Development of Computational Methods for Rationalizing Chemical Lead Optimization and Compound Design

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Apotheker Dipl-Pharm.
Dimitar Yonchev
aus Samokov, Bulgarien

Bonn 2021

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath
2. Referent: Univ.-Prof. Dr. rer. nat. Michael Gütschow

Tag der Promotion: 09.09.2021
Erscheinungsjahr: 2021

Abstract

Chemical lead optimization (LO) plays an important role in pharmaceutical drug discovery. It represents a highly complex multi-objective process, in which the ultimate goal is to identify a suitable candidate molecule that satisfies a variety of often contradicting properties. This is facilitated by iterative synthesis and testing of structurally similar analog compounds, leading to the formation of analog series (ASs) with distinct characteristics contingent upon the project requirements. Decision-making during LO is largely driven by the subjective intuition, experience, and knowledge of medicinal chemists, which renders it challenging and prone to bias when assessing the overall progress of a campaign. While most of the *in silico* methods relevant to LO are mainly focused on characterizing single molecules, so far only a few approaches have been explicitly designed to evaluate its course at the level of entire series.

This thesis follows the gradual evolution of computational methodologies for data-driven LO rationalization based upon analyzing individual ASs. In addition, it explores different strategies for design and prioritization of new potential candidate compounds. Thereby, real-world LO campaigns are modeled by systematically extracting series from publicly available compound bioactivity data originating from medicinal chemistry literature and screening campaigns. The incremental development of novel cheminformatic approaches for quantifying the chemical saturation and structure-activity relationship progression of ASs culminates in the introduction of the Compound Optimization Monitor (COMO) as a diagnostic tool for holistic evaluation of different optimization aspects. Thus, by subjecting ASs to comprehensive scoring, their development stage can be estimated according to their distinct LO profile, providing a means for objective comparison and (de)prioritization as well as rationale for potential (dis)continuation criteria. Moreover, an elaborate system for virtual analog design and candidate selection is integrated into the method, in order to further enhance its potential for practical application and support medicinal chemists in deciding upon what compounds to synthesize next in a given series. Therefore, different *de novo* design strategies for generation of synthetically accessible focused compound libraries ranging from rule-based methods to generative deep learning models are investigated and combined with different approaches for compound activity prediction.

In memory of Dr. Andrey Karpov and Ivo "Fena" Vladimirov

Acknowledgements

First of all, I would like to express my gratitude to my supervisor Prof. Dr. Jürgen Bajorath for introducing me to the field of cheminformatics as well as for the scientific inspiration and devoted mentoring. His continuous scientific guidance, personal support, and encouragement have motivated me throughout my doctoral studies. I also thank Prof. Dr. Michael Gütschow for reviewing my thesis as a co-referent and the other members of my PhD committee. I am grateful to the Jürgen Manchot Stiftung, Düsseldorf for their financial support and for providing me with scientific freedom during my PhD studies.

I further want to thank all my friends and colleagues (current and former) from the Life Science Informatics Group for always being there for me and for all the unforgettable times we had together. Special thanks to Dr. Dilyana Dimova for guiding me through my first steps in cheminformatics, to Dr. Dagmar Stumpfe for her continuous support with preparing scientific presentations and manuscripts, and to Dr. Martin Vogt for the valuable scientific advice and fruitful discussions during our work together. Furthermore, I thank Dr. Tomoyuki Miyao and Dr. Ryo Kunimoto for being excellent team partners, Thomas Blaschke for his helpful technical advice and optimism, and Christian Feldmann for all the scientific inspiration and ideas that have arisen from our countless brainstorming sessions. I also want to thank Oliver Laufkötter for being a great climbing partner, Huabin Hu for making weekends at b-it less lonely, and Dr. Erik Gilberg for his moral encouragement.

Moreover, I owe my deepest gratitude to my wife Yana, my parents Daniela and Georgi, my sister Polina, my grandparents, and all friends for their unconditional support, encouragement, and inspiration in both good and bad times. To Yana, thank you for your wisdom, soothing words, and patience with all the late-night shifts and endless scientific discussions. Last but not least, I thank Michael Fischer for being a great team partner during my undergraduate pharmacy studies and Elislav Atanasov for the motivation while writing this thesis.

Contents

List of abbreviations

1D, 2D, 3D	One-, two-, three-dimensional
ADME	Absorption, distribution, metabolism, excretion
AE	Autoencoder
AS	Analog series
ASB	Analog series-based
CCR	Compound-core relationship
COMO	Compound Optimization Monitor
DL	Deep learning
DNN	Deep neural network
EA	Existing analog
ECFP	Extended-connectivity fingerprint
FP	Fingerprint
FW	Free-Wilson
GAN	Generative adversarial networks
GRU	Gated recurrent unit
InChI	IUPAC international chemical identifier
IND	Investigational new drug
LO	Lead optimization
LogD	Logarithmic distribution coefficient
LogP	Logarithmic octanol-water partition coefficient
LR	Linear regression
LSTM	Long short-term memory
MACCS	Molecular access system
ML	Machine learning
MMP	Matched molecular pair

MPO	Multi-parameter optimization
NBH	Neighborhood
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
pK_a	Acid dissociation constant (-log10)
pK_i	Constant of inhibition (-log10)
RF	Random forest
RL	Reinforcement learning
RNN	Recurrent neural network
SA	Synthetic accessibility
SAR	Structure-activity relationship
SARM	Structure-activity relationship matrix
SMILES	Simplified molecular-input line-entry system
SVM	Support vector machine
SVR	Support vector regression
Tc	Tanimoto coefficient
TL	Transfer learning
TPSA	Total polar surface area
VA	Virtual analog

Chapter 1

Introduction

1.1 Drug discovery

Pharmaceutical research and development is an integral part of modern medicine and healthcare.[1] Since the late 1800s drug development has become an increasingly complex task giving rise to a large high-tech multi-billion dollar industry.[2–5] Today, its importance is particularly evident in the midst of 2020's global COVID-19 pandemic.[6,7] In an unprecedented effort to develop vaccines and antiviral drugs within a shortest possible period of time, pharmaceutical research has largely benefitted from the application of advanced computational methods and artificial intelligence which significantly accelerate progression and reduce the cost of the undertaken campaigns.[8–10]

1.1.1 Development stages

The process of small molecule drug discovery and development, albeit very distinct in each case, generally consists of several stages as shown in **Figure 1**.[11] The first step is target identification which involves basic research in pathophysiology and chemical biology to explore and determine the pharmacological target(s) and the molecular mechanism of action associated with a disease phenotype.[12–14] During the following stage of hit identification, large chemical libraries of structurally diverse compounds[15] are screened for activity against the validated target in a series of biological assays.[16,17] In the subsequent hit-to-lead phase, compounds found to be active (hits) are examined more closely and the most promising (lead) molecules, which typically represent several different chemotypes, are prioritized.[18,19] In the next stage of lead optimization (LO), critical drug-related properties, which determine the pharmacodynamic, pharmacokinetic, and toxicological profile of a compound, are improved through directed chemical modifications in lead structures accompanied by extensive *in vitro* and *in vivo* testing.[20] Finally, if successful, preclinical drug development converges at the point of formal registration of an investigational new drug (IND) for subsequent clinical trials in humans.[21] If a drug is approved by a regulatory institution and granted market access, its post-approval usage and conferred therapeutic benefits must be continuously monitored according to pharmacovigilance guidelines to protect patients from risks and adverse side effects.[22] The estimated success rate of drug development projects is around 10%[23] and, given their considerable length and high financial cost, significant efforts are made to reduce attrition in each of the aforementioned stages.[24–27]

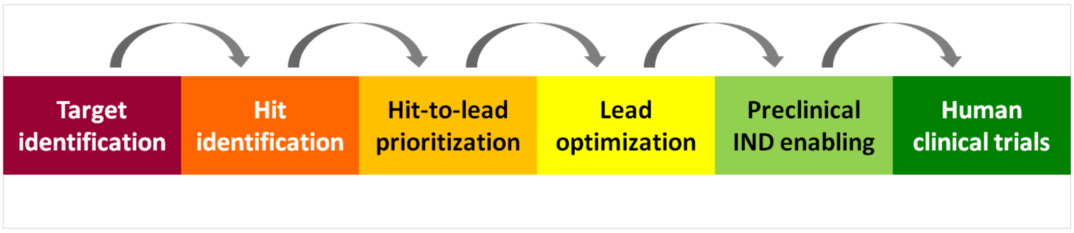


Figure 1: Drug discovery and development. An overview of the main stages of drug discovery and development is provided.

1.1.2 Computational methods

In the last several decades, a variety of *in silico* methods have been developed to augment and accelerate experimental medicinal chemistry campaigns for small molecule drug discovery.[28–32] For example, structure-based computational chemistry approaches such as molecular docking[33] and molecular dynamics[34] are related to the fields of structural biology and biophysics.[29] They require prior knowledge of the structure of both biological target and ligand(s) and rely on theoretical and quantum chemistry to derive mathematical models for explaining and predicting molecular interactions.[30]

In contrast, cheminformatics is primarily focused on *"the design, creation, organization, retrieval, analysis, dissemination, visualization, and use of chemical information"*.[35,36] As a data-centric discipline similar to bioinformatics,[37] it has emerged as an efficient tool to manage the exponentially increasing amounts of generated chemical information in the era of Big Data.[38–41] Cheminformatics gives rise to ligand-based approaches, where explicit knowledge of the pharmacological target structure is not imperative, since compound activity is implicated by chemical similarity to already known ligands.[30,39] The theoretical foundations for this are attributed to the similarity-property principle, which is a leading paradigm in drug discovery and states that structurally similar compounds demonstrate similar chemical and biological properties.[42] This serves as a basis for quantitative structure-property relationship (QSPR) and quantitative structure-activity relationship (QSAR) analysis methods, which aim at predicting changes in physicochemical properties and biological activity as functions of structural modifications in compounds.[43–46] The similarity-property principle is further utilized by ligand-based virtual screening, where potential new hit compounds are identified from large pools of existing or virtual molecules[47] based on their similarity to known reference ligands or on the presence of predefined structural features, such as pharmacophores which determine biological activity.[48,49] Furthermore, machine learning (ML)[50,51] and, more recently, deep learning (DL)[52–55] methods from computer science have been widely adopted, often in combination with Big Data, to address the aforementioned classical cheminformatic problems[46] as well as tasks such as chemogenomics,[56] *de novo* molecular design,[57,58] chemical reaction prediction,[59–61] and exploring alternative molecular representations.[62]

1.1.3 Medicinal chemistry databases

Recent advances in cheminformatics can be attributed not only to the growing computational power and algorithmic development but also to the increasing availability of chemical and biological data.[63] Notable examples for online databases relevant for drug discovery include DrugBank,[64] the Protein Data Bank,[65] UniProt,[66] SureChEMBL,[67] and ZINC.[68] Currently, the largest publicly available repository for compound bioactivity information is PubChem BioAssay.[69] It stores heterogeneous small molecule screening data in form of biological assays deposited by various research institutions worldwide. One of the main advantages of PubChem is that by extracting data from many different assays, information about compound test frequency can be obtained. Conversely, the ChEMBL database comprises only compound bioactivity information collected from the medicinal chemistry scientific literature.[70] Here, automatic data extraction is coupled with manual curation, which leads to improved data quality and homogeneity. Furthermore, ChEMBL provides a comprehensive uniform database vocabulary, which allows the definition of different data confidence criteria and enables efficient data mining.

1.2 Chemical lead optimization

Once a promising lead compound with desirable activity for a given target or disease phenotype has been identified among screening hits, it needs to be transformed into an efficacious and safe clinical IND candidate.[20] First and foremost, a suitable candidate must be highly potent i.e. display a sustainable level of activity at the pharmacological target(s), typically at a nanomolar concentration.[71,72] Its selectivity profile with respect to physiologically similar targets must be precisely determined, so that therapeutic efficacy is achieved only through explainable target modulation.[73,74] Furthermore, potential genotoxicity must be precluded and off-target activity against proteins, that are known to elicit adverse side effects (antitargets), must remain as low as possible.[75,76] Besides these pharmacodynamic criteria, pharmacokinetic properties, such as absorption, distribution, metabolism, and excretion (ADME), which govern the compound's behaviour in complex biological systems, must be also calibrated.[77] Among these, stable metabolism and predictable interactions with transporter proteins are of paramount importance for minimizing potential toxicological risks.[78–80] Hence, computational methods for prediction of compound interactions with antitargets,[81–83] ADME profile,[84,85] and drug metabolism[86,87] have been proposed to address some of these challenges.

