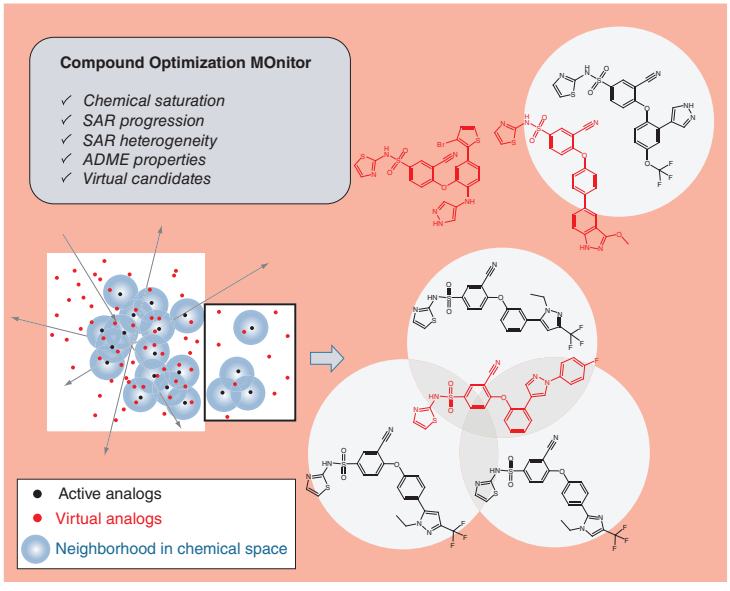
[1]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53113 Bonn, Germany

*Author for correspondence: Tel.: +49 228 736 9100; Fax: +49 228 736 9101; bajorath@bit.uni-bonn.de

Aim: Development of a new, practically applicable computational method to monitor progress in lead optimization. Computational approaches that aid in compound optimization are discussed and the Compound Optimization Monitor (COMO) method is introduced and put into scientific context. **Methodology & calculations:** The methodological concept and the COMO scoring scheme are described in detail. **Results & discussions:** Calculation parameters are evaluated, and profiling results reported for an ensemble of analog series. **Future perspective:** The dual role of virtual analogs as diagnostic tools for progress evaluation and as potential candidates for lead optimization is discussed. In light of this dual role, interfacing COMO with machine learning for compound activity prediction and prioritization of candidates is highlighted as a future research objective.

Lay abstract: In medicinal chemistry, new active compounds must be chemically converted into leads for drug development and further optimized to generate clinical candidates. This process is largely driven by subjective criteria, chemical intuition and experience. Only few computational approaches are available to support this process. COMO is introduced as the first computational method to directly evaluate progress made in lead optimization and aid in judging whether or not sufficient numbers of compounds have been generated.

Graphical abstract:



First draft submitted: 19 April 2019; Accepted for publication: 30 May 2019; Published online: 11 October 2019

newlands press

Keywords: chemical saturation • compound optimization • computational analysis • decision support • drug discovery • property evaluation • SAR progression

Hit-to-lead and lead optimization (LO) are central tasks in medicinal chemistry [1]. The ultimate goal of LO is the generation of clinical candidate compounds. Once a lead compound has been obtained that displays promising biological activity and structural features for further exploration, LO usually requires the generation of many analogs to further improve compound potency and other optimization-relevant properties [1]. LO is largely driven by the experience and intuition of medicinal chemists and is often perceived to be more of an art form than a science. For a given analog series (AS), the exploration and exploitation of structure–activity relationships (SARs) typically present new challenges that must be considered on a case-by-case basis. It is very difficult – if not impossible – to generalize LO strategies and predict the outcome of LO campaigns. A particularly critical issue during LO is estimating the odds of ultimate success for a given AS. Once large amounts of time and resources have been spent to further improve lead(s), discontinuing work on an AS is a difficult call to make in the practice of medicinal chemistry, understandably so. Consequently, LO efforts are often carried out for too long until they are finally suspended, and there is only very little external decision support available.

One would hope for decision support through computational analysis aiming to rationalize parts of the LO process. However, there currently are only few computational methods available to support LO, beyond compound potency prediction [2,3]. For example, multiobjective optimization is frequently applied to combine and weigh different compound properties and score candidate compounds [4,5]. In addition, statistical methods can be used to evaluate SAR progression or prioritize compounds that make positive contributions toward LO [6,7]. However, none of these few computational approaches with utility for LO is capable of assessing when an AS might be saturated, and generating more compounds would be unlikely to yield further progress. To these ends, new computational methodologies are required.

We have spent considerable efforts developing computational concepts for the evaluation of chemical saturation of ASs [8,9] and the combination of saturation and SAR progression analysis [10]. These concepts have provided the foundation of the Compound Optimization MOnitor (COMO) methodology presented herein. COMO is designed to address the questions to what extent an AS is chemically saturated, if there is further potential for SAR progression and if attractive candidate compounds still exist. It employs an intuitive scoring scheme comprising multiple score components to quantitatively assess LO progress and provide decision support for medicinal chemistry. In the following sections, the COMO methodology is detailed and its application to a panel of ASs from medicinal chemistry reported.

Methodology & calculations

Methodological concept

COMO was designed to evaluate LO efforts by assessing how extensively and densely chemical space around an AS is covered (chemical saturation) and whether potential for SAR progression is detectable. Key components of the approach include the use of virtual analog (VA) populations for a given AS to chart series-centric chemical space and the generation of chemical neighborhoods (NBHs) of active analogs [8]. VAs serve a dual purpose as diagnostic molecular entities and potential candidates for further optimization. Moreover, the application of the NBH concept makes it possible to distinguish between overlapping and nonoverlapping NBHs as a measure of compound density, map locations of VAs and characterize their SAR environments [9]. For an AS, the potential of further SAR progression is evaluated in a VA-dependent manner by determining local SAR discontinuity [9] as well as in a VA-independent manner by assessing global SAR heterogeneity. To quantify chemical saturation and SAR progression, two pairs of complementary and chemically interpretable scores are designed. One of these pairs yields a combined saturation score. In addition, a multiproperty score is introduced, taking into consideration that different compound properties must be balanced during late stages of LO. The methodological concept of COMO is illustrated in Figure 1. We note that NBHs are defined on the basis of distance relationships between compounds in chemical reference space, as further explained below, and that no similarity metrics are applied.

Virtual analogs

For a given AS, a set of VAs is generated using a newly developed computational enumeration scheme:

(i) From all bioactive compounds in ChEMBL (release 24) [11] with available high-confidence activity data (252,779 compounds in total), 16,575 unique substituents with up to 13 nonhydrogen atoms were systematically extracted using matched molecular pair fragmentation of exocyclic single bonds [12] on the basis of retrosynthetic rules [13]. These substituents provide a pool for VA design.

(ii) From an AS, all substituents including hydrogen atoms attached to the common core structure are collected. For each AS, the proportion of hydrogen atoms among all substituents is determined, which represents the AS-specific likelihood of hydrogen substitutions. It is calculated by dividing the number of hydrogens found in analogs across all substitution sites by the total number of substituents collected for a given AS.

