Integrating computational lead optimization diagnostics with analog design and candidate selection

Dimitar Yonchev1 & Jürgen Bajorath*,1
1Department 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 7369 100; Fax: +49 228 7369 101; bajorath@bit.uni-bonn.de

Aim: Combining computational lead optimization diagnostics with analog design and computational approaches for assessing optimization efforts are discussed and the compound optimization monitor is introduced. Methods: Approaches for compound potency prediction are described and a new analog design algorithm is introduced. Calculation protocols are detailed. Results & discussion: The study rationale is explained. Compound optimization monitor diagnostics are combined with a thoroughly evaluated approach for compound design and candidate prioritization. The diagnostic scoring scheme is further extended. Future perspective: Opportunities for practical applications of the integrated computational methodology are described and further development perspectives are discussed.

Lay abstract: Compound optimization is a central task in medicinal chemistry, which has many potential pitfalls. Computational approaches that help to better understand and guide chemical optimization efforts are highly desirable, but only a few are currently available. We have aimed to develop a computational methodology that combines, for the first time, the evaluation of progress in chemical optimization with the design of new candidate compounds.
Chemical optimization efforts play a central role in the practice of medicinal chemistry [1]. During lead optimization (LO), many analogs of initially prioritized active compounds must typically be generated until candidate status is reached. However, despite large compound numbers, work on analog series (ASs) must often be terminated when required optimization criteria cannot be reached. The need to abandon large-magnitude LO efforts results in a significant loss of time and resources, which causes major problems for medicinal chemistry. Accordingly, any approaches that help to evaluate LO projects and estimate the odds of success are highly desirable. For this purpose, computational evaluation of LO is particularly attractive. However, in addition to quantitative structure–activity relationship (QSAR) approaches that are long used to predict the potency of newly designed analogs [2,3], only few computational methods are currently available that aid in planning or assessing LO efforts [4–9]. These methods include multiparameter optimization and other statistical techniques to evaluate compound property progression or identify compounds that strongly contribute to structure–activity relationships (SARs) [4–7]. None of these approaches provide comprehensive LO diagnostics or combines data analysis with molecular design. Recently, a conceptually different computational methodology has been introduced to address the questions if an AS might be chemically saturated and if further SAR progression might be expected [8,9]. The analysis makes it also possible to estimate if sufficient numbers of analogs have been generated for a given series. Hence, the evaluation of chemical saturation and SAR progression was combined to provide decision support during LO [8]. These efforts have led to the development of the compound optimization monitor (COMO) program [9]. By design, the COMO approach is diagnostic in nature, similar to two other SAR evaluation methods [6,7]. However, a special feature of COMO’s additional chemical saturation analysis component is that it utilizes populations of virtual analogs (VAs) to chart chemical space for given ASs. These VAs are specifically generated for each AS and might thus also be evaluated as candidate compounds for synthesis. Accordingly, the COMO method might be further extended to compound design and the prediction of preferred candidates. This would provide a unique methodological combination of chemical saturation and SAR diagnostics with prospective compound design. However, achieving this goal requires the incorporation of approaches for AS-specific VA selection and candidate prediction. Herein, we report the extension of COMO to include the design and prioritization of candidate compounds for LO.

Methods

COMO diagnostic concept

The COMO approach, its scoring scheme and parameter optimization have been described in detail [9]. In the following, a summary of the COMO concept is presented as a basis for rationalizing its extension.

COMO evaluates chemical saturation and SAR progression of ASs by determining how extensively and densely chemical space around a given series is covered. In addition, COMO determines if significant potency variations among existing analogs (EAs) and increases in potency are observed during LO. The assessment relies on defining chemical neighborhoods (NBHs) of EAs and on using populations of VAs for given ASs to map NBHs and surrounding chemical space. VAs are currently generated using a pool of more than 32,000 unique substituents with at most 13 heavy atoms that were extracted from bioactive compounds on the basis of retrosynthetic criteria. For VA generation, the core structure of an AS is isolated while retaining substitution site information through atom indices. Then, predefined numbers of VAs are enumerated according to retrosynthetic rules using the substituent library. At individual substitution sites, an AS-specific likelihood of hydrogen substituents is taken into account [9]. All EAs and the corresponding VA population are then projected into a chemical reference space where overlapping and nonoverlapping NBHs of EAs are analyzed and their VA content is determined. Then, potency variations of EAs with overlapping and populated NBHs are quantified.

This analysis concept yields multiple scores for evaluating LO progression. COMO key scores account for chemical saturation and SAR progression. The chemical saturation score S is composed of two components quantifying the coverage and density of chemical space.
The coverage score \( C \) is defined as the proportion of VAs that populate NBHs of EAs:

\[
C = \frac{n_N}{n_V} \quad \text{(Eq. 1)}
\]

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

In addition, a term \( d_{\text{mean}} \) is introduced as the number of overlapping NBHs containing VAs (\( NBH_{O,VA} \)) relative to the total number of VAs falling into NBHs of EAs:

\[
d_{\text{mean}} = \frac{NBH_{O,VA}}{n_N} \quad \text{(Eq. 2)}
\]

The density score \( D \) with range \([0,1]\) is then calculated as:

\[
D = 1 - \frac{1}{d_{\text{mean}}} \quad \text{(Eq. 3)}
\]

Chemical saturation score \( S \) combines coverage and sampling density of chemical reference space and is obtained as the harmonic mean of score components \( C \) and \( D \):

\[
S = \frac{2CD}{C + D} \quad \text{(Eq. 4)}
\]