The simultaneous balancing of often contradicting compound properties represents a complex multi-parameter optimization (MPO) problem and is typically facilitated by iteratively synthesizing and testing different structural analogs of a lead compound in comprehensive screening cascades.[88–90] Thus, chemical series with large numbers of analogs,[91] frequently resulting from bioisosteric transformations,[92] may emerge from different lead compound chemotypes until, ultimately, an IND is proposed.[20] The recurrent need for accelerating and reducing the cost of long experimental campaigns has led to the introduction of *in silico* approaches for scoring and (de)prioritizing individual compounds.[93–95]

Critical physicochemical properties that need to be optimized include, among others, molecular weight, octanol-water partition coefficient (LogP), distribution coefficient at pH 7.4 (LogD), polar surface area, acid dissociation constant (pK_a), number of hydrogen bond donors and acceptors.[95] Different combinations of these have been reported to influence the ADME and toxicity of compounds and consequently, various "rules of thumb" and property filters have been proposed as guidelines for prioritizing potential "drug-like" candidates.[96–100] In addition, more sophisticated methods, such as desirability functions,[101,102] Pareto optimization,[103] and probabilistic scoring[104,105] have been employed for MPO. Furthermore, ligand efficiency indices have been developed as semi-empirical metrics combining free target-ligand binding energy with different physicochemical properties.[106–109] The underlying premise is that over the course of LO increase in compound potency often tends to be a concomitant effect of increasing molecular size and/or lipophilicity, which is in turn associated with higher risk of failure due to unfavorable ADME or toxicity.[110] Finally, some entirely empirical composite scores such as the quantitative estimate of drug-likeness[111] and the relative drug likelihood[112] have been derived based on the distributions of physicochemical properties of a set of approved oral drugs.

Nevertheless, due to the ambiguity in the definition of a drug and its often subjective association with regulatory approval status, the entire concept of drug-likeness and the strict application of rules, filters, and metrics arising from it have been disputed.[112–114] As has been repeatedly shown, the molecular properties of drugs do not represent some special confined property space, which is exclusively populated by successfully marketed compounds and must therefore be pursued on all accounts during LO, but instead have constantly evolved over time.[113,115–117] Hence, coveted compound characteristics are more contingent upon the specific requirements of an individual LO project than an abstract notion of drug-likeness.[113]

1.3 Virtual compound libraries

As outlined above, chemical exploration during LO is limited to a confined region of biologically relevant chemical space.[118] The latter, in turn, represents only a minute fraction of the chemical universe.[119] Theoretical estimates of the magnitude of the entire chemical space exceed 10^{60} organic small molecules, however the largest part of it is considered of no relevance for drug discovery.[120] Computational methods for efficient navigation and exploration of chemical space include, among others, the enumeration of all chemically feasible compounds with up to 17 heavy atoms,[121] identification of drug-like subspaces,[122] network-based analysis,[123] chemography,[124] and generative DL modeling.[125] Extending the chemical space of existing compound collections with both diverse and focused virtual molecular libraries is of prime interest for early drug discovery and is largely driven by the availability of chemical reactions.[126–129] Therefore, *in silico* assessment of the synthetic feasibility of generated virtual compounds needs to outperform or, at least, be in consensus with human knowledge and intuition.[130–132] In light of these considerations, different methods for computational estimation of synthetic tractability have been proposed based upon retrosynthetic rules or molecular complexity.[133,134] An efficient hybrid algorithm combining both approaches is the empirically derived synthetic accessibility score (SAscore), which takes into account historical synthetic knowledge instead of explicit retrosynthetic templates.[135] Hereby, the frequency of occurrence of individual chemical fragments is correlated to their quantitative contribution to synthetic accessibility while non-standard structural features, stereochemical complexity, and increasing molecule size are penalized.

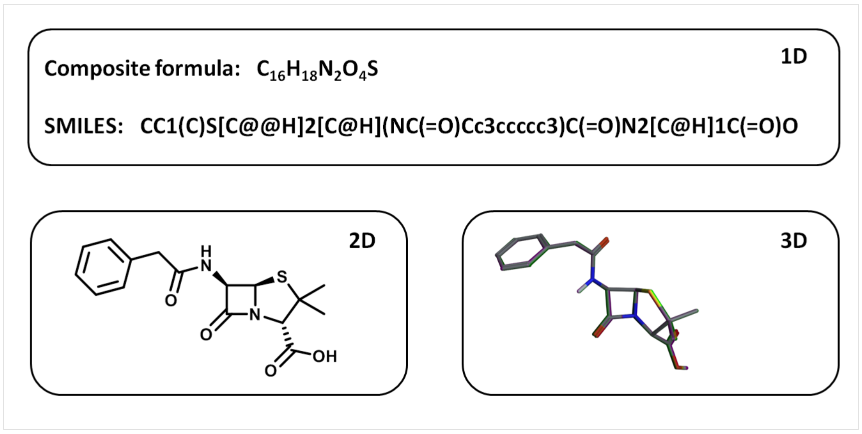
The term *de novo* design refers to "*the application of computational methods to automatically generate new compound structures in the search for an optimal compound*".[95] As such, *de novo* design aims to reduce experimental efforts by accomplishing the following three tasks: create virtual molecules, score and filter them, and subsequently optimize the sampling strategy according to the predefined objective based on previous knowledge.[136] Various ligand-based *de novo* design methods have been developed to address one or more of these tasks.[137] For example, in order to ensure synthetic accessibility, structure generation can be accomplished by conducting virtual reactions on synthetic building blocks.[138,139] This represents a successful deterministic approach for expanding biologically relevant chemical space,[140] however it might be prohibitive for generating molecular structures beyond the strict constraints of a relatively small set of robust organic reactions defined by medicinal chemists.

Alternative design methods can create new synthetically amenable compounds according to empirical rules derived from compound transformations typically undertaken in medicinal chemistry campaigns.[141,142] Such data-driven approaches allow the (de)prioritization of certain structural modifications based on their frequency of occurrence in historical data.[143] In addition, adaptive methods, such as evolutionary algorithms, can be utilized to select candidates via iterative cycles of optimization and sampling.[144–147] Successful applications of MPO-centric *de novo* approaches include the design of selective[148] and polypharmacological[149] compounds. Other methods, such as inverse QSAR/QSPR,[150,151] are still of limited use for drug discovery.[152]

1.4 Molecular representations

Encoding of structural features and/or properties of molecules is typically application-dependent and of variable degree of complexity.[153] In general, molecular representations can be categorized into one-, two-, and three-dimensional (1D, 2D, and 3D, respectively) as illustrated in **Figure 2**.[154] Linear notations, such as simple composition formulae, Simplified Molecular Input System (SMILES)[155] or IUPAC international chemical identifier (InChI),[156] are examples for 1D representations that emphasize either on human-readability or efficient computational processing. SMILES account for atom and bond types, aromaticity, branching, and stereochemistry by following specific syntax and canonicalization rules,[157,158] whereas InChI can, in addition, also encode different tautomers of a molecule.

The most popular and intuitive forms of molecular representation are 2D chemical graphs, which describe the specific molecular structure and topology.[159] Here, individual atoms and bonds are represented as nodes and edges, respectively, and can be optionally annotated with stereochemical information regarding their relative spatial arrangements. However, molecular con-formation, surface, and volume are best described using explicit 3D representations.[160] These consider specific steric and electronic properties including exact coordinates, bond orders, charges, and hybridization states.[160] Efficient computational processing of 2D and 3D molecular graphs is typically facilitated by storing them as connectivity tables in different file formats.[159,161]



Composite formula: $C_{16}H_{18}N_2O_4S$

SMILES: CC1(C)S[C@@H]2[C@H](NC(=O)Cc3ccccc3)C(=O)N2[C@H]1C(=O)O

Figure 2: Molecular representations. Shown are different molecular representations of the antibacterial drug benzylpenicillin in 1D, 2D, and 3D formats.

1.4.1 Descriptors

Molecules can be also described in terms of numerical values by applying different mathematical models on their 1D, 2D or 3D representations.[162–165] A large variety of molecular descriptors have been introduced in order to address different problems in cheminformatics.[166,167] For example, 1D descriptors calculated from linear molecular notations are only limited to bulk properties, such as simple atom counts or molecular weight. The more complex 2D descriptors are derived from molecular graphs and are used to quantify topological characteristics, such as total polar surface area (TPSA),[168] or approximate experimentally measured properties, such as LogP.[169] Lastly, 3D descriptors can be derived from specific molecular conformations or quantum mechanics calculations. Since individual descriptors are computed as single values condensing the information from entire molecular representations, they are typically aggregated into multi-dimensional molecular feature vectors, which constitute the dimensions of a chemical reference space.[166] The individual vectors are then used as coordinates to define the position of each molecule and quantitatively estimate its spatial proximity to other molecules.

1.4.2 Fingerprints

In contrast to that, molecular fingerprints (FPs) are a special category of descriptors, which directly represent the entire molecule as a numerical vector of homogeneous features.[170] Both length of the vector and type of the chemical feature can differ according to the FP definition.[171] Furthermore, FPs can be subdivided into binary and non-binary (counted).[172] In binary FPs, the presence or absence of a certain chemical feature determines whether its corresponding position in the bit vector is set to 1 or 0, respectively. In counted FPs, the frequency of occurrence of that feature is recorded instead. An example for a fixed-length binary FP is the Molecular ACCess System (MACCS),[173] which is a substructure FP that accounts for the presence or absence of predefined structural patterns (keys), each corresponding to a certain position in a 166-bit molecular vector. Other types of binary FPs, that go beyond the constraints of preset structural features, include combinatorial FPs such as the extended-connectivity fingerprints (ECFPs).[174] These circular FPs make use of a hashing function to encode specific atom environments as structural features based on traversing all possible molecular subgraphs in a layered fashion within a defined diameter.[175] Thus, the dimension of an ECFP is determined by the numerical range of the hashing function. It is typically too large to allow the representation of these FPs as explicit bit vectors. However, its dimensionality can be optionally reduced to a fixed-length vector via a special folding operation. **Figure 3** illustrates some of the principles for calculating descriptors and FPs.

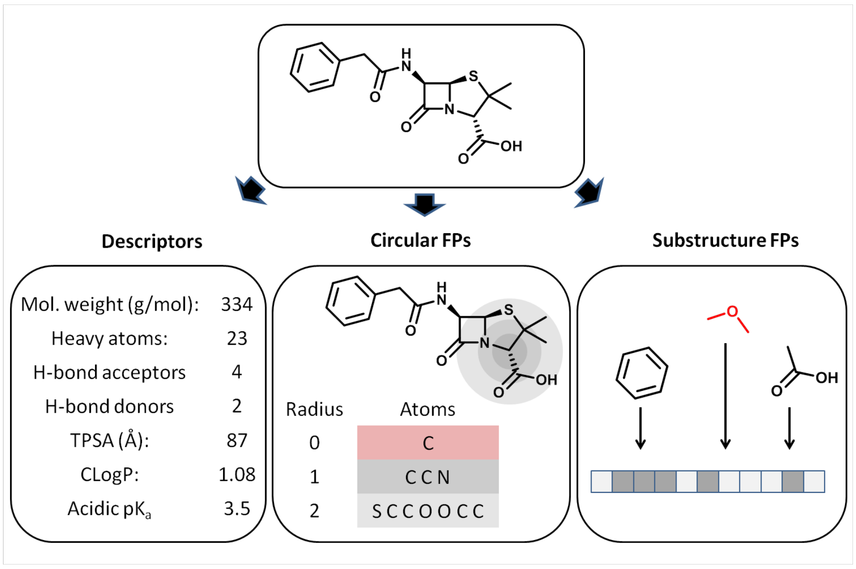


Figure 3: Molecular descriptors and fingerprints. An exemplary calculation of physicochemical descriptors (left), ECFPs with diameter of 4 atoms (middle), and MACCS FPs (right) for benzylpenicillin.

1.5 Structure-activity relationships

During LO, an analog series (AS) typically evolves around one or several congeneric core structures, also referred to as scaffolds, which are generally responsible for the underlying biological activity.[176–179] Medicinal chemists typically introduce structural modifications by exchanging functional groups (R-groups) at certain substitution sites in the AS core, in order to achieve a desired compound activity. The exploration and exploitation of such structure-activity relationships (SARs) is of crucial importance in medicinal chemistry because it can provide some orientation in the trajectory of a LO project. Moreover, the increasing amounts of available compound bioactivity data have enabled systematic large-scale SAR analysis beyond the scope of individual LO campaigns.[180]

In contrast to predictive methods such as QSAR, descriptive approaches aim to retrospectively deconvolute and visualize available SAR information.[181,182] Popular methods for SAR visualizaiton include SAR matrices and graphs, chemical space networks, and activity landscapes.[183] Furthermore, the SAR index can be employed to quantify distinct SAR characteristics for sets of compounds active against specific targets.[184] For instance, gradual changes in biological activity as a response to chemical modifications of varying magnitude can be rationalized as SAR continuity.[184] In a LO context, such predictable "linear" SAR behaviour might be desirable for fine-tuning physicochemical properties while trying to maintain relatively constant potency levels. Conversely, SAR discontinuity translates into large potency fluctuations resulting from minor alterations in chemical structure[184] with activity cliffs[185] representing the most notable examples. An activity cliff represents a pair of structurally similar compounds that exhibit a large potency difference.[186] Discontinuous SARs are naturally information-rich and indicative for progression but might be undesirable in some MPO scenarios due to higher probability of an unexpected "steep" decrease in compound potency. The concepts of SAR (dis)continuity are illustrated in detail in **Figure 4**.