(iii) The set of 16,575 substituents is used to enumerate VAs on the basis of the following rules:
- Substituents are permitted to contain at most 13 nonhydrogen atoms and the total size of a VA (including all substituents) is limited to at most 1.5-times the size of the corresponding core.
- For each substitution site, the subset of qualifying substituents is determined by testing whether the resulting bond meets retrosynthetic rule(s); 12 of 13 previously defined rules [13] are considered (excluding olefinic double bonds).
- ASs with single and multiple substitution sites are investigated. In the case of single substitution sites, VAs are enumerated using all qualifying substituents (including a hydrogen atom). If an AS has multiple substitution sites, VAs are generated by randomly decorating each site with a hydrogen or a qualifying nonhydrogen substituent on the basis of the AS-specific likelihood of hydrogen substitutions according to (ii).

For the analysis reported herein, 10,000 unique VAs were generated for each AS with multiple substitution sites. For ASs with single substitution sites, between 5191 and 9850 unique VAs per series were obtained, depending on the number of qualifying substituents.

Scoring system

The COMO scoring scheme consists of two categories of scores accounting for chemical saturation and SAR progression, respectively, yielding four score components. In addition, a property score is introduced to balance multiple optimization-relevant compound properties, which can be flexibly selected for a given compound class and optimization task.

Chemical neighborhood radius

For each active analog, the NBH radius is set herein to the first percentile of the distribution of pairwise distances between VAs in chemical reference space. This setting has been selected on the basis of test calculations reported below. Distance between two compounds in chemical space is calculated as the Euclidian distance between their descriptor (feature) vectors.

Since VA populations are much larger than existing ASs, they mostly determine coverage of chemical space, which rationalizes the consideration of VA distance relationships for NBH definition [8,9]. VAs might map to nonoverlapping NBHs, overlapping NBHs or outside of NBHs, which is quantitatively accounted for through scoring as detailed below.

Chemical saturation

For a given AS and the corresponding VA population, coverage C of chemical space is quantified as the proportion of VAs that fall into NBHs of active analogs:

$$C = n_N / n_V$$

Here, n_N and n_V refer to the number of VAs in NBHs and the total number of VAs, respectively. The C score has the range (0,1).

Furthermore, a subset of VAs in NBHs might be located in overlapping NBHs.

The more densely the chemical space is covered by active analogs, the larger the total number of overlapping NBHs becomes and the larger the likelihood will be that VAs in NBHs map to overlapping NBHs.

Accordingly, d_{mean} is defined as the number of overlapping NBHs containing VAs (NBH_{O_VA}) relative to the number of VAs falling into NBHs:

$$d_{\text{mean}} = \text{NB}H_{O\_VA}/n_N$$

It is normalized to the density score D having the range (0,1):

$$D = 1 - d_{mean}^{-1}$$

Combined coverage and sampling density of chemical reference space is a measure of chemical saturation. Accordingly, the saturation score S is defined as the harmonic mean of score components C and D:

$$S = 2\text{CD}/(C + D)$$

SAR progression

If a VA is present in overlapping NBHs of multiple active analogs, the magnitude of potency variations among these analogs indicates the degree of SAR discontinuity across the associated NBHs. The parameter $\overline{\Delta}_i$ is introduced to capture the potency range of m_i active analogs that form overlapping NBHs containing a VA. For a given VA in overlapping NBHs, $\overline{\Delta}_i$ is computed as the mean potency difference over all pairs of m_i active analogs (pot_j and pot_k represent the logarithmic [log] potency of compound j and k, respectively):

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

The SAR progression score P is then calculated as the mean over all VAs in NBHs applying a weighting scheme $w_i = \frac{1}{m_i}$ if $m_i > 1$ and $w_i = 0$ if $m_i = 1$:

$$P = \frac{1}{\sum_{i=1}^{n_N} w_i} \sum_{i=1}^{n_N} w_i \overline{\Delta}_i$$

If follows that only VAs in overlapping NBHs contribute to P. The score is a measure of local SAR discontinuity across overlapping NBHs containing VAs. For P, large values are obtained when VAs map to overlapping NBHs of active analogs with large potency fluctuations. In such regions, virtual candidates might yield highly potent compounds. Accordingly, large P values indicate potential for further SAR progression.

Herein, we introduce an additional SAR measure to complement VA-centric progression scoring. The underlying idea is to relate the potency distribution of active analogs forming overlapping NBHs to the mean potency of the entire AS. The measure accounts for global SAR heterogeneity and is hence termed H score. It is calculated as the difference between the weighted mean potency of active analogs forming individual clusters of overlapping NBHs and the mean potency of the complete AS (pot_i is the log potency of compound i and pot_{AS} the log potency of the AS):

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

For each active analog i, the weighting factor w_{Ni} represents the number of active analogs that form overlapping NBHs with analog i. Thus, active analogs with increasing numbers of overlapping NBHs make increasingly large contributions to the H score. We note that H can be positive or negative, depending on whether the weighted mean potency of analogs with overlapping NBHs is larger or smaller than the mean potency of the entire AS. Increasingly positive or negative H values are indicative of increasing SAR heterogeneity at the AS level. By contrast, scores close

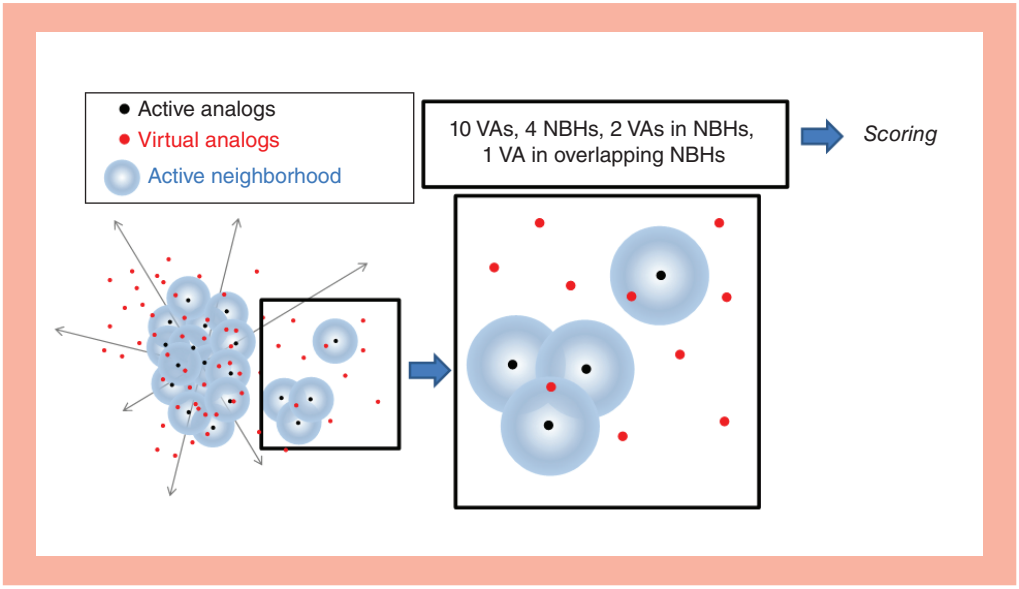


Figure 1. Methodological concept of compound optimization monitor. Shown is an analog series (black dots) with compound NBHs (gray spheres) and VAs (red dots) in n-dimensional chemical reference space. A small section of populated chemical space is enlarged, which contains four active analogs, their NBHs, and 10 VAs. Eight of the ten VAs are located outside of NBHs and two VAs map to different NBHs. Three of the four NBHs are overlapping and contain one of the VAs. This NBH and compound classification scheme provides the basis for the calculation of chemical saturation and structure–activity relationship progression scores.
NBH: Neighborhood; VA: Virtual analog.