Furthermore, SAR progression is assessed by determining potency variations of EAs sharing VAs in overlapping NBHs, which provides a measure of SAR discontinuity of a given AS. For a given VA, parameter \( \overline{\Delta}_i \) accounts for the potency range among \( m_i \) associated analogs. It is calculated as the mean potency difference over all pairs of \( m_i \) EAs. In addition, \( pot_j \) and \( pot_k \) represent the logarithmic potency of analog \( j \) and \( k \), respectively:

\[
\overline{\Delta}_i = \frac{2}{m_i(m_i - 1)} \sum_{j,k=1}^{m_i} \left| pot_j - pot_k \right| \quad j < k \quad \text{(Eq. 5)}
\]

The SAR progression score \( P \) is then calculated as the mean over all VAs in NBHs using their \( \overline{\Delta}_i \) values and a weighting scheme \( w_i = \frac{1}{m_i} \) if \( m_i > 1 \) and \( w_i = 0 \) if \( m_i = 1 \):

\[
P = \frac{1}{\sum_{i=1}^{n_N} w_i} \sum_{i=1}^{n_N} w_i \overline{\Delta}_i \quad \text{(Eq. 6)}
\]

The COMO calculations reported herein were carried out as described previously [9] using a seven-dimensional (7D) chemical reference space and a population of 2000 VAs per AS. In each case, \( S \) and \( P \) scores were calculated to illustrate the characterization of ASs.

AS

For our analysis, new ASs with activity against a given target and available high-confidence potency measurements of inhibition constant (\( K_i \)) or half maximal inhibitory (IC\(_{50}\)) values were extracted from ChEMBL (version 25) [10]. The ASs were identified using a previously reported algorithm [11] following matched molecular pair (MMP) fragmentation [12,13] of bioactive compounds on the basis of retrosynthetic rules [14,15]. An MMP is defined as a pair of compounds that are only distinguished by a chemical modification at a single site [12]. Systematic fragmentation of exocyclic single bonds generates MMP cores and substituent fragments. Following fragmentation, the AS identification algorithm assembles series with a shared core and single or multiple substitution sites [11]. For our analysis, the 24 largest ASs with more than 100 compounds (max. 264) and multiple (two to six) substitution sites were considered. They contained more compounds than previously investigated ASs and were active against 16 distinct targets. Table 1 summarizes their composition.
### Methodology

#### Yonchev & Bajorath

**Table 1. Analog series.**

| AS ID | Target name                          | ChEMBL Target ID | # Subst. Sites | # EAs | # FW EAs | # FW VAs |
|-------|--------------------------------------|------------------|----------------|-------|----------|----------|
| 1     | Serine/threonine-protein kinase mTOR | 2842             | 2              | 153   | 27       | 264      |
| 2     | Acetyl-CoA carboxylase 2             | 4829             | 2              | 112   | 0        | 218      |
| 3     | Acetyl-CoA carboxylase 2             | 4829             | 6              | 149   | 33       | 3812     |
| 4     | Adenosine A2b receptor               | 255              | 6              | 129   | 72       | 392      |
| 5     | GABA receptor alpha-5 subunit        | 5112             | 2              | 193   | 118      | 1647     |
| 6     | Purinergic receptor P2Y12            | 2001             | 2              | 237   | 145      | 2766     |
| 7     | Vanilloid receptor                   | 4794             | 3              | 162   | 0        | 0        |
| 8     | Mitogen-activated protein kinase kinase 12 | 1908389   | 2              | 111   | 44       | 1844     |
| 9     | S-lipoxygenase activating protein    | 4550             | 3              | 259   | 162      | 5204     |
| 10    | S-lipoxygenase activating protein    | 4550             | 2              | 100   | 4        | 96       |
| 11    | Epidermal growth factor receptor erbB1 | 203          | 2              | 106   | 0        | 306      |
| 12    | Sodium channel protein type IX alpha subunit | 4296    | 3              | 146   | 40       | 2860     |
| 13    | Acetyl-CoA carboxylase 2             | 4829             | 3              | 100   | 10       | 145      |
| 14    | Proteinase activated receptor 4      | 4691             | 5              | 117   | 8        | 212      |
| 15    | Acetyl-CoA carboxylase 2             | 4829             | 3              | 128   | 54       | 319      |
| 16    | p53-binding protein Mdm-2            | 5023             | 4              | 149   | 72       | 381      |
| 17    | Acetyl-CoA carboxylase 2             | 4829             | 3              | 116   | 81       | 331      |
| 18    | Sodium channel protein type IX alpha subunit | 4296      | 5              | 151   | 17       | 1736     |
| 19    | P2X purinoceptor 3                   | 2998             | 6              | 102   | 84       | 336      |
| 20    | MAP kinase ERK2                      | 4040             | 2              | 264   | 0        | 262      |
| 21    | Tyrosine-protein kinase SYK          | 2599             | 5              | 173   | 43       | 5755     |
| 22    | Prostaglandin E synthase             | 5658             | 3              | 168   | 63       | 3893     |
| 23    | Tyrosine-protein kinase SYK          | 2599             | 3              | 168   | 72       | 1887     |
| 24    | S-lipoxygenase activating protein    | 4550             | 2              | 126   | 0        | 124      |

The table summarizes the composition of ASs used herein and reports the proportion of existing analogs and newly generated virtual analogs that qualify for Free-Wilson potency prediction, as discussed in the text. ‘# Subst. Sites’ reports the number of substitution sites per AS.