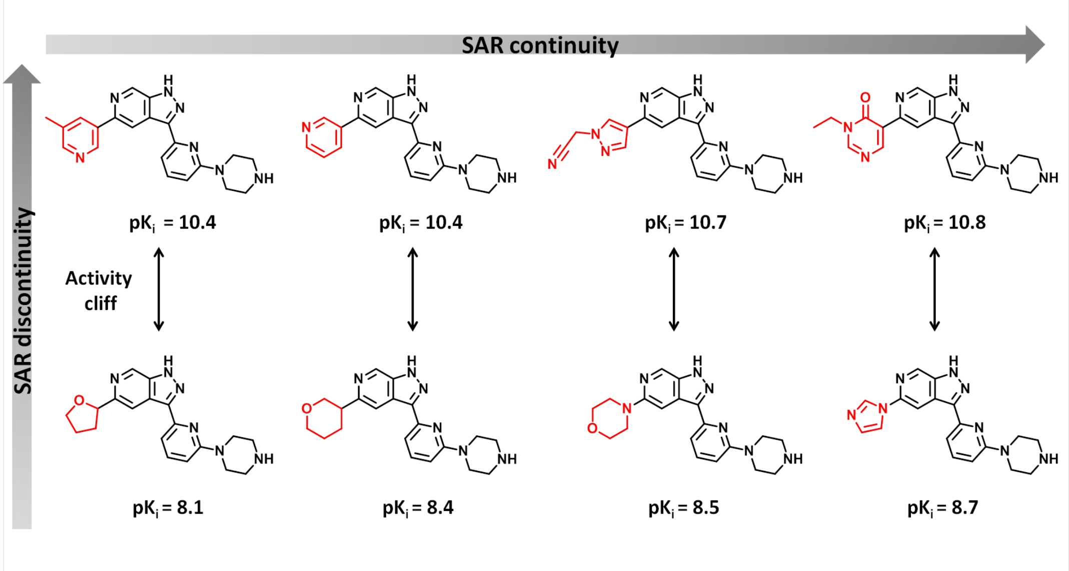


Figure 4: SAR (dis)continuity. Displayed are four highly (above) and four weakly (below) potent inhibitors of serine/threonine-protein kinase PIM1 and their corresponding potency values as negative logarithmic (log10) constants of inhibition (pK$_i$). Each structural modification in a compound (red) induces a different response in activity, corresponding to SAR continuity (left to right) or discontinuity (bottom to top).

1.5.1 Molecular similarity

The concept of molecular similarity is fundamental to cheminformatics due to its implications for the similarity-property principle. Therefore, proper assessment of molecular similarity has been a widely studied topic in cheminformatics as it often governs the success of virtual screening campaigns and QSAR/QSPR analysis.[187,188] A variety of measures have been developed to quantitatively estimate molecular similarity.[189,190] One approach is based upon the calculation of distances between compounds.[43] For example, the Euclidean distance d between two compounds P and Q in n-dimensional descriptor vector space can be calculated as follows:

$$d\,(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{1}$$

where p_i and q_i are the i-th descriptor of P and Q, respectively. Since distance in vector space is complementary to similarity, compounds with small distances to each other are regarded as similar and *vice versa.*[191] The most popular - albeit not universal - similarity measure in cheminformatics is the FP-based Tanimoto coefficient (Tc) also known as the Jaccard index.[192,193] It can be calculated for both binary or counted FPs and accounts for the percentage of shared structural features between two molecules. For two compounds with n-dimensional FPs A and B the similarity is computed as follows:

$$\mathrm{Tc}\,(A,B) = \frac{\sum\limits_{i=1}^{n}\min(a_i,b_i)}{\sum\limits_{i=1}^{n}\max(a_i,b_i)} \tag{2}$$

where a_i and b_i denote the i-th feature of FP A and B, respectively. Because the outcomes of similarity calculations are dependent on the choice of similarity measures and molecular descriptors/FPs, no universal threshold value for molecular (dis)similarity exists, although some empirically derived suggestions have been proposed.[190,194]

Alternatively, a more qualitative approach for the assessment of molecular similarity, that resonates with chemical intuition, is the identification of substructure relationships, such as matched molecular pairs (MMPs).[190,195] A MMP is defined as a pair of compounds that only differ at a single substitution site.[196] Systematic exploration of the effects of such structural transformations on compound properties is a widely used technique in medicinal chemistry and cheminformatics.[197–199] Algorithmic MMP generation relies on systematic compound fragmentation, which can be optionally designed to take synthetic feasibility into account by following certain retrosynthetic rules.[200,201] Given the subjective nature of the molecular similarity concept, qualitative and quantitative approaches do not necessarily overlap in their explanatory power, as depicted in **Figure 5**, and should be therefore viewed as complementary rather than mutually exclusive.[190]

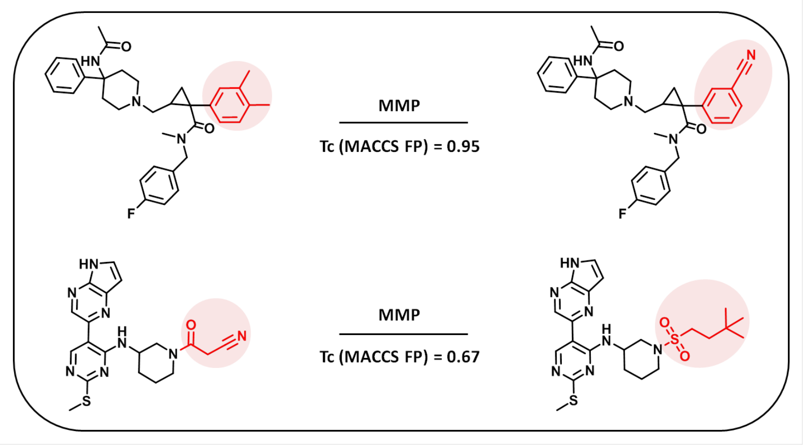


Figure 5: Molecular similarity. Two exemplary MMPs are displayed, for which the corresponding Tanimoto (Tc) similarity values calculated on the basis of MACCS substructure FPs differ significantly.

1.5.2 Analog series

Because no universal definition of an AS exists, computational SAR exploration employs molecular similarity concepts to group compounds together into series under consideration of their biological activity.[180–182] When undertaking chemical transformations in a series, medicinal chemists typically rely on so-called R-group tables, where different combinations of core and substituent structures can be monitored.[20] Naturally, some of the most popular and intuitive methods for extraction and visualization of SARs and ASs include extensions of conventional R-group tables[202,203] or SAR matrices (SARMs).[204–206]

Alternatively, ASs can be identified by organizing similar compounds around individual cores with one or more substitution sites.[207] One such approach relies on systematic clustering of MMPs, such that each compound in a series forms a MMP with at least one other analog.[208] This method enables the extraction of AS-based (ASB) scaffolds, thereby ensuring that all conserved structural characteristics of the corresponding ASs are retained.[209,210] In another approach based on individual compound-core relationships (CCRs),[211] ASs are formed independently of MMPs. In this case, compounds are first subjected to multi-step fragmentation and subsequently all possible cores are matched without considering information about substitution sites. Analogs are organized into series around the smallest possible generalized core, to which the initial substitution site information is then reversely mapped. Thus, an AS core represents the generalized maximum common substructure of all associated analogs as depicted by the example in **Figure 6**. Both described methods rely on systematic compound fragmentation, which can be optionally guided by retrosynthetic[212] rules and are independent of molecular representations and similarity metrics.

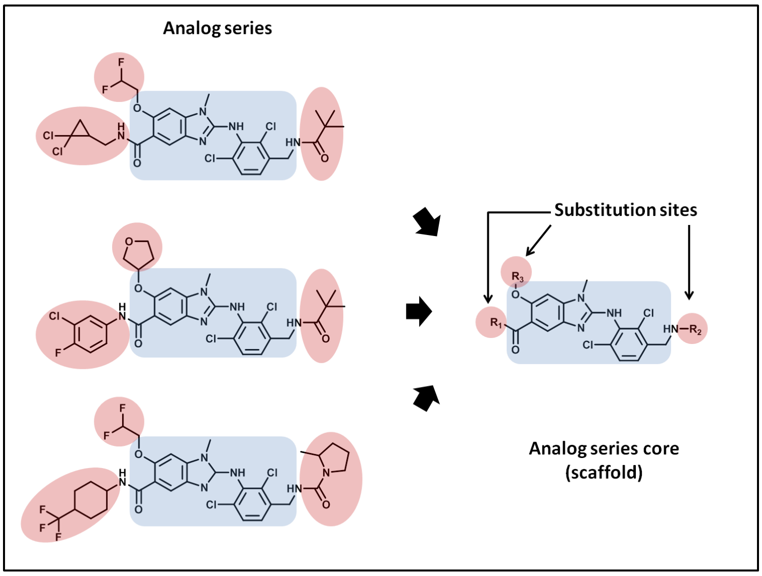


Figure 6: Analog series. Three compounds (left) active against the human prostaglandin E synthase sharing the same core structure (blue) are organized in an exemplary AS. The extracted scaffold contains three substitution sites, where R-groups are exchanged (red).

1.5.3 Chemical neighborhoods

Following the aforementioned definitions, chemical space covered by an AS can be viewed as core-centric. Depending on the variation of R-groups at each substitution site, analogs may map to different areas of AS-relevant chemical space. Hence, chemical neighborhoods (NBHs) of highly similar analogs can emerge as subclusters in individual series.[213] This is typically a likely consequence of medicinal chemists focusing their efforts on potentially more promising candidates and their nearest neighbors during LO.[20,197] Chemical NBHs provide an intuitive local structural context and thus a more granular view on subtle SAR patterns in a given series.[195,214]

The chemical NBH concept is also utilized in the Free-Wilson (FW) additivity principle that is widely used in medicinal chemistry.[215,216] Accordingly, individual R-groups at different scaffold substitution sites contribute to changes in compound activity in an independent and additive manner. Thus, a medicinal chemist is able to approximate the potency of a compound prior to its synthesis, given that the potency values of its nearest neighbors with corresponding R-groups are known.[216] FW-type calculations are particularly useful as local "mini-QSAR" models for activity predictions in MMP-based chemical NBHs.[217]

As outlined in the example in **Figure 7**, MMPs are formed between analog I and II, I and III, II and IV, III and IV, respectively, and the net potency difference associated with each R-group exchange is highlighted accordingly (orange or blue). Given the potency values of analogs I, II, and III are known, the potency of analog IV can be approximated by adding the individual net contributions of directed R-group modifications. Here, the activity of analog IV is predicted retrospectively and its real experimentally measured activity is provided for comparison. Importantly, such potency predictions can be derived from different NBHs (if available) for one and the same compound. Thus, their accuracy can be statistically evaluated and potential outliers can be disregarded when calculating the average estimated potency.

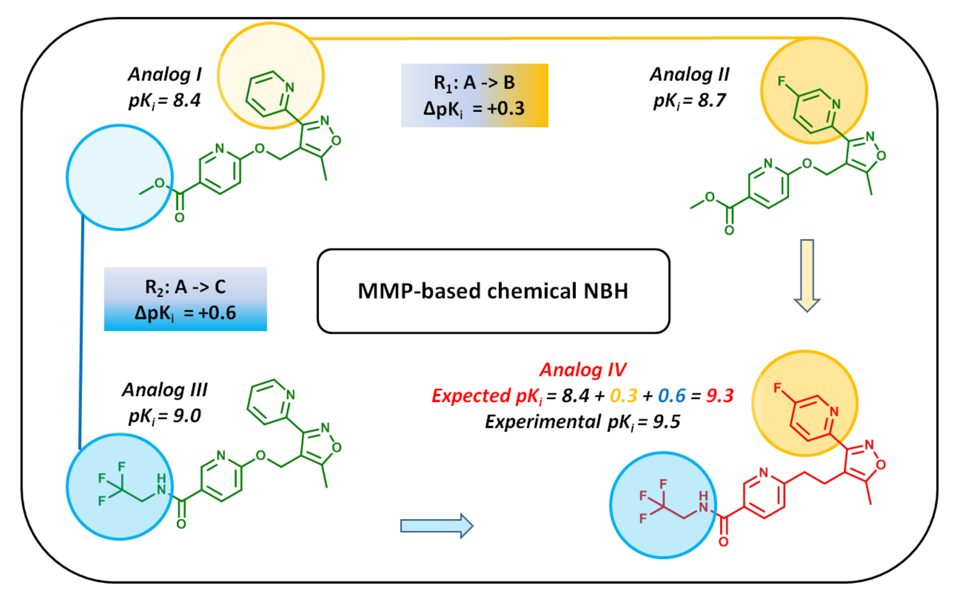


Figure 7: Free-Wilson additivity principle in chemical NBHs. Shown is a four-membered MMP-based chemical NBH comprising inhibitors of GABA receptor alpha-5 subunit together with their potency values as negative logarithmic (log10) constants of inhibition (pK_i). Detailed description is provided in the text. The figure is adapted from the publication in Chapter 6 of this thesis.

1.6 Machine learning

Predicting the biological activity of compounds is one of the main applications of statistical and ML modeling in cheminformatics.[46] Ligand-based QSAR has been traditionally employed for classification or regression problems, such as differentiating between active and inactive molecules or predicting their exact potency values, respectively.[45] In the context of an AS, where all compounds are designed as close structural analogs of an active lead molecule and hence generally expected to be (at least weakly) active, regression modeling represents a more likely application scenario than classification because it does not require an arbitrary potency threshold for separating data into active and inactive compounds in the first place.