to zero reflect low SAR heterogeneity. This characteristic renders the VA-independent H score complementary to the P score. Local SAR discontinuity, as indicated by P values, can be related to global SAR heterogeneity, as indicated by H values. By comparing these scores, potential differences between local and global SAR characteristics can be detected for ASs.

Multiproperty score

During late stages of LO, multiple optimization-relevant properties must typically be balanced while retaining potency. Therefore, we further extend the scoring scheme through the introduction of a multiproperty (M) score, which is calculated for active analogs. For scoring, descriptors of physicochemical properties of choice can be selected. In our current study, five property descriptors are chosen including the number of rotatable bonds in a molecule, the logarithmic octanol/water partition coefficient, aqueous solubility, topological polar surface area and MW. These descriptors represent a subset of those used for defining a chemical reference space, as described below. For the descriptors, preferred, acceptable and undesired value ranges are defined following the calculation of Absorption, Distribution, Metabolism, Excretion (ADME) traffic lights [14] and scored accordingly. For each active analog, property values are calculated and a penalty score of 0 (preferred), 1 (acceptable) or 2 (undesired) is assigned to each value. For MW, an ADME-relevant halogen atom correction is introduced as suggested [14]. For each compound, descriptor-based penalty scores result in a cumulative score of 0–10. For an AS, the M score is then calculated as the mean cumulative penalty score.

Chemical reference space

For profiling of ASs, a chemical reference space is required. For the assessment of chemical saturation and VA-dependent SAR progression, an intuitive, seven-dimensional descriptor space was found to yield results very similar to those obtained in higher-dimensional and more complex space representations [10]. This space was generated using descriptors accounting for molecular properties known to be relevant for ligand–target interactions including the number of hydrogen bond donors, acceptors, rotatable bonds, the logarithmic octanol/water partition coefficient, aqueous solubility, topological polar surface area and MW. The descriptors were calculated as described [9]. This chemical reference space is used herein. Distances between compounds in chemical space were calculated as the

Euclidean distance between descriptor vectors following unit-variance scaling on the basis of the VA population of a given AS.

We note that for both chemical reference space design and multiproperty scoring, different sets of descriptors can be selected, given individual preferences and/or requirements of specific applications.

Calculations

ASs with 50 or more compounds and available high-confidence activity data were extracted from ChEMBL (release 24) using a computational AS identification method [15]. Compounds of qualifying ASs were distinguished by one or more substituents. 72 ASs were obtained that were active against 35 unique targets and contained 50–148 analogs per series (a total of 5430 compounds). These ASs included 29 series with single and 43 series with multiple (two to six) substitution sites. For each AS, VAs were generated as described above.

Parameters for COMO calculations include the chemical reference space, VA design strategy, size of VA populations and the NBH radius. For the ASs used herein, test calculations were carried out by systematically varying the size of VA populations and NBH radii to further investigate the influence of these parameter settings on scoring.

Results & discussion

The COMO methodology was designed to combine computational evaluation of chemical saturation and SAR progression potential with the aid of VA populations, as illustrated in Figure 1. The use of VAs is essential for analyzing chemical space and NBH coverage as well as for assessing the density of coverage. For active analogs, NBHs are generated and overlapping NBHs are identified. Then, it is determined if VAs fall into single or overlapping NBHs, which provides the basis for calculating C, D, S and P scores. Different from P scores, complementary H scores for SAR characterization do not take VAs into account but also rely on the notion of overlapping NBHs. By contrast, M scores only depend on properties directly calculated for ASs and not on the COMO formalism. Figure 2 shows exemplary compounds from an actual AS, their NBHs and VAs. Four active analogs (black) and three exemplary VAs (red) are selected. Three active analogs form overlapping NBHs into which one of the VAs falls. In addition, another VA is located in the NBH of an isolated active analog and the third VA maps outside of the NBHs.

Virtual analogs

Because VAs play a dual role as diagnostic chemical entities as well as potential candidate compounds, their design requires careful consideration. Compared with conventional enumeration strategies for virtual libraries [16,17] and our previously applied method [9], the VA generation approach introduced herein emphasizes synthetic accessibility of VAs and a balanced size distribution. On the basis of visual inspection, these VAs are typically sound from a medicinal chemistry perspective and can be readily considered as candidates for optimization efforts.

Parameter settings

In addition to selecting a suitable chemical reference space, key calculation parameters for COMO include the size of VA populations and the radius of NBHs, as discussed above. Preferred parameter settings can be determined on the basis of test calculations. Figure 3A shows mean S scores for our ASs, VA populations of increasing size and increasing NBH radii. As one would expect, S scores tend to increase with increasing NBH radii. However, for a given radius, the scores are surprisingly stable for increasing number of VAs, which is a consequence of the inter-VA distance-dependent definition of the NBH radius. Figure 3B shows the distributions of individual S scores for increasing NBH radii in the presence of a constant number of 3000 VAs. For a percentile of 1.0, an intermediate score distribution is observed for our ASs ensemble with a median S score of close to 0.3. Figure 3C shows mean P scores for VA populations of increasing size and increasing NBH radii. Mean P scores for our ASs ensemble are distributed over a fairly narrow scoring range (from 0.4 to 0.6). For small NBH radii, mean P scores are slightly more variable than S scores for increasing numbers of VAs, but the scores become essentially constant when about 3000 (or more) VAs are used. Figure 3D reports the distributions of individual P scores for increasing NBH radii in the presence of 3000 VAs, which are much more similar to each other than the corresponding distributions of S scores. On the basis of the test calculations reported in Figure 3, the NBH radius was set to the first percentile of inter-VA distances for all subsequently reported calculations, and 3000 VAs were consistently used.