AS: Analog series; EA: Existing analog; FW: Free-Wilson; ID: Identification; VA: Virtual analogs.

#### Linear & ridge regression

Linear regression (LR) is the simplest and most widely used statistical approach for numerical value predictions [16]. In QSAR modeling, LR is applied assuming the presence of linear relationships between numerical chemical features and biological activity [3]. The predictive performance of LR models inevitably suffers from outliers [16] and has limited predictive ability in the presence of nonlinear SARs in training and/or test sets [3]. To address the outlier problem, a penalty on model weights can be introduced. This requires optimizing the penalized residual sum of squares defined as:

\[
\min_w ||X_w - y||^2 + \alpha||w||^2
\]  

(Eq. 7)

Here, \(X_w\) is the estimated target value, \(y\) the true target value, \(w\) the weighting coefficient and \(\alpha\) the regularization parameter determining regularization strength. This regularized least squares LR approach is generally referred to as ridge regression (RR) [17,18], which was applied herein as an advanced LR technique.

#### Support vector regression

Support vector machine (SVM) [16,19] is a supervised machine learning algorithm that is widely used in chemical informatics [3]. SVM was originally introduced as a method for binary object classification (class label prediction) and ranking. The SVM algorithm aims to separate positive and negative training instances in a given feature space via a hyperplane having the largest possible margin [19]. If linear separation is not possible in a given feature space, kernel functions are applied to project the training data into higher dimensional feature spaces where linear separation might become possible [19].
Support vector regression (SVR) \[20\] is a variant of the SVM approach. Instead of optimizing a separating hyperplane for classification, a regression function is derived for predicting numerical values:

\[
 f(x) = \sum_i (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{Eq. 8}
\]

Here, \(\alpha\) and \(\alpha_i^*\) are support vectors representing the vector \(w\) derived from a convex optimization procedure, \(K(x_i, x)\) is the kernel function applied to the input feature vectors and \(b\) the bias parameter derived from the convex optimization procedure \[20,21\]. SVR is capable of fitting a LR function for nonlinear SARs by increasing the feature space dimensionality. Therefore, it has become a method of choice for nonlinear QSAR modeling and potency prediction \[3\]. Critical parameters during model building include the regularization term \(C\) and the \(\epsilon\)-insensitive tube \[20,21\]. The \(\epsilon\) parameter determines the maximally permitted prediction error during training and regularization term determines the trade-off between model complexity and error penalization.

**Free-Wilson formalism**

Free-Wilson (FW) analysis is based upon the premise that chemical modifications in series of compounds are independent of each other and that associated potency changes are additive \[22,23\]. The additivity assumption represents an approximation because there might also be cooperativity between substitution sites. However, in practice, the FW approximation often holds, providing a basis for meaningful compound potency predictions \[23\]. Principles of FW analysis are illustrated in Figure 1. Exemplary EAs with activity against the GABA receptor alpha-5 subunit are shown and their experimentally measured logarithmic Ki values are given. Structural relationships between EAs were established by searching for MMPs. In this example, analog A forms distinct MMP relationships with analog B and C as a consequence of structural modifications at the first and second substitution site (R1 and R2), respectively. By contrast, analogs B and C do not form an MMP because they differ at both substitution sites. Analog B and C share the same substituent at R2 while the pyridine ring at R1 in A is fluorinated in B, which results in a potency increase \(\Delta pK_i = +0.3\). Conversely, analogs A and C share the same substituent at R1 while the methyl ester function at R2 in A is replaced by a tri-fluoro ethyl amide in C. This modification results in a potency increase of \(\Delta pK_i = +0.6\) for analog C. Combining the structural modifications that convert analog A to B and A to C, respectively, results in a new analog X. This analog forms an MMP with B and C, respectively, but not with analog A that differs from X at two sites. Following FW principles, the potency of analog X can be predicted on the basis of analog A by adding the potency changes accompanying the conversions of A to B and C, respectively, as illustrated in Figure 1. Accordingly, the pKi value predicted for analog X is 9.3 (i.e., \(8.4 + 0.3 + 0.6\)). A, B and C form the Free-Wilson neighborhood (FW NBH) of analog X. For a given FW target compound (such as X), multiple qualifying FW NBHs may exist. In this case, potency predictions are typically averaged over all qualifying FW NBHs.

For this FW prediction example, analog X was ‘virtualized’ (i.e., considered as a VA) since it also belonged to the AS with activity against the GABA receptor alpha-5 subunit. The predicted value of pKi = 9.3 was only slightly lower than the experimentally observed potency of pKi = 9.5, illustrating the utility of FW predictions when the additivity approximation applies.