1.6.1 Regression models for compound activity prediction

Building a ML model requires each compound to be represented by a molecular feature vector x, composed of descriptors or FPs as input variables, and the corresponding potency, usually as a negative logarithmic (log10) value, serving as output variable y.[43] Furthermore, compound data is typically partitioned into training, internal validation, and external validation sets.[218] The training set is used to fit models with varying hyper-parameters, which are then evaluated on the internal validation set. Once a model with an optimal parameter setting has been selected, it is used to retrospectively predict the potency values in the external validation set. A double (internal and external) cross validation procedure with multiple different random data splits is typically carried out in order to avoid overfitting and ensure model robustness.[219,220] Finally, overall model quality is assessed by juxtaposing predicted and expected values and computing performance measures. The most widely used measure for regression problems is the coefficient of determination R^2 defined as:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3}$$

where, y_i and $\hat{y}_i$ are the expected and predicted values of instance i, respectively, and $\bar{y}$ is the mean expected value of the underlying data set. The maximum value for R^2 is 1, which translates into ideal model performance, whereas a value of 0 or a negative value indicates that the model is performing as good as or worse than simply predicting the same value $\bar{y}$ for every instance.

Some of the most commonly used statistical methods and ML algorithms in cheminformatics include linear/logistic regression (LR),[46,50] random forests (RFs),[221,222] support vector machines (SVMs),[223,224] and deep neural networks (DNNs).[225,226] A LR aims to fit a line through the input data distribution by minimizing the sum of the squared residuals between predicted and expected output values and assigning weight coefficients as follows:

$$f(x) = wx + b \tag{4}$$

where, x and w are the input variable and its corresponding weight, respectively, and b is the y-intercept. Advanced LR methods, such as ridge regression, are introduced to alleviate deteriorating model performance that is caused by outliers in the data distribution, by penalizing model weights.[227]

The predictive ability of LR models is typically limited to the presence of linear SARs in a compound data set and therefore more sophisticated ML techniques, such as SVMs, are employed for nonlinear QSAR modeling.[45] Initially introduced as a binary classification method, SVMs are supervised ML models which maximize the margins of a multi-dimensional hyperplane that best separates positive from negative class instances.[228] If linear separation in the underlying feature space is not possible, SVMs allow the use of kernel functions to calculate the relationships between data projections as vector dot products in a higher dimensional feature space, where linear separation might be feasible.[223] This so-called "kernel trick" circumvents the computationally expensive explicit mapping of data instances in (potentially infinite) higher dimensions. Commonly used kernel functions include linear, polynomial, radial basis, and Tanimoto kernels, with the latter often being used in combination with binary FPs in cheminformatics.[229]

Support vector regression (SVR) is an extension of the SVM algorithm designed for numerical predictions (**Figure 8**).[230] Hence, a real-valued output is computed from a feature vector for each data instance and compared to the corresponding real value. Deviations are permitted only within a predefined error range ϵ and are otherwise penalized. This so-called "ϵ-insensitive tube" and the regularization parameter C, which controls the subtle trade-off between error penalization and model complexity, are hyper-parameters, which typically need to be optimized during model training.[231] The final regression function is defined as:

$$f(x) = \sum_i (\alpha_i - \alpha_i^*) K(x, x_i) + b \tag{5}$$

where x_i are the support vectors and a_i, a_i^* the associated weights obtained from solving a convex optimization problem; $K(x, x_i)$ is the kernel function applied to the input feature vector x and support vector x_i, and b is the bias parameter. Thus, SVR relies on the kernel trick to fit a LR function to nonlinearly distributed data and can be utilized for modeling more challenging prediction scenarios that arise from the presence of discontinuous SARs.[232]

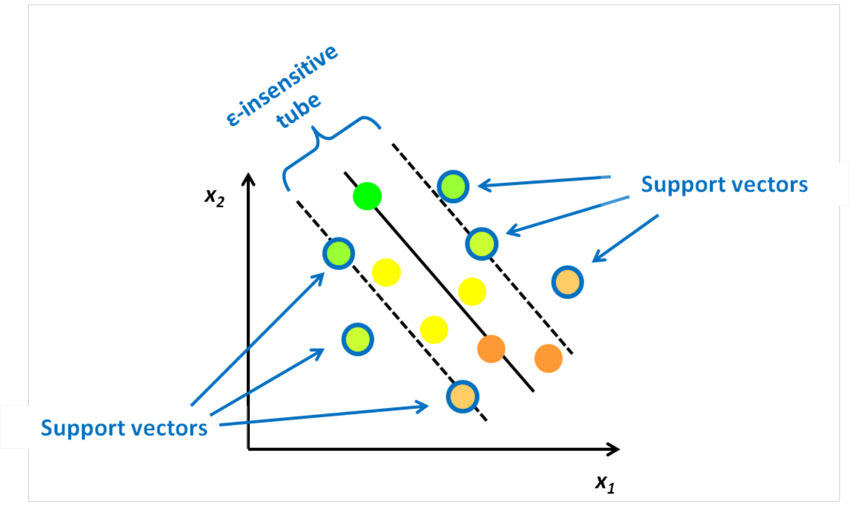


Figure 8: Support vector regression. A schematic illustration of the principle of SVR. Circles represent data points (e.g. compounds) with increasing numerical values (e.g. potency values) as color gradients from red to green.

1.6.2 Generative deep learning methods

In the last several years, generative DL methods[225,233] have become increasingly popular in molecular *de novo* design due to their ability to automatically generalize and infer knowledge from large amounts of chemical structure and reaction data.[234,235] For example, reinforcement learning (RL), where the "decision-making policy" of a generative model is trained to maximize the reward for sampling compounds according to predefined desirability parameters. has been successfully applied for MPO-guided compound design.[236,237] Another strategy for generating focused compound libraries relies on transfer learning (TL), where knowledge acquired for solving one task is conveyed to solving related tasks.[238,239] Hence, models are first pre-trained on large compound data sets representing a broad chemical space and then fine-tuned by further training on specific small data sets, in order to sample compounds within a desired confined area of chemical space.[240–242] Commonly used algorithms for generative modeling include autoencoders (AEs),[243–245] generative adversarial networks (GANs),[246,247] and recurrent neural networks (RNNs)[248,249] which can operate on either molecular graph[250,251] or textual[252,253] compound representations.

One of the most extensively explored and computationally inexpensive generative approaches utilizes RNNs in combination with SMILES sequences.[57] An RNN is a type of a DNN which can operate on sequential data types in a time-dependent manner.[225,233] A DNN consists of multiple nodes (also called neurons or units) which receive, transform, and pass forward numerical information.[254,255] Nodes receiving the same input data and operating in parallel are grouped into layers and the output of each layer is transmitted to the next one according to the way adjacent layers are connected to each other (fully or partially).[254] DNN architectures consist of an input layer, two or more hidden layers in the middle, and an output layer. Input variables are formatted by the input layer, transformed during forward propagation through the hidden layers, and finally converted into output values by the output layer.[254] During training the weights of the network are iteratively modified via a backpropagation algorithm with the objective of minimizing the error between predicted and expected output values.[256] Model overfitting can be decreased by artificially injecting "noise" into the network through dropouts that randomly convert subsets of output values in each layer to zero.[257]

1.6.3 Recurrent neural networks

In an RNN, in addition to the training data input, hidden nodes also receive information about their respective "hidden states" from previous time steps (sequentially behind the current input).[255] This additional temporal dimension allows RNNs to process sequences of arbitrary length and complexity and hence they have been successfully exploited in domains such as natural language processing,[258] speech recognition,[259] formal language,[260] and computer code generation.[261] The most widely used node types in RNNs include long short-term memory (LSTM)[262] or gated recurrent unit (GRU)[263] neurons, which contain memory cells stabilizing the value gradients during training. In the context of *de novo* design, RNNs are trained to learn the universal SMILES syntax of entire molecules[264] or individual fragments[265] and (re)produce chemically meaningful strings. Due to their internal "memory" of previously seen sequence elements, they are able to capture complex contextual patterns in SMILES, such as side chains or ring openings/closures.[266] Thereby, presenting the RNN model with different SMILES variants of one and the same molecule during training can serve as a computationally efficient data augmentation technique, which reduces model overfitting caused by learning strict canonical representations and increases the proportion of syntactically valid SMILES by improving the generalization capability of the model.[267]

In the first step of the training procedure, individual SMILES characters (or character combinations such as "Cl" or "[nH]") are typically tokenized by converting them into so-called "one-hot" numerical representations before presenting them to the model.[268] Thereby, each unique sequence element is assigned a k-dimensional bit vector, where only a single position specifically corresponding to that element is set to 1, and k is equivalent to the model vocabulary size i.e. the number of possible tokens that can be sampled by the model (**Figure 9**). Furthermore, special "start" and "end" tokens are added as placeholders before and after the SMILES string to define its limits.

The tokenized SMILES sequence is then injected to the model one element at a time, beginning with the start token, and propagated along the network.[255] At time step t, a hidden state h_t, which contains d hidden nodes and receives one-hot input vector x_t $(k \times 1)$, comprises an input W_X $(d \times k)$, a hidden state W_H $(d \times d)$, and an output W_Y $(k \times d)$ weight matrix. These three weight matrices remain constant across all time steps within a single RNN loop. At each time step the model computes its current hidden state vector h_t $(d \times 1)$ as follows:

$$h_t = a(W_H \cdot h_{t-1} + W_X \cdot h_t) \tag{6}$$

where a is a nonlinear activation function, h_{t-1} $(d \times 1)$ is the hidden state vector from the previous time step, x_t is the input vector from the current time step, W_H and W_X are the hidden state and input matrices, respectively. In the special case when the start token x_1 is passed, the initial hidden state vector h_0 can be set to a null vector or pre-conditioned to initiate a more focused sampling within a certain output domain.[269]

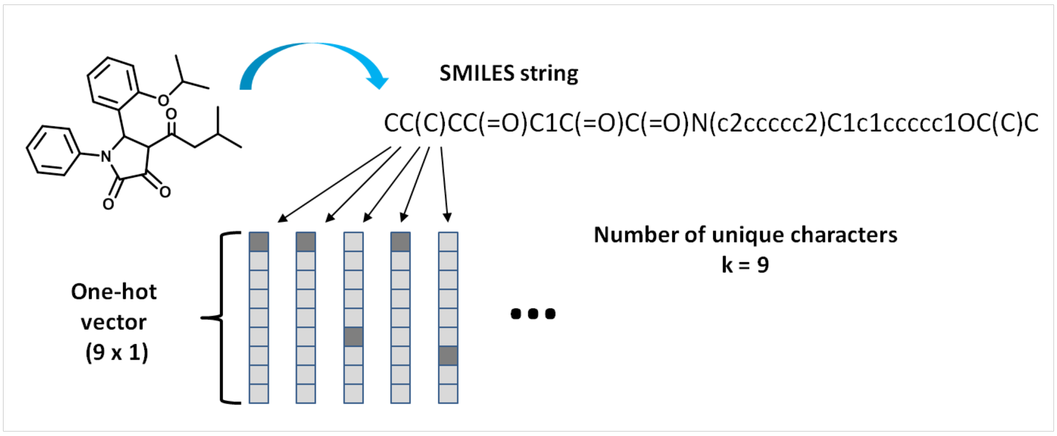

SMILES string
CC(C)CC(=O)C1C(=O)C(=O)N(c2ccccc2)C1c1ccccc1OC(C)C
Number of unique characters
k = 9
One-hot
vector
(9 x 1)
N
O
O
O
O

Figure 9: Tokenization of SMILES. A schematic illustration of tokenizing individual syntactic elements in a SMILES string.

Since the model objective is to predict the next character of the sequence, the hidden state vector h_t obtained from the equation above is used to calculate the current output $\hat{y}_t$ as follows:

$$\hat{y}_t = g(W_Y \cdot h_t) \tag{7}$$

where g is the output function and W_Y is the output weight matrix. The predicted value is then used as input x_{t+1} at the next time step $t+1$. Alternatively, the expected output may be presented to the model instead, in an approach referred to as "teacher's forcing", which leads to faster model convergence.[270] This cycle is repeated for T time steps with T being the maximum permitted length of the SMILES string. For sequences shorter than T, the resulting buffer can be padded with start or end tokens.

Once the entire SMILES string is sampled, its deviation from the expected sequence is calculated as the total loss.[271] Thereby, the probability distributions of the real y_t and predicted $\hat{y}_t$ values are compared at each time step t and summed for the whole sequence as:

$$J(y, \hat{y}) = -\sum_{t=1}^{T} y_t \cdot \log(\hat{y}_t) \tag{8}$$

where J is the categorical cross entropy loss function, while y and $\hat{y}$ are the expected and predicted output sequences for all T time steps, respectively. Finally, an RNN training loop is completed by backpropagation, where the gradients for adjusting the model weights are computed with respect to the total training loss.[271] **Figure 10** provides a schematic overview of the RNN training procedure.