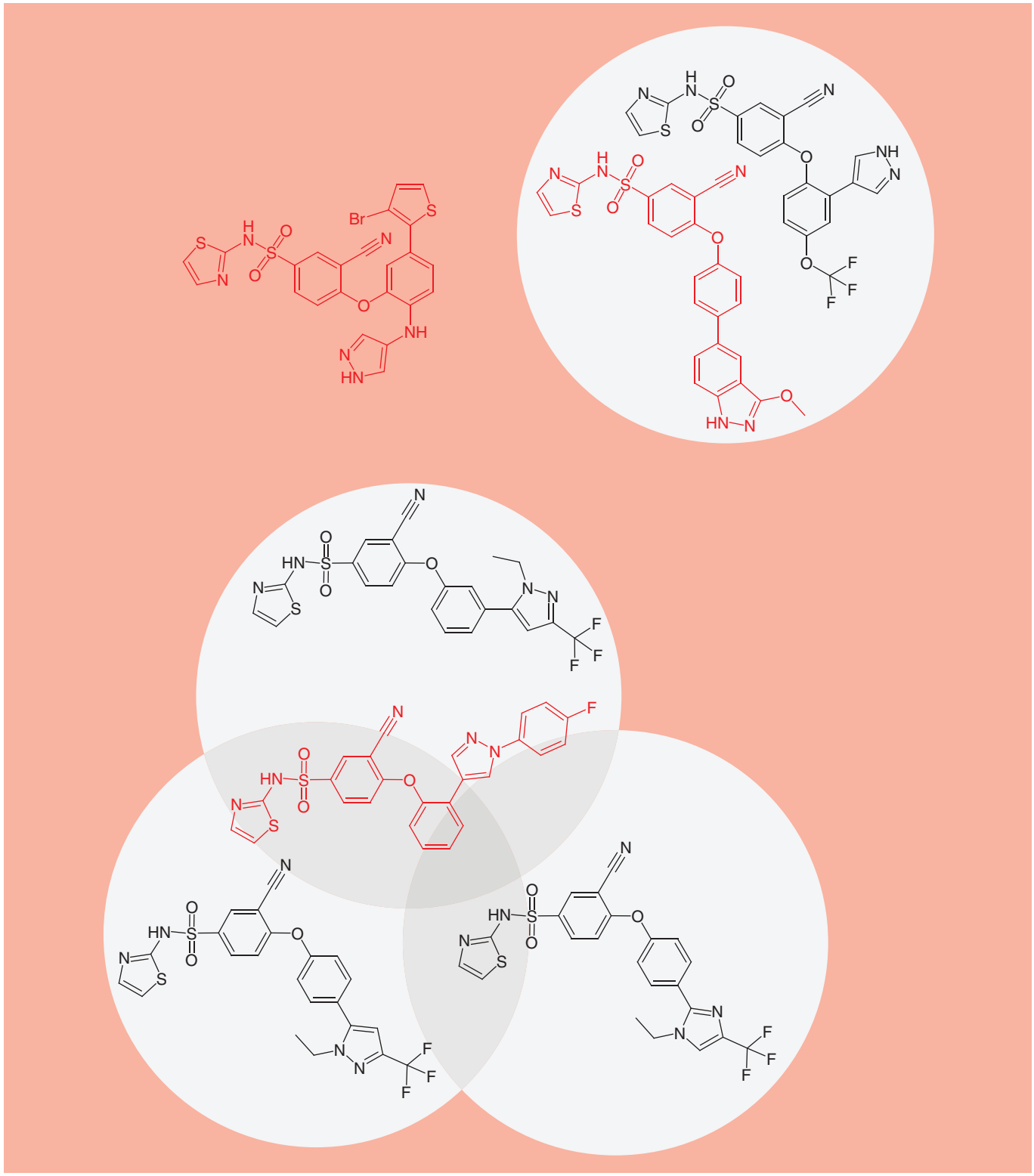


Figure 2. Exemplary active analogs, neighborhoods and virtual analogs. For an analog series of sodium channel protein type IX α subunit ligands, exemplary compounds, their neighborhoods and virtual analogs are shown according to Figure 1.