**Generation of FW analogs**

To complement COMO-derived VA populations a new algorithm was implemented to generate VAs suitable for FW analysis (termed FW VAs). The FW VA algorithm consists of the following steps:

- **Step 1**: Given an AS core structure with indexed substitution sites, all EAs and their site-specific substituents are collected.
- **Step 2**: For EAs, all possible MMPs are generated and organized in a MMP network (using the Python Networkx package \[24\]) where nodes represent EAs and edges pairwise MMP relationships.
- **Step 3**: For each MMP, exchanged and conserved substituents are stored and assigned to the MMP edge in the network. Exhaustive search for FW NBHs (according to Figure 1) is performed and detected FW NBHs are stored.
- **Step 4**: For each FW NBH, the direction of the MMP-defining substitution exchanges is determined according to Figure 1 (i.e., A to B, A to C) and the corresponding newly introduced substituents are recorded.
- **Step 5**: For each FW NBH, new substituents are added to the AS core (using the OpenEye toolkit \[25\]) generating a new FW VA for the NBH.
- **Step 6**: Unique FW VAs associated with one or more FW NBHs are retained.
Figure 1. Principles of Free-Wilson analysis. Shown are four analogs from the same AS that are active against GABA receptor alpha-5 subunit (AS 5; ChEMBL target ID 5112). For each compound, its logarithmic experimental potency ($p_{Ki}$) value is reported. In addition, the core structure of the AS is depicted in the center and the two substitution sites $R_1$ and $R_2$ are highlighted in yellow and blue, respectively. Corresponding substituents in analogs are colored accordingly. Individual potency contributions of directed substitutions are reported as $\Delta p_{Ki}$ values. The figure illustrates the principles of Free-Wilson predictions of compound potency.

AS: Analog series.

FW NBHs define existing FW analogs (FW EAs) that are also sampled.

For each AS, varying numbers of FW NBHs, FW EAs and FW VAs were obtained, depending on the underlying MMP distribution. The calculations identified FW EAs for further analysis and generated FW VAs for potency prediction and candidate selection.

Potency predictions

Regression models

For each AS, QSAR models using RR and SVR were independently generated via double (internal and external) cross-validation [26] using scikit-learn [27]. For each analog, the extended connectivity fingerprint with bond diameter 4 [28] was calculated and folded into a 1024-bit feature vector using RDKit [29] as a molecular representation. Initially, each AS was randomly partitioned into training/test data (80%) and external validation sets (20%) 35 times to ensure statistically sound model evaluation. In addition, it was monitored that the potency of each FW EA from a given series was externally predicted at least once using RR and SVR. Training and test data were subjected to fivefold internal cross-validation. During internal cross-validation optimal hyper-parameters were selected for each model. These hyper-parameters were subsequently used for prediction of the external validation set for the same independent trial. For SVR, a parameter grid of 18 $C$ and 5 $\epsilon$ values was optimized in combination with the Tanimoto kernel [30]. For RR, seven different $\alpha$ values were tested during hyper-parameter optimization. The RR
and SVR models were also used for predicting the potency of FW VAs for each AS. Predictions were averaged over all models.

**FW predictions**

For each AS, compounds forming FW NBHs were identified. Each participating FW EA was virtualized and its potency was predicted as the mean over all FW NBHs in which it occurred. Analogous predictions were carried out for newly generated AS-specific FW VAs.

**Model evaluation**

The performance of QSAR models can be evaluated using different statistical measures [3]. Herein, the coefficient of determination termed \( R^2 \) was used as the most popular measure, which is defined as:

\[
R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{Eq. 9}
\]

Here, \( y_i \) is the true value of instance \( i \), \( f_i \) the predicted value of instance \( i \), and \( \bar{y} \) the mean of all true test instance values. The numerator represents the residual sum of squares and the denominator is the total sum of squares. The maximal value of \( R^2 \) is 1, which results from perfect correlation between predicted and true values. A value of 0 (or negative value) for \( R^2 \) means that the performance of a model is equal to (or worse than) simple value averaging.

**Results & discussion**

**Study goal**

COMO was originally designed as a diagnostic approach to aid in the evaluation of LO efforts by combining quantitative assessments of chemical saturation and SAR progression. Chemical saturation analysis utilizes AS-dependent VA populations to chart chemical space around an AS. Such VAs might thus be assigned a dual purpose as diagnostic chemical entities and as potential candidates for AS expansion. This dual role provides the opportunity to generate a unique computational approach that combines LO diagnostics with compound design and candidate prediction. The corresponding workflow includes the analysis of optimization characteristics of ASs, identification of series with further development potential and use of predictive models to screen AS-specific VA populations for preferred candidate compounds. Extending COMO for combined diagnostic AS analysis and prospective series expansion was the major goal of our study.

**Diagnostic scoring**

For our analysis, new ASs were assembled that contained at least 100 compounds. As reported in Table 1, the majority of these ASs consisted of 100–200 analogs. The three largest ASs comprised 237, 259 and 264 compounds, respectively. Hence, newly identified ASs were of considerable size. Initially, it was investigated if these ASs displayed different characteristics suitable for our analysis. Therefore, COMO scores were calculated. Figure 2 compares S and P scores for the ASs. COMO scoring clearly distinguished between ASs, revealing different degrees of chemical saturation and SAR progression that did not correlate with AS size. None of the ASs displayed a combination of high chemical saturation and low SAR progression, which would represent a termination criterion [9]. Hence, all ASs were still expandable through the generation of new analogs and were thus suitable for our analysis.