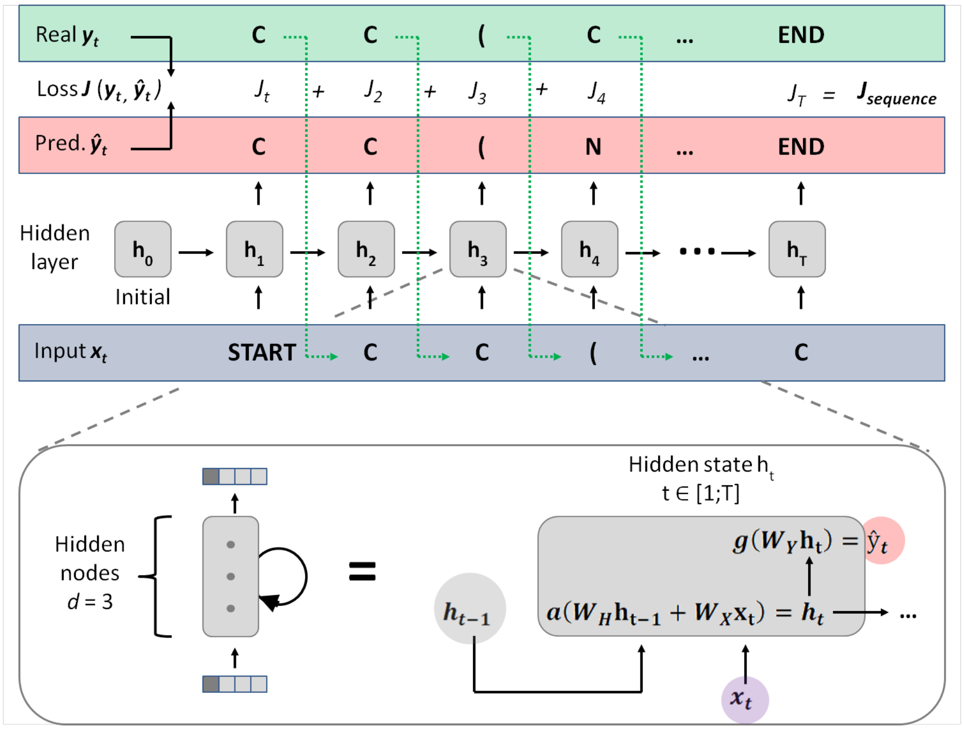


Figure 10: Recurrent neural network. Depicted are two different views of the basic RNN architecture and the model training process. **(bottom)** A hidden layer with three hidden nodes receives a tokenized one-hot vector as input x_t and a hidden state vector h_{t-1} from its previous hidden state. Based upon that, the current hidden state vector h_t is computed, used to calculate an output one-hot vector $\hat{y}_t$, and then passed along to the next hidden state. **(top)** An input sequence is injected to the model one step at a time beginning with a "START" token and an initial hidden state h_0. The next sequence element is predicted by the model and compared to the expected output at the current time step using the loss function (the output sequence is shifted one time step further from the input sequence). The expected output is used as input for the next time step (green dotted lines). In this example, the first three SMILES characters ("C", "C", "(") as well as the last token ("END") are predicted correctly, whereas the fourth sequence element ("C") is predicted incorrectly ("N")

.

1.7 Thesis outline

This dissertation follows the iterative development of *in silico* methods for rationalizing chemical LO and compound design on the basis of individual ASs. *Chapter 2* shortly elaborates on already existing similar approaches reported in the literature and places the scope of the methods developed in this thesis in a scientific context. *Chapter 3* to *Chapter 7* include five original publications in a chronological order representing the evolution of the methods described herein. *Chapter 3* examines the influence of varying calculation parameters on a newly proposed concept of chemical saturation within ASs *Chapter 4*, a method combining improved assessment of chemical saturation and novel SAR progression scoring is introduced for further series characterization. *Chapter 5* presents the COMO methodology and its application for categorizing ASs according to their LO development stage. *Chapter 6* elucidates the extension of COMO with an integrated strategy for design and prioritization of new analog molecules. In *Chapter 7*, DL models are utilized for generating focused virtual compound libraries for ASs profiled with COMO and the outcome is compared to previously explored rule-based *de novo* design strategies. Finally, *Chapter 8* reiterates over the most important findings and concludes this thesis.

Chapter 2

Overview of Computational Methods for Series-based Evaluation of Progress in Lead Optimization

Given its compound-centric nature, the LO process aims at finding a single "ideal" candidate at the intersection of multiple properties. Consequently, computational methods that can be applied in the context of LO, such as QSAR/QSPR, MPO techniques or drug-likeness measures, are predominantly focused on evaluating and predicting individual compound characteristics. However, these approaches are of limited use for assessing LO progress in general. In practice, it is difficult to predict the trajectory of a growing AS and the outcome of a LO campaign. Moreover, knowledge transfer between different programs is not necessarily straightforward since the blueprint of each project is likely to have its own specific endpoints. Objective evaluation is further complicated by the fact that as compound series grow larger, it tends to become more difficult to discontinue them, especially after a significant amount of effort has been already invested. Hence, an AS is usually terminated only when insurmountable roadblocks are met. However, decisions on whether to continue can be often taken at a much earlier stage, provided that progression can be objectively monitored. In light of these challenges, *in silico* methods for rationalizing the course of LO on the basis of entire ASs are highly desirable in order to support decision-making in a more data-driven and less biased manner.

To these ends however, only few such methods have been previously introduced. One of these approaches is LO attrition analysis, where progression of individual ASs is determined by the proportions of experimentally tested analogs passing or failing certain predefined filters.[272] Another method is LO telemetry, which employs statistical risk assessment of individual compounds in order to derive a global model for calculating the odds for success of different ASs and visualizing important milestones in LO convergence.[273] Furthermore, in a different approach, the SARM data structure is utilized as a chemically intuitive tool for monitoring SAR progression in evolving series by quantifying the amount of non-redundant SAR information being added with newly synthesized compounds.[274]

In addition to these approaches, which are relatively simple, robust, and powerful in their ability to discriminate between different AS phenotypes, a novel computational method for estimating chemical saturation in ASs has been introduced.[275] Accordingly, it quantifies how extensively the chemical space of an individual series has been explored through synthetic efforts. Thereby, AS-relevant chemical space is delineated by populations of (not yet synthesized or tested) virtual analogs (VAs), which are projected alongside already existing analogs (EAs) in a descriptor-defined chemical reference space allowing for calculation of inter-compound distance relationships. Based on those, local chemical NBHs around EAs are derived, leading to the development of a dual scoring scheme for categorizing series into different stages of LO.

In conclusion, the latter method proposes the concept of chemical saturation as a measure for distinguishing between ASs and serves as a starting point for the development of the methods presented in this thesis.

Chapter 3

Computational Assessment of Chemical Saturation of Analogue Series under Varying Conditions

Introduction

The initial method for assessment of chemical space saturation has been introduced as a proof-of-concept for discrimination between compound series of different LO stages. In this case, exemplary ASs have been computationally extracted from biological assay data (PubChem) containing both active and inactive compounds against a given target. Depending on that, two types of chemical NBHs have been defined based on distances between EAs and/or VAs, giving rise to a global and a complementary local saturation score.

In this chapter, the initially developed local saturation score is modified and the robustness of the dual scoring scheme explored under varying parameter settings. In addition to the previously analyzed ASs from PubChem, new series are extracted from medicinal chemistry literature, comprising exclusively active analogs. Alternative molecular descriptors constituting the chemical reference space, in which compounds are projected, are compared. Furthermore, differently designed VA populations are explored and the influence of varying numbers of VAs used for scoring is benchmarked. Finally, the chemical saturation behaviour of growing ASs is analyzed.

My main contribution to this work was carrying out all calculations under varying conditions and providing statistics and visualization for subsequent analysis.

ACS OMEGA

Article

Cite This: ACS Omega 2018, 3, 15799−15808

http://pubs.acs.org/journal/acsodf

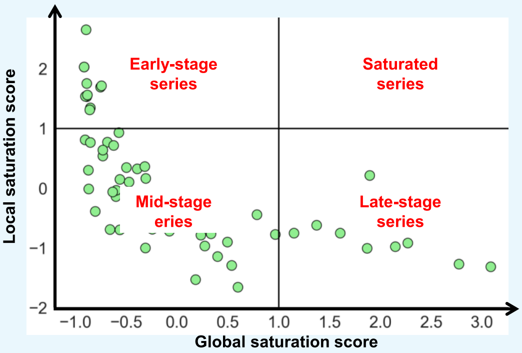
Computational Assessment of Chemical Saturation of Analogue Series under Varying Conditions

Dimitar Yonchev, Martin Vogt, Dagmar Stumpfe, Ryo Kunimoto,[†] Tomoyuki Miyao,[‡] and Jürgen Bajorath*

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

S Supporting Information

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



1. INTRODUCTION

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

So far, only a few approaches have been introduced to evaluate the generation of analogue series. These approaches include multiparameter optimization (MPO) using desirability functions,[1] attrition curves,[2] or risk statistics,[3] which balance multiple compound properties. However, MPO does not directly assess progression saturation of analogue series, but it prioritizes candidate compounds with desirable property profiles. MPO can hence be applied to indirectly evaluate series progression by prioritizing candidate compounds with preferred property combinations that can still be obtained. Regardless of the MPO approach that is applied, a pool of candidates for evaluation must be generated separately. In addition, numerical[3,4] and graphical[5] SAR analysis methods have been introduced to monitor SAR progression of evolving compound series and evaluate whether newly generated analogues yield further SAR information. These approaches may also suggest preferred candidate compounds[3] or provide diagnostics for SAR landscapes of compound data sets.[4,5]

Previously, we have introduced a computational concept to more directly and quantitatively assess progression saturation of analogue series using virtual analogue populations that are mapped into chemical reference space together with exiting analogues.[6] The approach is focused on the assessment of chemical progression saturation, but not SAR progression, and addresses the following questions: Do existing analogues provide sufficient chemical space coverage? Are virtual candidate compounds available that have a high likelihood of activity? As discussed in detail below, the methodology requires the use of chemical reference spaces and virtual candidate compounds to quantify chemical saturation. The degree to which the results of computational saturation analysis depend on such parameters is yet to be explored.

Received: August 18, 2018
Accepted: November 8, 2018
Published: November 20, 2018

DOI: 10.1021/acsomega.8b02087

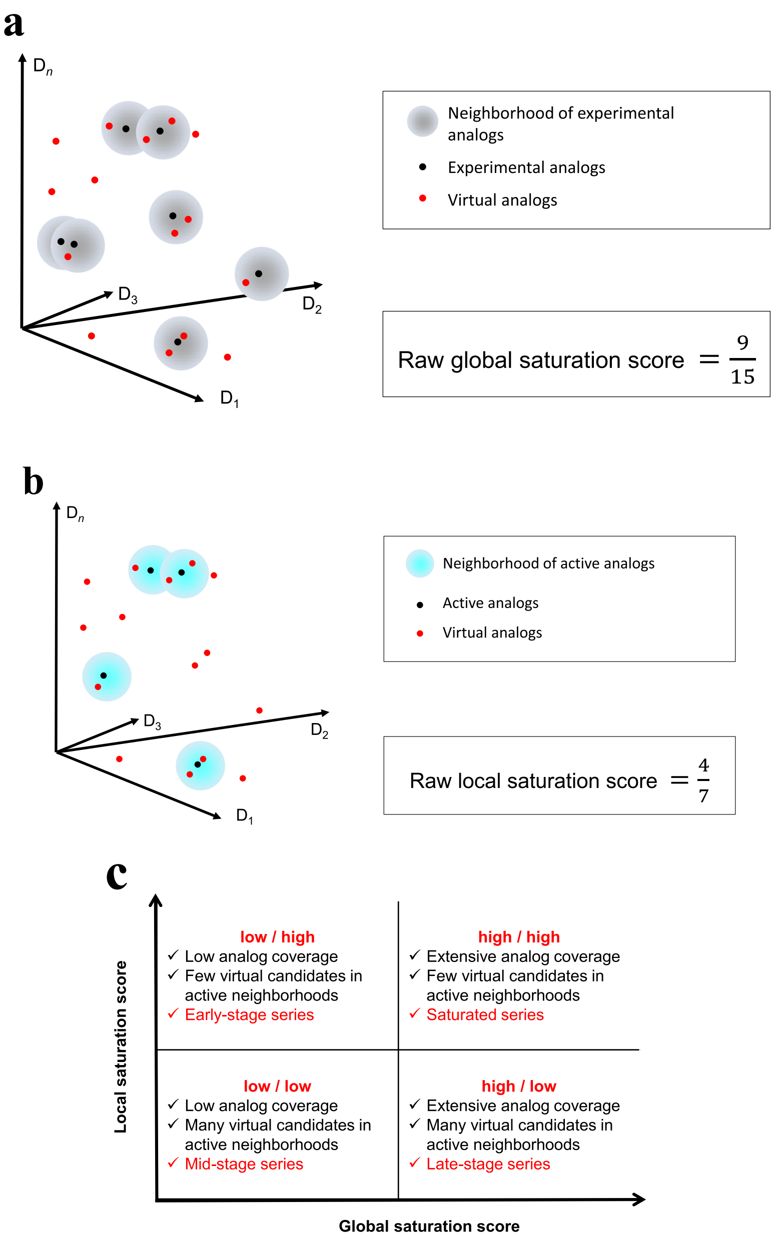

a
D_n
D_3
D_2
D_1
Neighborhood of experimental analogs
Experimental analogs
Virtual analogs
Raw global saturation score = 9/15
b
D_n
D_3
D_2
D_1
Neighborhood of active analogs
Active analogs
Virtual analogs
Raw local saturation score = 4/7
c
low / high
✓ Low analog coverage
✓ Few virtual candidates in active neighborhoods
✓ Early-stage series
high / high
✓ Extensive analog coverage
✓ Few virtual candidates in active neighborhoods
✓ Saturated series
low / low
✓ Low analog coverage
✓ Many virtual candidates in active neighborhoods
✓ Mid-stage series
high / low
✓ Extensive analog coverage
✓ Many virtual candidates in active neighborhoods
✓ Late-stage series
Local saturation score
Global saturation score

Figure 1. Saturation scores and categorized combinations. (a, b) Calculation of the raw global and local saturation scores is illustrated, respectively. The coordinate system represents a chemical reference space containing an analogue series and virtual candidates. "D" stands for descriptor. Each descriptor adds a dimension to the space. (c) Combinations of global and local saturation scores are categorized as indicators of different progression saturation stages. Figure panels were adapted from ref 6.