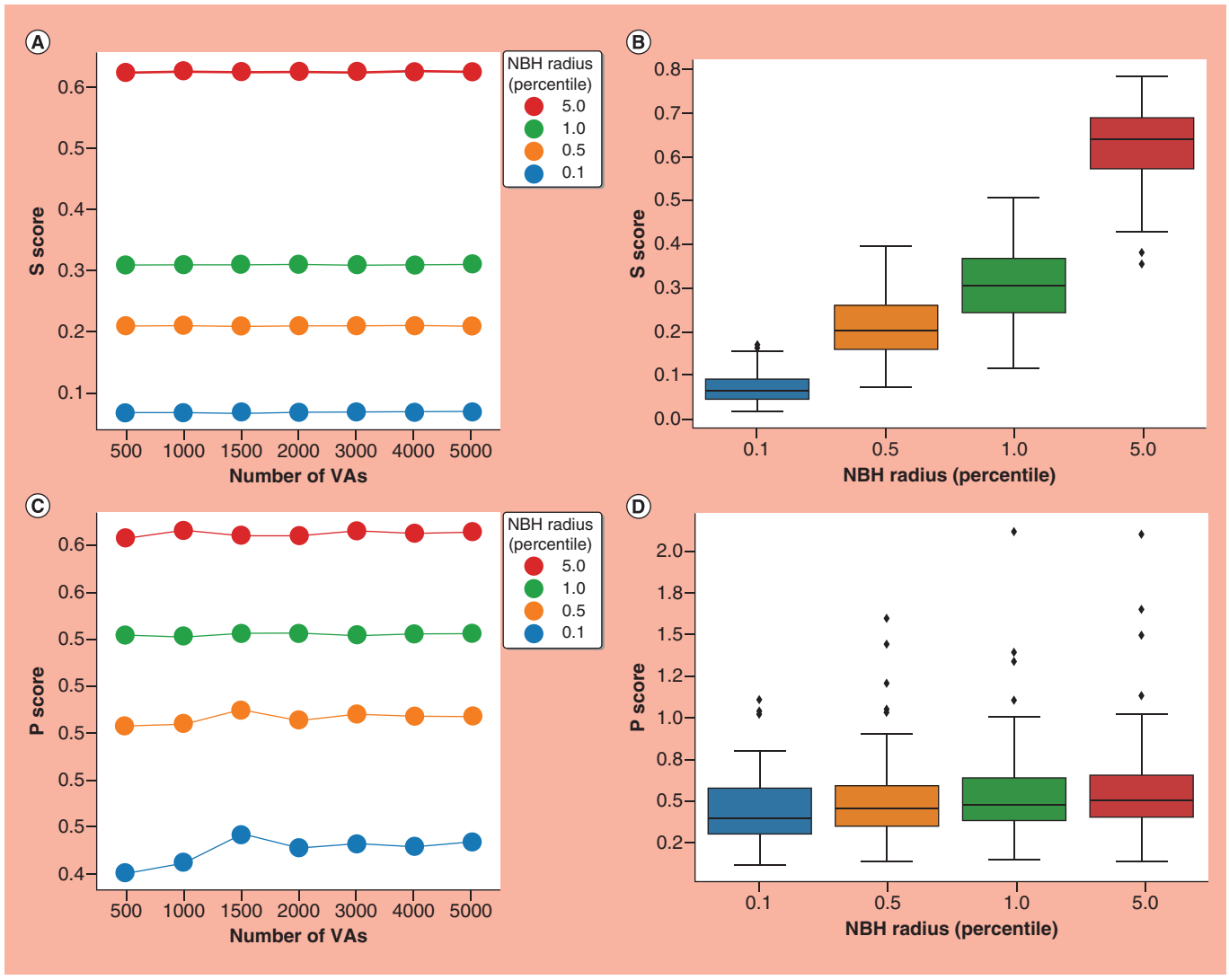


Figure 3. COMO calculation parameters. (A) Reports mean S scores for the set of 72 analog series as a function of virtual analog populations of increasing size over increasing neighborhood radii. **(B)** Shows box plots representing the distribution of S scores across all analog series for increasing neighborhood radii in the presence of a constant number of 3000 virtual analogs. **(C)** Reports mean P scores corresponding to **(A)** and **(D)** reports the distribution of P scores corresponding to **(B)**.
COMO: Compound optimization monitor; NBH: Neighborhood; VA: Virtual analog.

Score distributions

Figure 4 shows the distributions of all six COMO scores for the 72 ASs. Figure 4A compares distributions of C and D scores, which are combined to yield the S score. The distributions reveal that the ASs studied herein mostly have limited coverage of chemical reference space (i.e., a low proportion of VAs falling into their NBHs) but a high density of coverage (i.e., many VAs map to overlapping NBHs). Furthermore, P scores of the ASs ensemble preferentially populate an intermediate range (Figure 4B). By contrast, the H score is narrowly distributed close to 0, hence indicating the absence of significant SAR heterogeneity detectable with this score (Figure 4C). Nonetheless, the tendency to yield positive or negative H scores can be rationalized for these ASs, as discussed in the next section. The M scores mostly populate an intermediate range, with few outliers having high (unfavorable) scores (Figure 4D).

SAR heterogeneity

Figure 5 shows network representations for different ASs (with <60 compounds) in which analogs are represented as nodes (color coded by potency) and edges indicate the formation of overlapping NBHs. These networks show

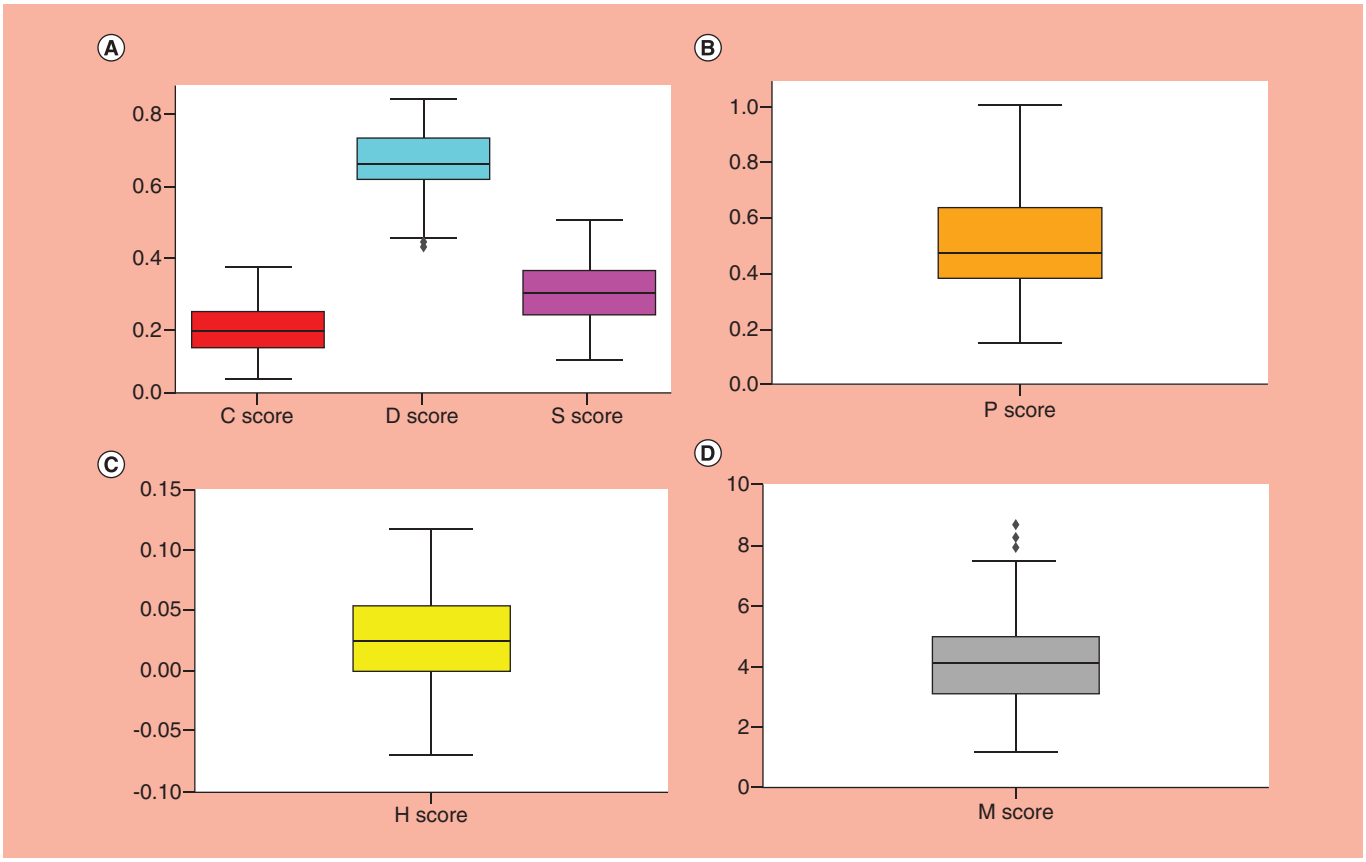


Figure 4. Score distributions. Shown are box plots representing the distributions of the six compound optimization monitor scores for all 72 analog series calculated using a constant neighborhood radius (first percentile) and 3000 virtual analogs.

that only a fraction of analogs have overlapping NBHs, which is an important observation from a methodological viewpoint. In addition, the networks reveal possible origins of SAR heterogeneity. For example, the network of the AS in Figure 5A contains two clusters of densely connected and mostly weakly potent analogs, which results in a negative H score. The network in Figure 5B reveals clusters of compounds with varying potency, which essentially mirror the potency distribution across the AS, resulting in an H score close to 0. By contrast, the network in Figure 5C contains clusters formed by mostly highly potent analogs and, in addition, a large number of singletons with varying potency. The clusters with potent analogs are responsible for producing a positive H score. Hence, increasing SAR heterogeneity detected by H scoring is straightforward to rationalize on the basis of network views. Scoring of larger ASs than those currently available (i.e., ASs with more extensive cluster formation) will help to determine if the H score should be numerically adjusted.