**Regression QSAR models**

A prerequisite for meaningful screening of AS-specific VA populations is the derivation of accurate QSAR models for given ASs. Accurate models make it possible to carry out meaningful predictions for VAs and prioritize candidate compounds for further exploration. Therefore, we generated standard RR and SVR models for all 24 ASs and evaluated their predictive performance. In our study, no decision tree methods were considered, given that SVR is a widely applied standard in the QSAR field. The results are shown in Figure 3 and reveal that model performance was highly heterogeneous, depending on the AS. For the majority of ASs, no predictive regression models were obtained. Models with \( R^2 \) values exceeding 0.6 were only observed in a few cases. Overall, there was a slight increase in prediction accuracy for the more complex SVR over the simple RR models. Limited prediction accuracy of regression models for ASs is frequently observed if models are not iteratively fine-tuned for individual series. However, for systematic AS expansion, robust predictive models with meaningful accuracy are required. Clearly,
Figure 2. Compound optimization monitor diagnostic scores. The scatter plot compares COMO chemical saturation (S) and SAR progression (P) scores for ASs from ChEMBL (version 25). Each dot represents a series and is scaled in size according to the number of analogs. Different combinations of S and P scores distinguish ASs at different LO stages. AS: Analog series; COMO: Compound optimization monitor; LO: Lead optimization; SAR: Structure–activity relationship.

Figure 3. Performance of regression models. On the vertical axis, mean coefficients of determination (R²) for regression models are reported. The horizontal axis lists ASs using their IDs according to Table 1. For each AS, two R² values are given for RR (blue) and SVR (orange) models. AS: Analog series; ID: Identification; RR: Ridge regression; SVR: Support vector regression.

on the basis of our test calculations, limited accuracy of standard regression models prohibited their general use for our purposes. Therefore, alternative predictive approaches were explored.

FW predictions
We reasoned that FW-type predictions following the formalism illustrated in Figure 1 might provide an alternative. This assumption was based on the local nature of FW predictions involving separate NBHs. Locally confined
predictions would alleviate the need for building regression models of entire ASs that might be affected by the presence of SAR discontinuity. Therefore, we systematically searched the 24 ASs for FW EAs enabling local predictions. Varying numbers of up to 162 FW EAs were detected in 19 ASs (Table 1). For 18 of these ASs (one with only four FW EAs was excluded), systematic FW predictions were carried out. For this purpose, each FW EA was virtualized once in each NBH it occurred. Figure 4A shows R² values for FW and global SVR predictions. Compared with regression modeling, the results were much more promising. In this case, 11 of the 18 qualifying ASs yielded FW predictions with R² values in the range of >0.5 to 1.0 (>0.6 for seven ASs). Predictions on the seven remaining ASs with typically only small numbers of FW EAs and NBHs essentially failed. As a control, potency predictions for FW EAs using SVR models were extracted from all external validation sets and separately

Figure 4. Comparison of Free-Wilson and support vector regression predictions. Scatter plots compare R² values of FW and SVR predictions for individual ASs. Each dot represents an AS that is scaled in size according to the number of FW EAs. The diagonal corresponds to perfect correlation between calculated coefficients. (A) FW versus global SVR predictions (according to Figure 3). (B) FW versus SVR predictions on FW EA subsets. AS: Analog series; EA: Existing analog; FW: Free-Wilson; SVR: Support vector regression.
Figure 5. Experimental and predicted potency values. (A) Box plots compare experimental and predicted potency value distributions. Each triplet represents one of 11 ASs yielding predictive FW and SVR models. The y-axis reports log. potency values and the x-axis the number of FW EAs. Numbers in parentheses are AS IDs. Distributions of experimental potency values (green), mean FW predictions (red) and mean SVR predictions (orange) are reported for FW EAs. (B) Individual predictions are shown for four exemplary FW EAs (with ChEMBL IDs) from the same AS with activity against purinergic receptor P2Y12 (AS 6; ChEMBL target ID 2001). In the table inserts, the first row contains the experimental potency values of each analog and the second row the mean FW-predicted potency values (with the corresponding number of FW NBHs in parentheses). The third row contains the mean SVR-predicted potency values (with the corresponding number of individual prediction trials in parentheses).

AS: Analog series; EA: Existing analog; FW: Free-Wilson; ID: Identification; NBH: Neighborhood; SVR: Support vector regression.

evaluated, as shown in Figure 4B. Surprisingly, for the 11 ASs with promising FW predictions, the potency of FW EAs was also predicted with higher accuracy using SVR models than other external validation instances. These predictions were comparable with FW analysis. Spearman correlation coefficients of potency values predicted by FW analysis and SVR models were high, ranging from 0.82 to 0.98. These improvements might be attributable to nearest neighbor effects among FW EAs.

For the 11 AS, we also compared the experimental potency distribution of FW EAs with predicted distributions, as shown in Figure 5A. In the majority of cases, similar distributions and median values were observed. Notable differences between experimental and predicted potency distributions were only detected for three ASs. Figure 5B shows four exemplary FW EAs for which FW potency predictions over 36 to 51 NBHs and SVR predictions over six to 11 trials were available. These examples represented different levels of prediction accuracy. For one compound (top left), the experimental potency was exactly predicted by both FW and SVR. For another (bottom right), both methods under-predicted the experimental value by 0.8 log units. For the remaining two examples, one of the two approaches was slightly more accurate than the other. However, in all cases, the experimental potency was predicted well within an order of magnitude using both FW and SVR. Such predictions are meaningful taking experimental accuracy limits into consideration.

Thus, taken together, the results indicated that predictions of FW EAs focusing on local NBHs were much more promising than results obtained with regression models for entire ASs. Therefore, preference was assigned to FW analysis for compound potency predictions. Ultimately, such predictions must be carried out on VAs in order to prioritize candidate compounds. Therefore, in the next step, COMO VA populations were further analyzed.