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

2. MATERIALS AND METHODS

2.1. Methodology. *2.1.1. Concept.* Progression saturation of the analogue series is evaluated by comparing distributions of existing analogues and virtual candidate compounds for series expansion in chemical space[6] applying a neighborhood concept.[7,8] First, chemical space coverage of the analogue series is estimated by determining the proportion of virtual candidates falling into predefined neighborhoods of existing analogues. For this purpose, a global saturation score is calculated, as defined below. In this case, chemical neighborhoods of analogues are defined on the basis of nearest-neighbor distances between virtual compounds to measure global coverage of the chemical space. Second, the population of neighborhoods of active analogues (active neighborhoods) is assessed by determining virtual candidates falling into active neighborhoods. Therefore, neighborhoods are defined differently on the basis of median distances between active compounds in the chemical space and a local saturation score is calculated, as also defined below. Relating global and local compound distributions and resulting scores to each other makes it possible to evaluate progression saturation of the analogue series. The assessment is primarily focused on chemical progression saturation (rather than SAR progression), addressing the questions whether chemical space coverage by existing analogues is extensive and, in addition, whether a significant number of virtual candidates exist that are likely to be active.

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

$$\text{raw global saturation score } (S) = \frac{|\nu_{\text{Exptl}}|}{|V|}$$

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

The so-defined raw global saturation score, the calculation of which is illustrated in Figure 1a, measures the chemical space coverage of existing analogues and virtual candidates. Hence, the larger the score, the more virtual analogues map to neighborhoods of experimental analogues, indicating extensive coverage. Depending on the composition of a series, S may contain both active and inactive or only active analogues.

The raw local saturation score is defined as the ratio of the number of active analogues relative to the number of virtual analogues falling into the neighborhoods of active analogues

$$\text{raw local saturation score } (A) = \frac{|A|}{|V_{\text{active}}| + 1}$$

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

The raw global scores and logarithmically transformed local saturation scores are converted into conventional Z-scores on the basis of the mean and standard deviation of the score population of large sets of analogue series. Accordingly, the mean of the score population was subtracted from each raw score and the difference was divided by the standard deviation of the distribution, yielding the Z-score. Accordingly, the resulting Z-scores have a mean of 0 and standard deviation of 1. Global and local scores generally display low correlation.[6]

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

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

2.2. Analogue Series.

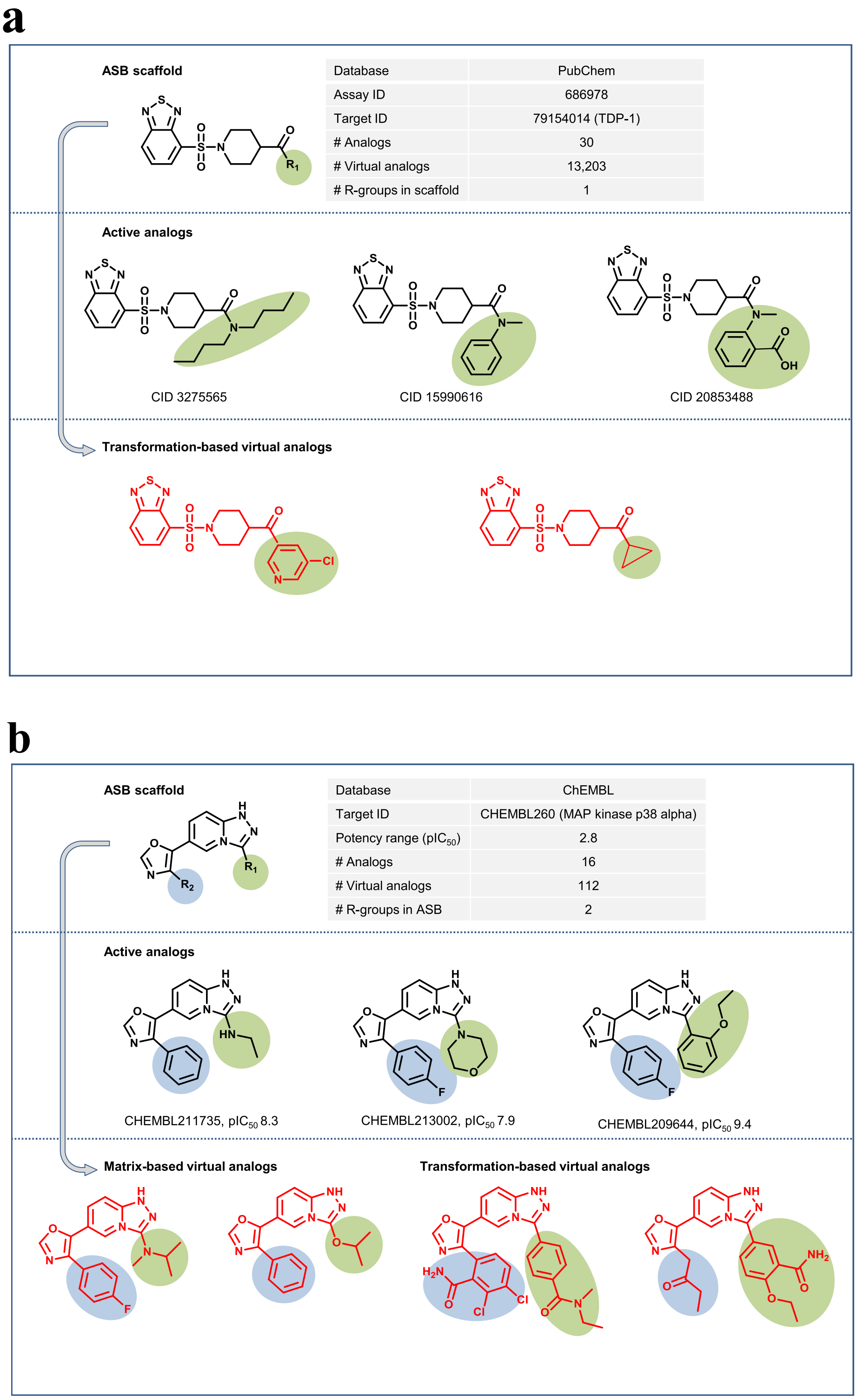
Analogue series were assembled using a computational method that identifies analogue series in compound data sets of any composition.[9] This method makes use of the matched molecular pair (MMP) concept.[10] An MMP is defined as a pair of compounds that are distinguished by a chemical modification at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

Compounds sharing the same RECAP-MMP core form a matching molecular series (MMS),[14] which represents an analogue series with a single substitution site.[9] By contrast, different MMSs sharing analogues form a series with multiple substitution sites.[9] In this case, substitution sites of the corresponding RECAP-MMP cores are transferred to shared analogues from which an analogue-series-based (ASB) scaffold[15,16] with multiple substitution sites[16] is extracted. This ASB scaffold then represents an analogue series with multiple substitution sites.

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

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

DOI: 10.1021/acsomega.8b02087



Database	PubChem
Assay ID	686978
Target ID	79154014 (TDP-1)
# Analogs	30
# Virtual analogs	13,203
# R-groups in scaffold	1

Database	ChEMBL
Target ID	CHEMBL260 (MAP kinase p38 alpha)
Potency range (pIC_{50})	2.8
# Analogs	16
# Virtual analogs	112
# R-groups in ASB	2

a
ASB scaffold
R_1
Active analogs
CID 3275565
CID 15990616
CID 20853488
Transformation-based virtual analogs
b
ASB scaffold
R_2
R_1
Active analogs
CHEMBL211735, pIC_{50} 8.3
CHEMBL213002, pIC_{50} 7.9
CHEMBL209644, pIC_{50} 9.4
Matrix-based virtual analogs
Transformation-based virtual analogs

Figure 2. Representative analogue series. Exemplary compounds of series from (a) PubChem and (b) ChEMBL are shown with associated target and potency information and corresponding virtual analogues (red). In addition, the composition of each series is reported and its ASB scaffold is shown.

computational progression saturation analysis.[6] They are termed PubChem series in the following.

2.2.2. Series from Medicinal Chemistry. A set of 64 analogue series yielding ASB scaffolds with multiple substitution sites were extracted from ChEMBL (release 23)[18] on the basis of high-confidence activity data. Accordingly, only compounds with direct interactions (type "D") with human targets at the highest assay confidence level (confidence score 9) were selected. As potency measurements, only specified equilibrium constants (K_i values) or IC_{50} values were considered. These analogue series, termed ChEMBL series in the following, were required to consist of 10−30 active analogues. They contained a total of 1422 compounds and covered 62 unique targets.

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

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

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

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

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

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

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

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

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

Table 1. Descriptors for Chemical Reference Spaces[a]

	descriptor name	description
set 1	a_acc	number of H-bond acceptor atoms
	a_con	number of H-bond donor atoms
	b_1rotN	number of rotatable single bonds
	log *P* (o/w)	log octanol/water partition coefficient
	logs	log solubility in water
	TPSA	topological polar surface area
	weight	molecular weight
set 2	petitjeanSC	topological shape index (diameter − radius)/radius
	rsynth	synthetic feasibility based on retrosynthetic rules
	PEOE_VSA_FNEG	fractional negative van der Waals surface area
	balabanJ	topological index (Balaban distance connectivity)
	PEOE_RPC+	relative positive partial charge
	PEOE_VSA_FPPOS	fractional polar positive van der Waals surface area
	a_nN	number of nitrogen atoms

[a] Descriptors used for the design of chemical reference spaces are described according to the Molecular Operating Environment with which they were calculated. Set 1 constitutes a seven-dimensional reference space, and sets (1 + 2) form a 14-dimensional space.

3. RESULTS AND DISCUSSION

3.1. Progression Saturation Analysis. The scoring scheme underlying progression saturation assessment is illustrated in Figure 1a,b. Global and local saturation scores

DOI: 10.1021/acsomega.8b02087

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

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

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

3.3. Alternative Chemical Reference Spaces. For the assessment of progression saturation, analogue series must be projected into chemical reference spaces. We first investigated the influence of alternative chemical space representations on score and series categorization. Figure 3 shows the comparison of 7-dimensional (Figure 3a) and 14-dimensional reference spaces (Figure 3b) for 80 PubChem series in the presence of 13 203 transformation-based virtual analogues. In both reference spaces, similar score combinations were obtained for analogue series, leading to a closely corresponding assignment of series to different progression stages (Figure 3c). In both reference spaces of different dimensionalities, the majority of series (60 vs 58) fell into the low/low global/local saturation score category, 13 series belonged to the high/low category, and only three series were assigned to the high/high category. As shown in Figure 4, equivalent observations were made for the 64 ChEMBL series in the presence of 13 203 transformation-based virtual analogues. In both chemical reference spaces, most series (43 vs 49) belonged to the low/low category. In this case, no series were assigned to the high/high category. Hence, both sets of PubChem and ChEMBL series were dominated by analogue series with mid-stage character. Only 17 PubChem and 12 ChEMBL series were found to belong to different categories in the 7- and 14-dimensional reference spaces, consistent with the closely corresponding distributions observed in Figures 3 and 4. Thus, scoring was stable in both reference spaces and very similar assignments were obtained, indicating that progression saturation assessment was not sensitive to chemical reference space variation. Furthermore, the score distributions in Figures 3 and 4 also reveal a significant spread of series across the scoring range, indicating the capacity of global and local scores to distinguish between different series. At least for analogue series from chemical optimization projects published in the medicinal chemistry literature, the observed prevalence of series with a mid-stage character would be expected.

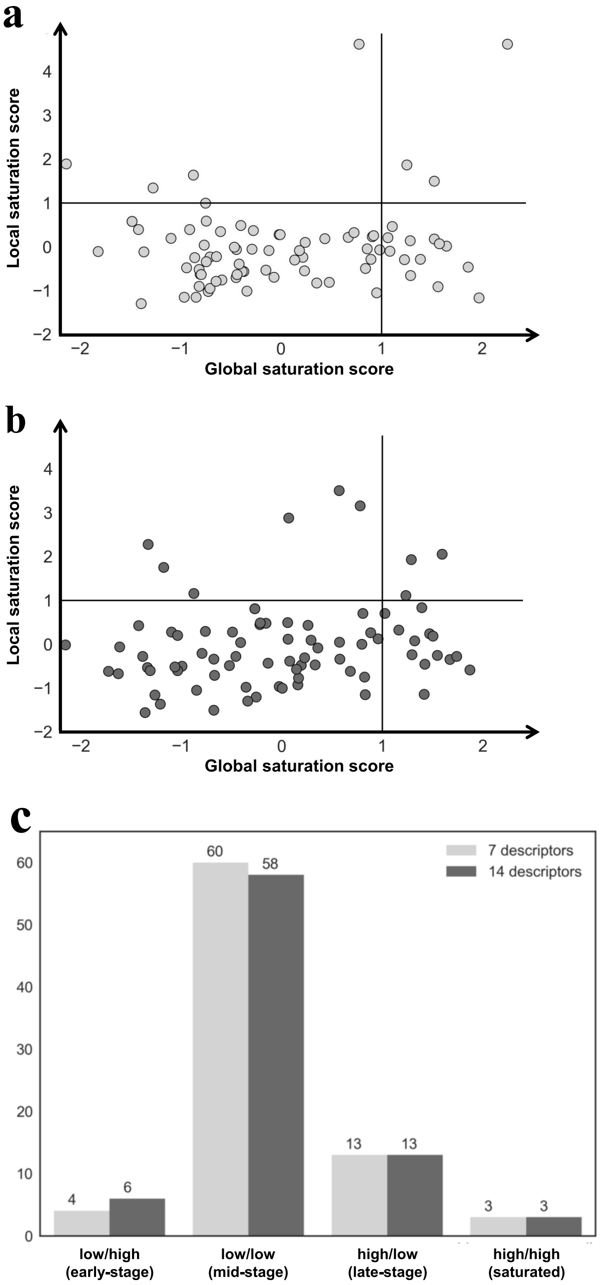


Figure 3. Analysis of PubChem series in different chemical reference spaces. Progression saturation of analogue series from PubChem with 13 203 virtual candidates is assessed in (a) 7-dimensional and (b) 14-dimensional reference spaces. Scatter plots report local and global saturation scores obtained for all series (each dot represents a series). (c) Assignment of series to different score combination categories (according to Figure 1c).