Score comparison

We next compare different COMO scores for individual ASs. Figure 6A shows the comparison of C and D scores. Different combinations are observed for ASs of varying size and, importantly, no correlation is detectable between these scores. This confirms that coverage of chemical space and the density of coverage are independent properties, which can contribute differently to the S score. In addition, Figure 6B compares P and H scores. ASs with increasing P score predominantly – but not exclusively – display positive H scores, indicating that compounds with overlapping NBHs on average exceed the potency of the entire AS; an interesting observation. Thus, nearest neighbors in an AS tend to have above average potency, which likely reflects the generation of close-in analogs once a potent compound is identified.

Figure 6C shows the comparison of S and P scores, which are central components of the COMO methodology. Importantly, no correlation between these scores is observed and ASs of similar size display different scores. The absence of correlation is a prerequisite for an unbiased assessment of chemical saturation and SAR progression.

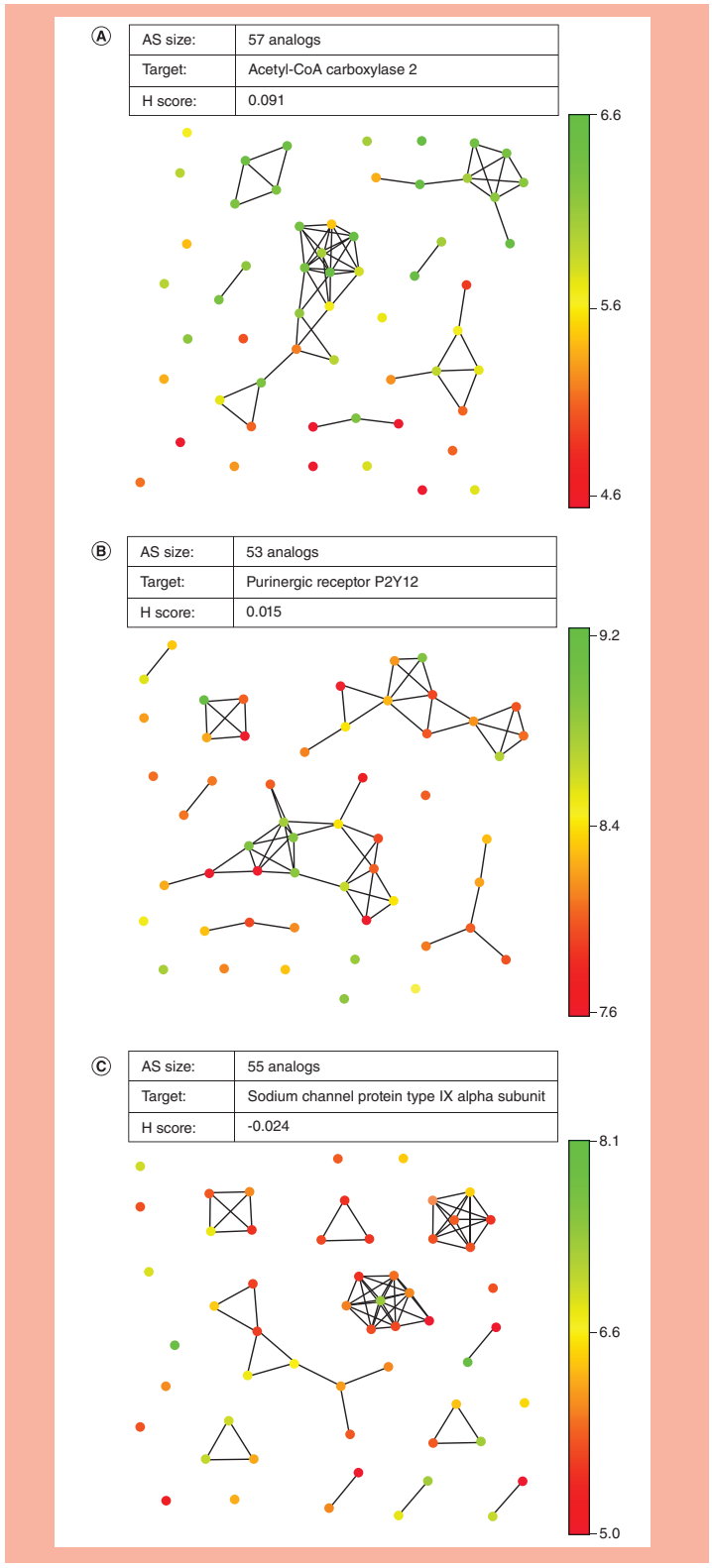


Figure 5. Neighborhood overlap-based analog networks. For different analog series, network representations are shown in which nodes represent compounds. Nodes are color coded by logarithmic compound potency (K_i or IC_{50}) values using a continuous color spectrum as indicated. In addition, edges between nodes indicate that the corresponding compounds have overlapping neighborhoods. In each case, the target of the analog series is specified, the number of analogs given and the H score reported. **(A)** Acetyl-CoA carboxylase 2 inhibitors, **(B)** purinergic receptor P2Y12 ligands and **(C)** sodium channel protein type IX α subunit ligands. For clarity, networks were drawn on the basis of a smaller neighborhood radius (0.1st percentile) than used for score calculations, which reduced the number of overlapping neighborhoods. Networks were computed with the Python wrapper of the Graphviz software using the 'neato' network layout [18].