FW VAs
The diagnostic VA populations generated for the 24 ASs and used to calculate the COMO scores in Figure 2 were screened for VAs that complemented FW NBHs. These FW VAs qualified for FW predictions. However,
diagnostic VA populations only contained few if any FW VAs. This was a likely consequence of using a large pool of diverse substituents for VA enumeration (see Methods). Therefore, to make FW analysis a practical option for potency prediction, we complemented diagnostic VA populations with new FW VAs. These VAs were specifically designed to complete FW NBHs in given ASs. Therefore, we implemented a new algorithm to generate FW VAs on the basis of EAs, as detailed in the Methods section. Application of this algorithm yielded between 100 and 5798 FW VAs for all but one of the 24 ASs, as reported in Table 1. Each AS contained multiple substitution sites with 45–265 available substituents that were recombined for FW analog generation. Accordingly, in some cases, large VA ensembles with several thousand compounds were obtained. Hence, through complementary analog design, COMO’s VA populations were significantly enriched with FW VAs as potential candidate compounds for AS expansion.

Pilot predictions
To further evaluate the general suitability of FW VAs for AS expansion, potency predictions were carried out using both FW analysis and SVR models for the 11 ASs for which predictions of FW EAs succeeded. The underlying idea was that FW VA ensembles should contain FW VAs having higher predicted potency than EAs. Such FW VAs would represent preferred candidates for experimental evaluation.

**Figure 5. Experimental and predicted potency values (cont.).** (A) Box plots compare experimental and predicted potency value distributions. Each triplet represents one of 11 ASs yielding predictive FW and SVR models. The y-axis reports log. potency values and the x-axis the number of FW EAs. Numbers in parentheses are AS IDs. Distributions of experimental potency values (green), mean FW predictions (red) and mean SVR predictions (orange) are reported for FW EAs. (B) Individual predictions are shown for four exemplary FW EAs (with ChEMBL IDs) from the same AS with activity against purinergic receptor P2Y12 (AS 6; ChEMBL target ID 2001). In the table inserts, the first row contains the experimental potency values of each analog and the second row the mean FW-predicted potency values (with the corresponding number of FW NBHs in parentheses). The third row contains the mean SVR-predicted potency values (with the corresponding number of individual prediction trials in parentheses).

**AS:** Analog series; **EA:** Existing analog; **FW:** Free-Wilson; **ID:** Identification; **NBH:** Neighborhood; **SVR:** Support vector regression.
Figure 6. **Potency predictions for Free-Wilson virtual analogs.** (A) Box plots compare experimental potency value distributions of 11 ASs according to Figure 5A with potency predictions of corresponding FW VA populations. The y-axis reports logarithmic potency values and the x-axis the number of FW VAs per series. Numbers in parentheses are AS IDs. The experimental potency distribution of all EAs per series is displayed in light green, the FW-predicted VA potency distribution in red and the corresponding SVR-predicted distribution in orange. (B) Exemplary VAs (middle and right) are shown that were predicted to have higher potency than the most potent EA (left) of an AS active against the P2X purinoceptor 3 (AS 19; ChEMBL target ID 2998). In beeswarm plots below (color-coded according to the box plots), the exemplary compounds are indicated using arrows.

AS: Analog series; EA: Existing analog; FW: Free-Wilson; ID: Identification; NBH: Neighborhood; SVR: Support vector regression; VA: Virtual analog.

Figure 6A compares the experimental potency value distribution of the 11 ASs with potency value distributions predicted for FW VAs using FW analysis and SVR. Predicted potency value distributions were generally lower than experimental distributions. In all but one case, the predicted median potency was lower than the experimental median. This was principally meaningful because FW VA ensembles should also contain a variety of inactive analogs. Consistent with this expectation, FW analysis predicted a number of FW VAs from different ASs to be inactive. However, the potency value distributions predicted by FW analysis typically covered a wide range. For each AS, at least a few FW VAs were consistently predicted by FW analysis to be more potent than the most potent EAs. This was an encouraging observation, providing a basis for FW VA prioritization in practical applications.

SVR-predicted distributions were generally narrower than FW distributions. Different from FW analysis, SVR is intrinsically limited to interpolative potency predictions falling within the range of training data. Thus, the potency of a few FW VAs that were predicted by SVR to be more potent than experimental analogs fell within the range of the permitted absolute prediction error of the models. Consequently, for only three ASs, potencies beyond the highest experimental value were observed.

Figure 6B shows the most potent compound from an AS representing a FW EA whose logarithmic potency value (pIC$_{50}$ = 8.4) was well predicted using both FW analysis (pIC$_{50}$ = 8.0) and SVR (pIC$_{50}$ = 7.8). In addition, two FW VAs of this compound are depicted that were predicted to be most potent by FW analysis (pIC$_{50}$ = 9.0) and SVR (pIC$_{50}$ = 8.5), respectively. These two FW VAs were only distinguished by a cyclopentyl ether to methyl cyclopropanyle ether substitution. Both analogs were predicted by FW analysis to be more potent than the FW EA.

Taken together, the result of pilot predictions on newly generated FW VAs indicated that candidates for AS expansion could be consistently selected on the basis of FW analysis. While the activity state of preferred FW
**Figure 6. Potency predictions for Free-Wilson virtual analogs (cont.).** (A) Box plots compare experimental potency value distributions of 11 ASs according to Figure 5A with potency predictions of corresponding FW VA populations. The y-axis reports logarithmic potency values and the x-axis the number of FW VAs per series. Numbers in parentheses are AS IDs. The experimental potency distribution of all EAs per series is displayed in light green, the FW-predicted VA potency distribution in red and the corresponding SVR-predicted distribution in orange. (B) Exemplary VAs (middle and right) are shown that were predicted to have higher potency than the most potent EA (left) of an AS active against the P2X purinoceptor 3 (AS 19; ChEMBL target ID 2998). In beeswarm plots below (color-coded according to the box plots), the exemplary compounds are indicated using arrows.