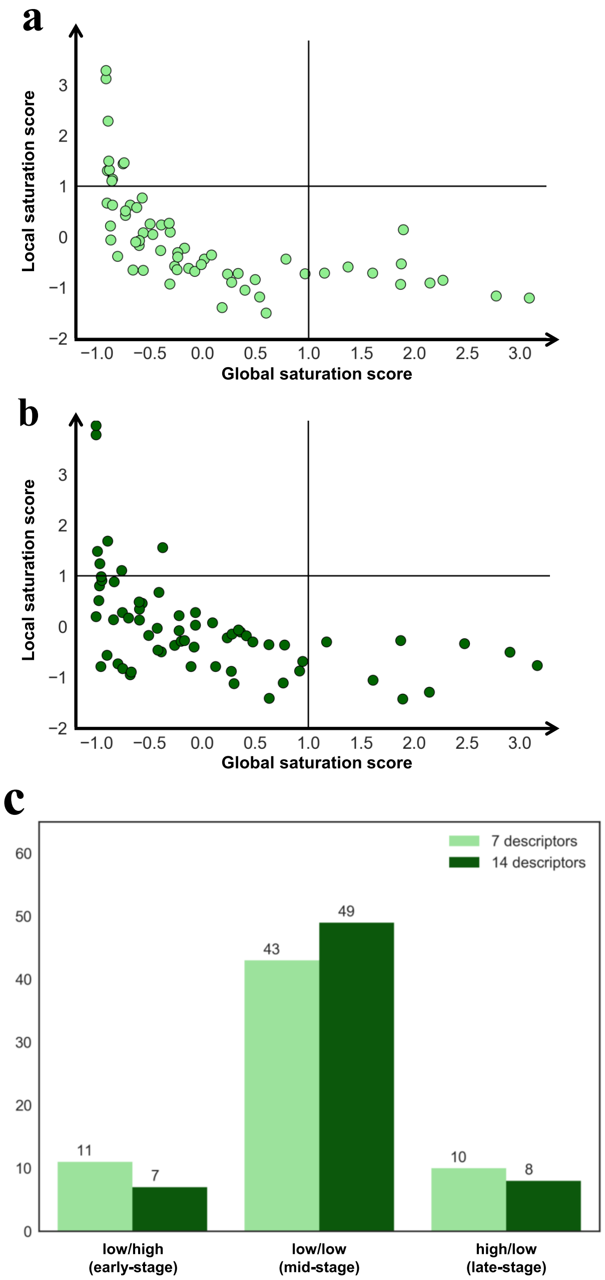


Figure 4. Analysis of ChEMBL series in different chemical reference spaces. Progression saturation of analogue series from ChEMBL with 13 203 virtual candidates was assessed in (a) 7-dimensional and (b) 14-dimensional reference spaces. (c) Assignment of series to different score combination categories.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

3.4. Virtual Analogues of Different Designs. Next, we compared different ensembles of virtual candidates for progression saturation assessment, which represented another parameter of the analysis. For ChEMBL series having multiple substitution sites, matrix-based virtual analogues were generated, yielding varying numbers of 101−501 candidates per series. For each series, the corresponding number of transformation-based virtual analogues were generated and saturation scores were calculated on the basis of alternative sets of virtual analogues in 14-dimensional reference space. Figure 5a,b shows the score distributions obtained for matrix- and transformation-based virtual candidates, respectively. The resulting assignment of series to score combination categories is shown in Figure 5c. Here, moderate changes were observed, predominantly for the low/high and low/low categories.

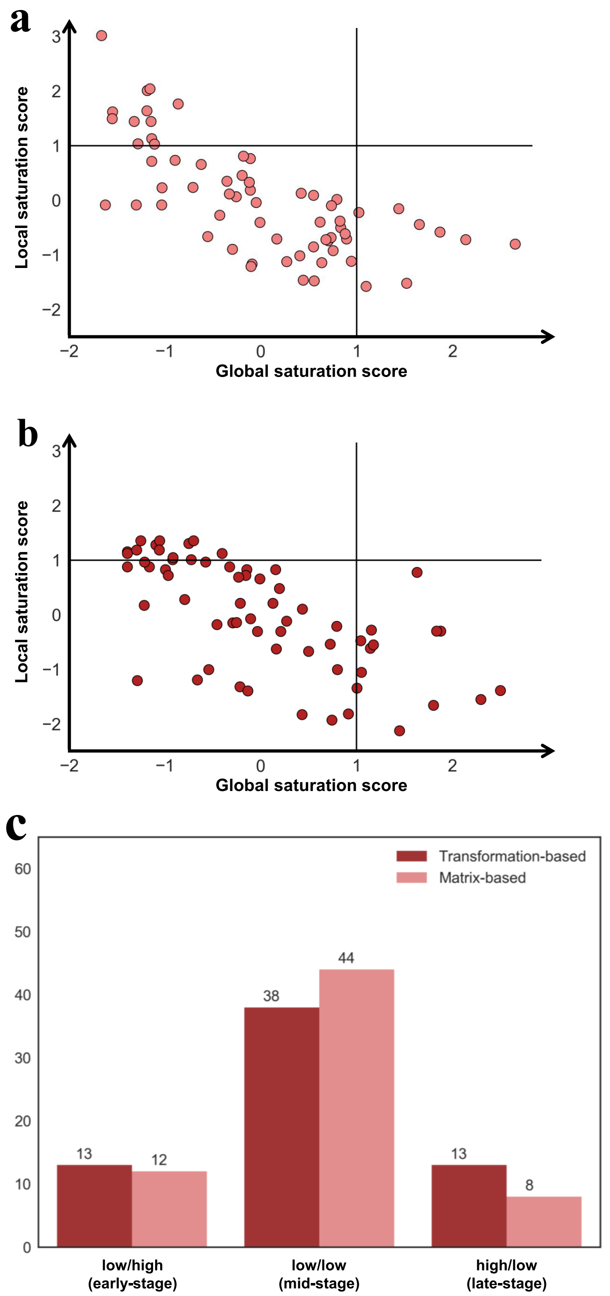


Figure 5. Progression saturation analysis using virtual analogues of different designs. Progression saturation of ChEMBL series was assessed on the basis of corresponding numbers of (a) matrix- and (b) transformation-based virtual candidates. Scatter plots report local and global saturation scores obtained for all series (each dot represents a series). (c) Comparison of the assignment of series to different categories. Here, the raw global saturation score calculation was modified using the top 10% of nearest virtual neighbors for determining the neighborhood radius to account for the smaller number of virtual candidates.

Progression saturation assessment in the presence of transformation-based virtual analogues assigned 13 series to the high/low and 38 series to the low/low category. By contrast, assessment in the presence of matrix-based virtual candidates yielded eight high/low and 44 low/low series. In addition, 13 and 12 low/high series were obtained using transformation- and matrix-based virtual analogues, respectively. Thus, there was a shift from late-stage series toward series with the mid-stage character for matrix-based virtual analogues compared to that of transformation-based candidates. By design, matrix-based virtual analogues were structurally closer to existing analogues than transformation-based virtual candidates, which contained structurally more diverse substituents. Hence, matrix-based candidates should be more likely to map to neighborhoods of active analogues than transformation-based virtual analogues, which would result in lower raw saturation scores. This was consistent with the observation that 22 series contained no transformation-based virtual analogues within neighborhoods of active analogues, whereas all series contained at least a few matrix-based virtual candidates in active neighborhoods. However, lower raw saturation scores do not necessarily translate into significant differences in category assignments because the category of a series is determined on the basis of the magnitude of its global and local scores relative to the scores of the other analogue series. Mapping of transformation- or matrix-based virtual analogues delineates the chemical space of an analogue series. One might argue that the space covered by matrix-based virtual compounds, which are closely related to existing analogues, might more accurately reflect the space that is relevant for a given series.

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

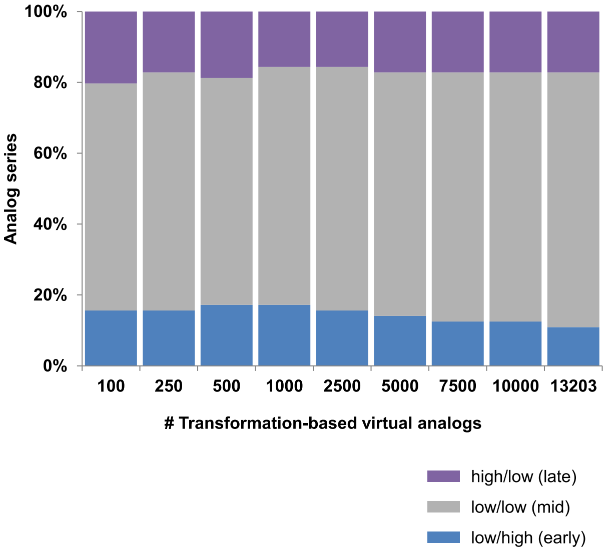


Figure 6. Category distributions for increasing numbers of virtual analogues. For increasing numbers of virtual analogues, score combination categories for ChEMBL series are reported as percentages. As in Figure 4, the top 10% of nearest neighbors were used for determining the neighborhood radius for calculating the global saturation score.

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

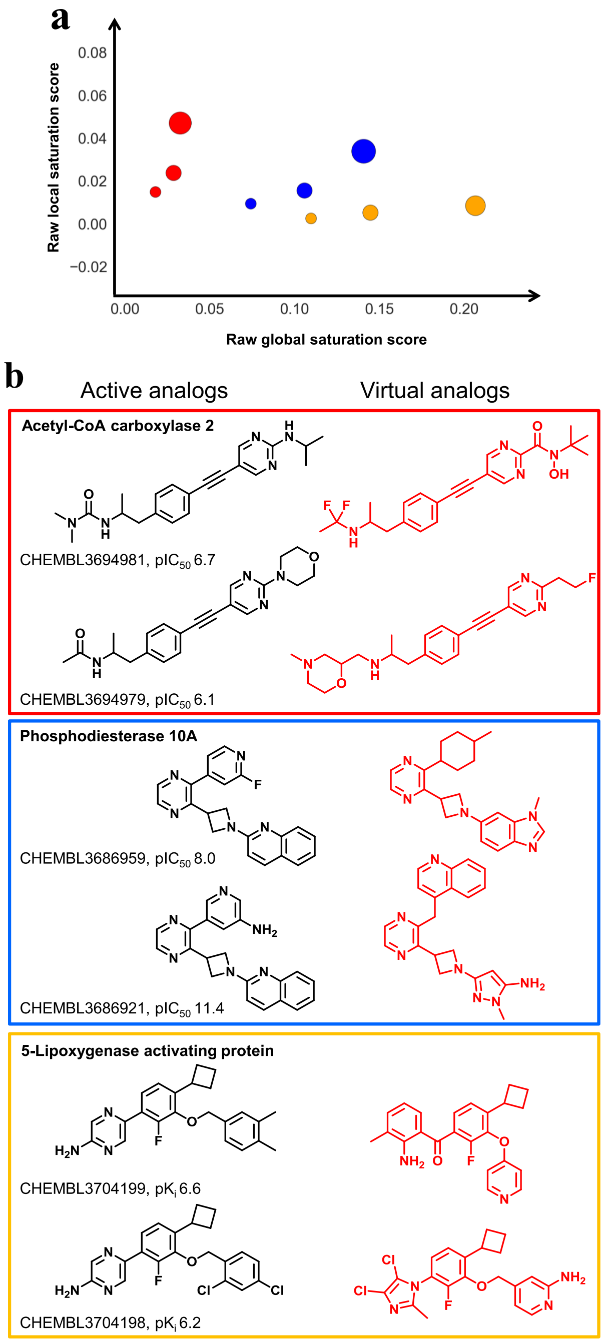


Figure 7. Categorization of evolving analogue series. (a) For three large analogue series, raw global and local saturation scores are reported for subsets of increasing size (small dots, 30 analogues; medium-size dots, 60 analogues; and large dots, all analogues of a series). Theses series include inhibitors of acetyl-CoA carboxylase 2 (red; 126 analogues), phosphodiesterase 10A (blue; 146), and 5-lipoxygenase activating protein (orange; 100). (b) Representative analogues from each series are shown together with transformation-based virtual nearest neighbors (red). The color code of boxes separating analogue subsets from different series corresponds to (a). In addition, the target of each series and the ChEMBL ID and potency value of each analogue are reported.

the right displayed late-stage character and the highest level of progression saturation.