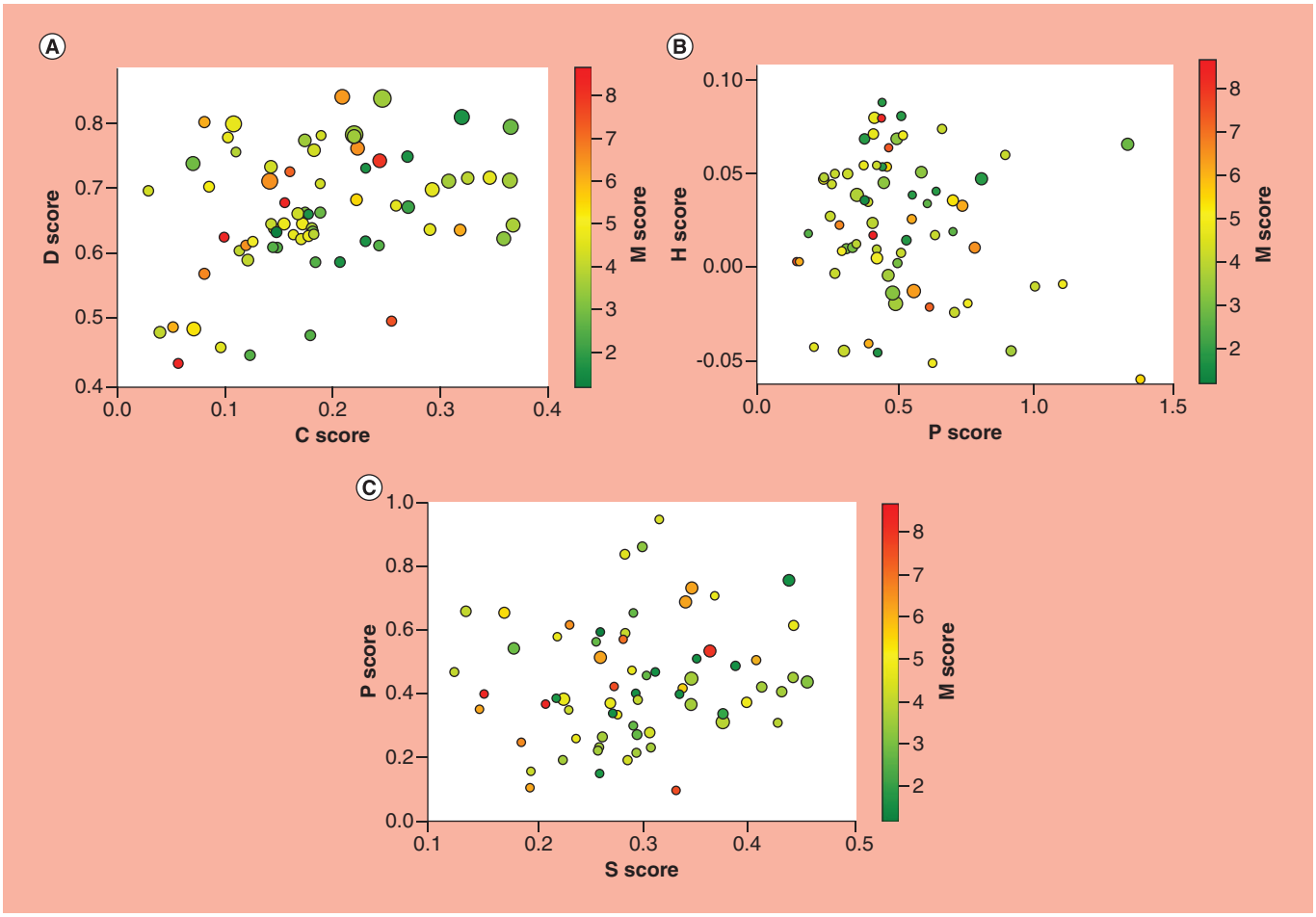


Figure 6. Score comparison. Scatter plots compare the distributions of different COMO scores for the set of 72 AS. Each dot reprscoresesents an AS. Dots are scaled in size according to the number of compounds per AS and color coded according to M scores using a continuous color spectrum as indicated. **(A)** C versus D, **(B)** P versus H and **(C)** S versus P scores.
AS: Analog series. COMO: Compound optimization monitor.

However, one would also expect a tendency that increasing numbers of active analogs should increase the degree of chemical saturation of an AS. Although the magnitude of such effects is influenced by the compound class under study and the number of substitution sites per AS, our findings also reflect this expectation. For example, 15 of the 18 ASs with highest S scores (representing the highest quartile of the S score distribution) exceed the median number of 66 analogs per AS. These 15 ASs contain nine of a total of 14 ASs with >100 analogs. Moreover, it is also important to note that the S- and P-score combinations cover wide scoring ranges, hence indicating high differentiation potential for the small to moderately sized ASs studied here, lending credence to the scoring scheme.

Figure 6 also shows that ASs have rather different M scores, ranging from favorable to unfavorable scores, and that these scores are not related to other COMO scores, as expected.

Score interpretation

On the basis of characteristic combinations of chemical saturation and SAR progression scores, ASs can be assigned to different LO stages, as illustrated in Figure 7. The ASs falling into the lower left quadrant of the plot are characterized by low S and low P scores. It follows that these ASs are still little explored chemically and do not display detectable SAR progression. Such series are at very early stages of chemical exploration and must be further extended and to better understand their potential. Furthermore, ASs in the upper left quadrant have low S and high P scores. Hence, these series are also still at early stages of chemical exploration, but already display significant potential for SAR progression. Accordingly, they represent promising candidates for further development.

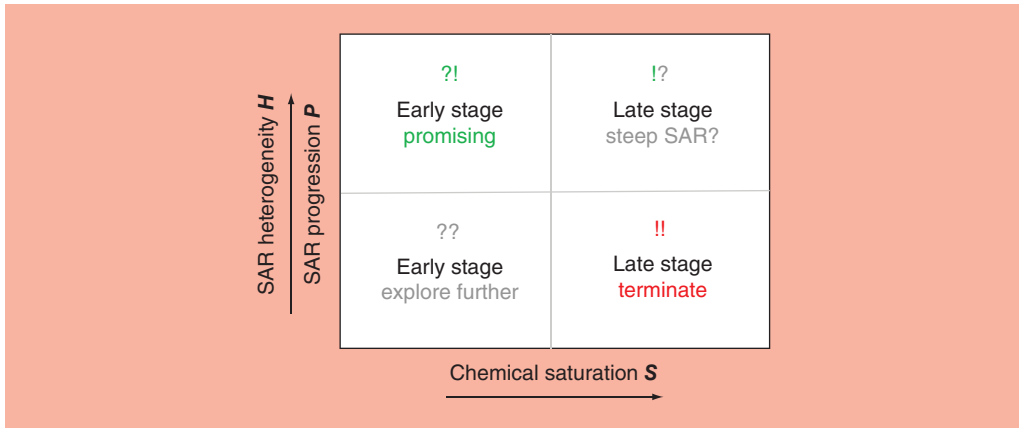


Figure 7. Interpretation of score combinations. The schematic representation illustrates combinations of COMO scores of different magnitude and their interpretation. Characteristic score combinations are used to assign analog series to different lead optimization stages.
COMO: Compound optimization monitor.

ASs in the upper right quadrant have high S and P scores, indicating that they are chemically far advanced but still have potential for SAR progression. This observation can be interpreted in different ways; for example, by generating additional analogs, further SAR progression might occur and more potent compounds might be identified. On the other hand, this score combination might also be indicative of steep SARs at late stages of LO. Such SARs features are undesired when multiple properties must be balanced while retaining compound potency. Therefore, caution is advised when ASs with high S and high P scores are detected, and follow-up analyses should be considered. For example, SAR responses of bioisosteric replacements should then be carefully analyzed during multiproperty optimization.

Finally, ASs in the lower right quadrant of the plot have low P and high S scores. Accordingly, they are chemically saturated and display very little potential for further SAR progression. Thus, given the absence of potential caveats associated with steep SARs, such ASs can serve as a basis for ADME-oriented multiproperty optimization once a desirable potency level of the lead candidate(s) has been achieved. Multiproperty optimization is supported by multiproperty scoring, as reported herein, and relies on retaining desirable potency levels, which is favored by low remaining SAR discontinuity. However, ASs with high S and low P scores may also represent candidates for discontinuation if no highly potent compound(s) have been identified, despite the extensive saturation of analog space. Furthermore, in some instances, additional help in judging whether or not desirable potencies levels have been achieved is provided by positive or negative H score, which support decision making during later stages of LO.