AS: Analog series; EA: Existing analog; FW: Free-Wilson; ID: Identification; NBH: Neighborhood; SVR: Support vector regression; VA: Virtual analog.

VAs remains unknown prior to experimental evaluation, the calculations revealed potential candidates. Their prioritization was further supported by meaningful predictions of FW EA potency.

It is important to note that the generation of FW VAs does not yield novel substituents because the substituents are sampled from existing compounds. Instead, novel core-substituent combinations are obtained. By design, FW VAs are enumerated to enable frequent FW predictions.
Figure 7. Free-Wilson neighborhood saturation scores. N scores are shown for ASs yielding predictive models as a function of increasing FW EA fraction, defined as the proportion of FW EAs among all EAs. Dots represent ASs that are scaled in size by the number of analogs per series and color-coded according to the number of algorithmically generated FW VAs per series.

AS: Analog series; EA: Existing analog; FW: Free-Wilson; VA: Virtual analog.

**FW centric saturation diagnostic**

The newly introduced FW VA algorithm made it also possible to further extend COMO’s diagnostic scoring scheme by focusing on the saturation of FW NBHs. This additional scoring opportunity provided a close link between NBH characteristics and prospective design.

The number of FW EAs and FW VAs per AS depends on pairwise relationships between EAs captured by MMPs. Increasing numbers of FW NBHs per FW EA support reliable potency predictions. To quantitatively assess these distributions, we introduce an additional FW NBH saturation score N, which quantifies the saturation of an AS with FW NBHs:

\[
N = 1 - \frac{n_{FW\text{ EA}}}{n_{FW\text{ NBH}}}
\]  
(Eq. 10)

Accordingly, increasing N values result from increasing numbers of FW NBHs per FW EA. In addition to AS size, this also increases the statistical likelihood to identify FW VAs. Large N scores indicate the presence of NBH behavior among EAs and the potential to further expand ASs with prioritized and correctly predicted FW VAs. Figure 7 reveals that N scores of ASs typically increased with increasing proportions of FW EAs among EAs. Accordingly, this measure of NBH content was a meaningful addition to COMO’s scoring repertoire. Furthermore, nearly all ASs for which well-performing predictive models were obtained produced high N scores indicating reliable potency predictions. Such predictions can only be obtained in the presence of SAR continuity, which also provides a basis for optimization of other properties during later stages of LO. This is the case because in the presence of SAR continuity, substitutions will lead to moderate changes in potency, making it possible to balance multiple properties.

**Conclusion**

In this work, the diagnostic COMO approach was used as a platform to develop the first computational methodology for combining the assessment of progress in LO and expansion of ASs. To these ends, complementary strategies for analog design and potency prediction were explored. FW analysis was found to be a preferred approach for potency prediction across different ASs. Given its local nature, interpretability, and low computational complexity,
FW analysis was attractive from several points of view. However, to enable extensive FW predictions for candidate prioritization, VA populations needed to be enriched with FW VAs. This was accomplished by developing a new dual-purpose algorithm to search for FW EAs and generate AS-specific FW VAs as source for candidate compounds. For FW VAs, FW analysis yielded predictions covering a wide potency range, including FW VAs predicted to be more potent than EAs. Algorithmic generation of FW NBH also led to the introduction of a new NBH-based saturation score. This score is applicable to estimate the likelihood of obtaining FW VAs and FW predictions over multiple NBHs. Taken together, our results indicate that computational LO diagnostics, analog design and candidate prioritization can be effectively integrated.

Future perspective
Combining LO diagnostics with analog design has significant potential for practical applications. ASs can be profiled on a large scale and series with strong potential for further development can be selected. In addition, parallel series can be monitored for chemical saturation and SAR progression characteristics during late stages of LO and close-in VAs can be generated. COMO offers new opportunities to closely link AS evaluation and expansion. Assessment of chemical saturation and SAR progression has been extended by FW NBH centric scoring to identify ASs that have potential for further expansion through FW analysis. For evolving series, FW EAs can be identified using our new algorithm and then systematically predicted to assess potency prediction accuracy. For qualifying ASs, the dual-purpose algorithm can be applied to generate FW VAs. The resulting FW VA ensembles provide the basis for a second round of potency predictions to prioritize candidates for synthesis. As we have shown, SVR models also yield consistently more accurate potency predictions for FW EAs than other EAs. Hence, SVR also merits consideration for prediction of FW VAs. FW analysis was shown to produce predictions covering wide potency ranges, typically including candidates with higher predicted potency than EAs. By contrast, potency ranges predicted using SVR were smaller. However, for potency predictions of FW VAs, both FW analysis and regression modeling might best be applied in parallel to determine if most potent FW VA candidates from a given ensemble are consistently predicted. The combined diagnostic and compound design approach can be practically applied to ASs of any source. Future refinements and extensions of the methodology will have several focal points. A major limiting factor for analog prioritization is the dependence of potency prediction accuracy on the nature of ASs, which represents a general problem in the QSAR field. Hence, for an attractive AS, it might not be possible to generate reasonable predictive models to guide analog design. Accordingly, it will be beneficial to further explore and characterize SAR features that limit prediction accuracy. Any potential progress in the area is highly desirable. Furthermore, specifically for our methodology, an area of high priority for future development will be the extension of diagnostics and predictions to other LO-relevant molecular properties. Such development efforts are currently hindered by limited availability of high-quality data beyond potency measurements in the public domain. It is hoped, however, that such data will become increasingly available in the near future, in particular, through academic drug discovery efforts and/or increasing collaborations between the pharmaceutical industry and academia.

Author contributions
J Bajorath conceived the study; J Bajorath and D Yonchev planned the analysis; D Yonchev performed the analysis; D Yonchev and J Bajorath analyzed the results and prepared the manuscript.

Acknowledgments
D Yonchev is supported by the Jürgen-Manchot-Stiftung, Düsseldorf.

Financial & competing interests disclosure
The authors have no 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. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

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

Open access
The work is licensed under the Creative Commons Attribution 4.0 License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
Executive summary

- Lead optimization (LO) plays a central role in medicinal chemistry but is vulnerable.
- Computational approaches providing decision support are rare.
- The compound optimization monitor (COMO) method quantitatively assesses optimization progress.
- LO diagnostics and compound design have not yet been combined.

Methods

- Principles of COMO diagnostics are summarized.
- Key scores are explained.
- The identification of analog series is described.
- Different methods for compound potency prediction are compared.
- A new algorithm for FW (Free-Wilson)-oriented analog design is introduced.
- Model building and test calculations are detailed.

Results & discussion

- Study rationale and goals are emphasized.
- Alternative predictive models are evaluated and compared.
- Complementary virtual analog (VA) ensembles for FW analysis are generated.
- A new FW neighborhood centric COMO score is introduced.

Future perspective

- A workflow for practical applications of the extended COMO approach is provided.
- Areas for future development are highlighted.
- Extending the approach to multiple LO-relevant properties is a priority.

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. Cherkasov A, Muratov EN, Fourches D et al. QSAR modeling: where have you been? Where are you going to? J. Med. Chem. 57(12), 4977–5010 (2014).
   ● Extensive review of quantitative structure–activity relationship approaches.
4. Segall M. Advances in multi-parameter optimization methods for de novo drug design. Expert Opin. Drug Discov. 9(7), 803–817 (2014).
   ● Discussion of multiparameter optimization techniques.
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 Leon A, Dimova D, Bajorath J. Monitoring the progression of structure–activity relationship information during lead optimization. J. Med. Chem. 59(9), 4235–4244 (2015).
   ● Data structures for computational evaluation of 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. Vogt M, Yonchev D, Bajorath J. Computational method to evaluate progress in lead optimization. J. Med. Chem. 61(23), 10895–10900 (2018).
   ● Combination of chemical saturation and structure–activity relationship progression analysis.
9. Yonchev D, Vogt M, Bajorath J. Compound optimization monitor (COMO) method for computational evaluation of progress in medicinal chemistry projects. Future Drug Discov. 1(2), FDD15 (2019).
   ●● Introduction of compound optimization monitor, its diagnostic components, and scores.
10. 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).
11. 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).
12. Grifffen E, Leach AG, Robb GR, Warner DJ. Matched molecular pairs as a medicinal chemistry tool. J. Med. Chem. 54(22), 7739–7750 (2011).
13. 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).
   ● Versatile algorithm for the systematic generation of matched molecular pairs.
14. 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).

15. de la Vega de León A, Bajorath J. Matched molecular pairs derived by retrosynthetic fragmentation. *Med. Chem. Commun.* 5(1), 64–67 (2014).

16. Vapnik VN. *The Nature of Statistical Learning Theory*. 2nd Edition. Springer, NY, USA (2000).

17. Marquardt DW, Snee RD. Ridge regression in practice. *Am. Stat.* 29(1), 3–20 (1975).

18. Dorugade AV, Kashid DN. Alternative method for choosing ridge parameter for regression. *Appl. Math. Sci.* 4(9), 447–456 (2010).

19. Joachims T. Making Large-scale SVM Learning Practical. In: *Advances in Kernel Methods: Support Vector Learning*. Schölkopf B, Burges CJC, Smola AJ (Eds). 169–184 MIT Press, MA, USA (1999).

20. Drucker H, Burges C. Support vector regression machines. *Adv. Neural Inform. Process. Systems* 9(1), 155–161 (1997).

21. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Stat. Comput.* 14(3), 199–222 (2004).

22. Free SM, Wilson JW. A mathematical contribution to structure–activity studies. *J. Med. Chem.* 7(4), 395–399 (1964).

23. Kubinyi H. Free Wilson analysis. Theory, applications and its relationships to Hansch analysis. *Quant. Struct.-Act. Relat.* 7(3), 121–133 (1988).

24. Hagberg A, Swart P, Chult DS. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Laboratory, NM, USA. https://www.osti.gov/biblio/960616

25. OEChem TK. OpenEye Scientific Software, Inc, NM, USA (2012). https://www.eyesopen.com/oecchem-tk

26. Baumann D, Baumann K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J. Cheminform.* 6(1), e47 (2014).

27. Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12(Oct), 2825–2830 (2011).

28. Rogers D, Hahn M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50(5), 742–754 (2010).

29. RDKit: open-source cheminformatics (2019). http://www.rdkit.org

30. Ralaivola L, Swamidass SJ, Saigo H, Baldi P. Graph kernels for chemical informatics. *Neural Netw.* 18(8), 1093–1110 (2005).