Conclusion

The basic idea underlying the COMO approach is rationalizing LO efforts beyond subjective judgment, especially considering key questions whether or not enough compounds have been generated or further progress is likely. Therefore, COMO is designed to characterize ASs by combining the assessment of chemical saturation and SAR progression. This is facilitated through a scoring scheme comprising two pairs of complementary chemical saturation- and SAR-relevant scores. As an additional diagnostic, a multiproperty score is calculated for test compounds. COMO analysis can be conveniently carried out when ASs evolve over time and new compounds are added. Hence, progress can constantly be monitored and detected changes further analyzed.

Currently, there are no related computational approaches available. Hence, from this point of view, the COMO methodology is charting new territory in computational medicinal chemistry. As shown herein, the COMO scoring scheme distinguishes between different ASs and the scores are chemically interpretable. Moreover, the analysis of score combinations makes it possible to assign ASs to different LO stages and prioritize series for termination or further exploration. The COMO methodology has been extensively evaluated internally on ASs extracted from public domain resources, and results obtained so far indicate considerable potential for further practical applications.

Future perspective

In its current implementation, COMO captures the results of several efforts to develop new computational concepts for the assessment of chemical saturation and SAR progression. These concepts have been translated into an advanced scoring system to quantitatively assess LO progress. Of course, as is the case with any computational methodology, the COMO framework will be subject to further development and extension. For example, it is conceivable that the scoring scheme will be further refined once large ASs become available for profiling. Notably, the ASs we are currently able to extract for benchmark calculations from publicly available compounds are generally limited in size. For example, the ensemble of ASs used herein only contains only a limited number of series with >100 analogs. Nonetheless, in our previous proof-of-concept investigation [9] and our current study, individual ASs with large differences between chemical saturation and/or SAR progression scores have already been detected.

Another aspect to consider is that publicly available ASs might often originate from different sources and therefore reflect practical optimization efforts only to a limited extent. In drug discovery, LO campaigns often produce much larger ASs than investigated herein and it will be interesting to subject such series to comparative COMO analysis. Furthermore, profiling of evolving ASs following the sequence of optimization efforts will also be of considerable interest. Here, decision support provided by computational analysis might have an immediate impact. Indications are that the methodology has matured to the point that such applications can be carried out.

An attractive area for future research will be further exploiting the dual role of VAs as diagnostic chemical entities and potential candidates for chemical optimization, for which the current VA generation approach provides a foundation. For example, initial efforts are currently underway to combine COMO with machine-learning approaches to derive models for activity prediction. For moderately sized ASs, this is already feasible. Such models will then be used to predict VAs having the highest probability of activity and highest potential for further SAR progression, thus adding a compound design component to COMO's diagnostic repertoire.

Executive summary

Background
- Lead optimization (LO) is largely driven by chemical intuition and experience.
- Progress in LO is difficult to evaluate.
- Only a few computational methods are available to monitor LO.
- New computational concepts are required to provide decision support.

Methodology & calculations
- Compound Optimization Monitor (COMO) is introduced as a new approach for quantifying LO progress.
- The key question is addressed if enough compounds have been made.
- COMO's methodological concept and its scoring scheme are detailed.
- A new approach for the generation of virtual analogs is introduced.

Results & discussion
- Calculation parameters are evaluated.
- COMO results are presented for an ensemble of 72 analog series (AS).
- Score distributions are analyzed and compared.
- COMO-based assignment of ASs to different LO stages is discussed.

Future perspective
- The computational concept is subject to further extension.
- ASs from the public domain often have limited exploration potential.
- Practical applications on ASs from drug discovery will be a focal point.
- Combining COMO with machine learning is a topic for future research.

Author contributions

J Bajorath conceived the study; D Yonchev and M Vogt implemented the methods; D Yonchev carried out the analysis; D Yonchev, M Vogt and J Bajorath analyzed the results; J Bajorath prepared the manuscript; and all authors reviewed the manuscript.

Acknowledgments

The authors thank D Stumpfe for help with illustrations.

Financial & competing interests disclosure

D Yonchev is supported by the Jürgen-Manchot-Stiftung, Düsseldorf. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

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

References

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

1. *The Practice of Medicinal Chemistry (3rd Edition).* Wermuth CG (Ed). Academic Press-Elsevier, CA, USA (2008).
2. Lill MA. Multi-dimensional QSAR in drug discovery. *Drug Discov. Today* 12(23–24), 1013–1017 (2007).
3. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* 20(3), 318–331 (2015).
4. Segall M. Advances in multi-parameter optimization methods for *de novo* drug design. *Expert Opin. Drug Discov.* 9(7), 803–817 (2014).

• **Review of multiparameter optimization approaches.**

5. Munson M, Lieberman H, Tserlin E *et al.* Lead optimization attrition analysis (LOAA): a novel and general methodology for medicinal chemistry. *Drug Discov. Today* 20(8), 978–987 (2015).
6. 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.* 59(9), 4235–4244 (2015).

• **Computational diagnostic for evaluating structure–activity relationship progression.**

7. Maynard AT, Roberts CD. Quantifying, visualizing, and monitoring lead optimization. *J. Med. Chem.* 59(9), 4189–4201 (2015).

•• **Statistical framework for identifying key compounds during lead optimization.**

8. Kunimoto R, Miyao T, Bajorath J. Computational method for estimating progression saturation of analog series. *RSC Adv.* 8(10), 5484–5492 (2018).

•• **Introducing a computational concept for chemical saturation analysis.**

9. Yonchev D, Vogt M, Stumpfe D, Kunimoto R, Miyao T, Bajorath J. Computational assessment of chemical saturation of analogue series under varying conditions. *ACS Omega* 3(11), 15799–15808 (2018).
10. Vogt M, Yonchev D, Bajorath J. Computational method to evaluate progress in lead optimization. *J. Med. Chem.* 61(23), 10895–10900 (2018).

•• **Combining chemical saturation and structure–activity relationship progression analysis.**

11. Gaulton A, Bellis LJ, Bento AP *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40(D1), D1100–D1107 (2012).
12. Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* 50(3), 339–348 (2010).

• **Bond fragmentation algorithm for systematic generation of matched molecular pairs.**

13. Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP – retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 38(3), 511–522 (1998).
14. Lobell M, Hendrix M, Hinzen B *et al. In silico* ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem.* 1(11), 1229–1236 (2006).
15. Naveja JJ, Vogt M, Stumpfe D, Medina-Franco JL, Bajorath J. Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS Omega* 4(1), 1027–1032 (2019).
16. Leach AR, Hann MM. The *in silico* world of virtual libraries. *Drug Discov. Today* 5(8), 326–336 (2000).
17. Walters WP. Virtual chemical libraries. *J. Med. Chem.* 62(3), 1116–1124 (2019).
18. Gansner ER, North SC. An open graph visualization system and its applications to software engineering. *Software Pract. Exper.* 30(11), 1203–1233 (2000).