4. CONCLUSIONS

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

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

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

Composition of analogue series (PDF)

■ AUTHOR INFORMATION

Corresponding Author

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

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Present Addresses

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

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

Author Contributions

The study was carried out and the manuscript written with contributions of all authors. All authors have approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Dimitar Yonchev is supported by a fellowship from the Jürgen Manchot Foundation, Düsseldorf, Germany. The authors thank Chemical Computing Group and OpenEye Scientific Software for providing academic software licenses.

REFERENCES

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

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

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

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

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

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

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

(8) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186−3204.

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

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

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

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

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

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

(15) Dimova, D.; Stumpfe, D.; Hu, Y.; Bajorath, J. Analog Series-Based Scaffolds: Computational Design and Exploration of a New Type of Molecular Scaffolds for Medicinal Chemistry. *Future Sci. OA* **2016**, *2*, No. FSO149.

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

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

(18) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.

(19) *Molecular Operating Environment (MOE)*, version 2014.09; Chemical Computing Group ULC: 1010 Sherbooke St. West, Montreal, QC, Canada, 2018.

(20) *OEChem TK*, version 1.7.7; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2012.

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

(22) Gupta-Ostermann, D.; Bajorath, J. The SAR Matrix Method and its Extensions for Applications in Medicinal Chemistry and Chemogenomics. *F1000Research* **2014**, *3*, No. 113.

DOI: 10.1021/acsomega.8b02087

Summary

Herein, analysis of the chemical saturation of ASs has been carried out under varying conditions. Based on combinations of global and local saturation scores, the majority of the studied ASs have been categorised as having mid-stage LO character, regardless of whether they originated from biological screening data or medicinal chemistry literature. Furthermore, expanding the initial seven-dimensional chemical reference space with additional weakly correlated descriptors has not led to significant changes in category assignment of ASs, indicating that the method is rather insensitive to the choice of chemical reference space representation. The two separate VA design strategies explored herein have mainly reflected local saturation scores with matrix-based VAs mapping more closely to EA NBHs than transformation-based VAs. Furthermore, variations in the number of VAs used for score calculations have yielded only moderate changes in scores, which tend to become more stable with increasing VA population sizes. Lastly, scoring of growing ASs has revealed distinct chemical saturation trajectories for each of the three analyzed ASs, thus further corroborating the robustness and discriminative ability of the method.

Taken together, these findings were encouraging for searching for further use cases and methodological extensions, which can account for not only chemical saturation but also SAR progression in ASs.

Chapter 4

Computational Method to Evaluate Progress in Lead Optimization

Introduction

Assessing SAR (dis)continuity in ASs is of major interest in LO campaigns. SAR patterns within single ASs are not necessarily uniformly distributed i.e. some areas in chemical space may be predominantly characterized by continuous and others by discontinuous SARs, thus posing a limiting factor for global QSAR models. Precisely in such cases, the chemical NBH concept can provide a high-resolution view on distinct SAR environments and serve as helpful orientation for medicinal chemists. This principle has been utilized in SAR matrices where local NBHs are defined on the basis of MMP relationships between compounds. However, such NBHs represent discrete constructs due to the strict binary-like definition of an MMP (only a single-site transformation is allowed). Alternatively, the distance-based NBHs described in the previous chapters exhibit a more flexible character (depending on molecular representations and distance types) and represent attractive data structures for local SAR exploration.

In this chapter, chemical saturation analysis is coupled with estimation of SAR progression, which is quantified as a function of NBH-based SAR discontinuity. The two different NBH definitions that have been initially proposed, are combined into a single universal one and based on that, previously introduced chemical saturation scoring is further improved to yield more interpretable score combinations. In this study, chemical saturation is calculated as a function of the degree and density of chemical space coverage by EAs. In light of these methodological modifications, calibration of hyper-parameters, such as number of used VAs and NBH radii, is carried out once again on the basis of newly extracted ASs from ChEMBL. In contrast to the previous chapter, in this analysis only VAs obtained via AS core enumeration are considered for delineating AS-centric chemical space.

My main contribution to this work was the design and implementation of the new scoring system and the subsequent analysis of ASs.

Journal of Medicinal Chemistry

Article

Cite This: J. Med. Chem. 2018, 61, 10895−10900

pubs.acs.org/jmc

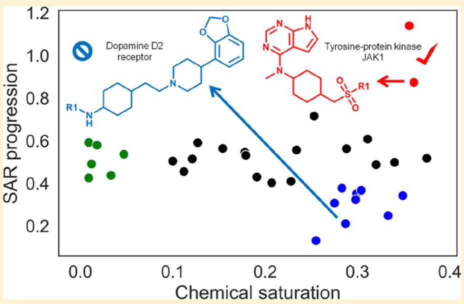
Computational Method to Evaluate Progress in Lead Optimization

Martin Vogt, Dimitar Yonchev, and Jürgen Bajorath*

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

Supporting Information

ABSTRACT: In medicinal chemistry, lead optimization is a critically important task and a highly empirical process, largely driven by chemical knowledge and intuition. Only very few approaches are available to guide and evaluate optimization efforts. It is often very difficult to understand when a compound series is exhausted and the generation of additional analogs unlikely to yield further progress toward potent and efficacious candidates. Rationalizing lead optimization remains an essentially unsolved problem. Herein, we introduce a new computational method to aid in evaluating whether sufficient numbers of analogs have been made and further progress is unlikely. The approach integrates the assessment of chemical saturation and structure−activity relationship progression of compound series. Easy-to-calculate scores characterize evolving analog series and identify candidates with high or low priority for further chemical exploration.



INTRODUCTION

One of the great challenges in medicinal chemistry is rationalizing lead optimization (LO).[1] At each stage of this time-consuming and costly process, a medicinal chemist must decide which compounds to make next in order to reach optimization milestones. Moreover, as an analog series (AS) evolves, it is difficult to judge the odds that further progress will be made in exploring and exploiting structure−activity relationships (SARs). During LO, multiple properties must be balanced to reach the stage when an active compound can be considered for in vivo studies. However, if a compound is not sufficiently potent, it will not become a candidate, irrespective of other features. This requirement makes SAR exploration a central task during early to mid-stages of LO. Given the complexity of LO, this process is largely driven by chemical knowledge, experience, and intuition. There are neither generally applicable optimization routines available nor methods to rationalize LO campaigns and estimate their outcomes. More often than not, roadblocks are encountered during LO and progress made is less than anticipated. However, once substantial efforts have been expanded to advance ASs, terminating a project is a difficult decision to make, even if it is questionable or unlikely that final goals are within reach. Given the essentially subjective nature of LO and the absence of unbiased evaluation criteria, it is often easier to make more analogs and hope to hit a home run than to discontinue a high-profile project. As a consequence, many LO campaigns are carried on for a long time and large numbers of analogs are generated, despite the lack of ultimate success.

In light of this situation, methodologies that help to characterize ASs and evaluate their potential are sought after in medicinal chemistry. Of course, it can hardly be expected that any single methodology will be capable of solving the LO problem. However, any method that can provide decision support during LO is desirable, especially when assessing project progression. Only very few approaches are currently available that can be considered in this context. Among computational methods, quantitative SAR (QSAR) analysis[2] has long been a standard for predicting active analogs for an evolving series. However, QSAR is not capable of assessing the progression of an AS. Going beyond compound potency prediction, statistical multi-objective optimization and analysis methods are applicable to suggest candidate compounds with desirable properties[3,4] or monitor compounds during LO that meet predefined optimization criteria.[4] Furthermore, coverage of chemical space around ASs has been estimated by defining neighborhoods (NBHs) of experimental analogs and screening these NBHs with virtual compounds.[5,6] Other approaches have been introduced to visualize SAR information[7] and monitor SAR characteristics of evolving data sets.[8] Furthermore, numerical SAR analysis functions can be applied to assess SAR progression[9,10] and statistical techniques to identify compounds that make significant contributions to positive progression and reduce the risk of failure.[11] However, although SAR visualization methods and numerical functions enable retrospective analysis, they cannot be used to predict whether further progress in LO might be made.

Herein, we introduce a computational method to estimate whether an AS can be further advanced by generating more analogs or whether this would be unlikely. The approach was designed to integrate the assessment of chemical saturation and SAR progression of ASs. Evaluation of our methodology was based on the largest ASs from medicinal chemistry sources

Received: October 19, 2018
Published: November 30, 2018

DOI: 10.1021/acs.jmedchem.8b01626

that we were able to identify in the public domain. Figure 1 shows representative examples.

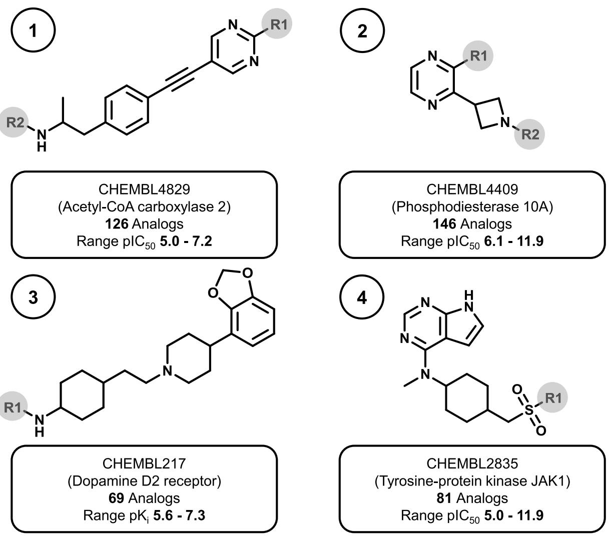


Figure 1. Exemplary ASs. Shown are AS-based scaffolds representing individual ASs. For each AS, the ChEMBL ID, target name, number of analogs, and their potency range are reported.

RESULTS

Methodological Concept. The new method is based on the following principle. To evaluate LO progress, it must be determined (i) how extensively chemical space around a given AS is covered and (ii) whether analogs display significant potency variations. The tasks in (i) and (ii) can be rationalized as determining *chemical saturation* and *SAR progression*, respectively. Varying potency of structural analogs indicates SAR discontinuity,[9] the apex of which are activity cliffs.[12] In the absence of SAR discontinuity, that is, when all analogs made have comparable potency values, there is no SAR progression. Combining (i) and (ii) enables the characterization of LO progress of ASs. In the presence of extensive chemical space coverage around an AS and the absence of SAR discontinuity, it is unlikely that additional analogs with further increased potency will be available. To assess (i) *chemical saturation*, populations of virtual analogs are generated for a given AS whose distribution serves as an indicator of chemical space coverage around the series. To evaluate (ii) *SAR progression*, potency changes among analogs that have overlapping NBHs with shared virtual analogs are quantified and weighted, as further rationalized below.

Central to the methodology is the definition of analog NBHs in chemical space. The vastness of chemical space makes the enumeration of all conceivable compounds infeasible, even if chemical space is confined to a single AS. For a given series, the space may be mapped using representative virtual analogs. This concept was adapted from a recent chemical saturation analysis,[5] modified and further extended. For quantification of parameters, it must be ensured that the sample of virtual analogs is large enough to yield consistent results. To evaluate compound distributions in AS-centered chemical space and across NBHs of analogs, distances are calculated. Specifically, the distance $d(a,b)$ between two compounds a and b is given by the Euclidean distance between two multidimensional vectors encoding molecular properties after unit-variance scaling on the basis of the virtual analog sample of an AS. A threshold distance, t, must be set to determine the NBH radius of an active analog. The derivation of the corresponding scoring scheme is presented in the following.

Scoring Scheme. On the basis of our methodological concept, a scoring scheme was developed for profiling ASs that addresses three important questions concerning the level of chemical saturation and SAR progression:

1. How extensively does an AS cover chemical space restricted by its core structure?
2. How densely does an AS sample cover chemical space?
3. How strongly do potencies of close analogs vary?

The first two questions relate to chemical saturation, whereas the third one relates to SAR progression. Corresponding parameters can be estimated from the structures and potency values of analogs comprising a given AS. The resulting scores are simple and intuitive. Their formal derivation is presented in the following.

Chemical Saturation. Sampled chemical space and NBH definition form the basis for quantifying a set of key parameters. Given an AS $A = \{a_1, a_2, ..., a_{n_A}\}$ consisting of n_A active analogs and a sampled chemical space $V = \{v_1, v_2, ..., v_{n_V}\}$ of n_V virtual analogs, the sets $V_i = \{v \in V \mid d(v,a_i) \le t\}$ are defined as virtual analogs falling into the NBHs of a_i. The union $V_N = \bigcup_{i=1}^{n_A} V_i$ then gives the set of all virtual analogs that are contained in an NBH of at least one active compound a_i. Let $n_N = |V_N|$ be the number of virtual analogs in NBHs. The coverage

$$C = \frac{n_N}{n_V}$$

with range $[0, 1]$ is the fraction of virtual analogs in the combined NBHs of all active compounds A and used to estimate coverage of relevant chemical space for a given AS.

Virtual analogs v of V_N can be contained in one or more NBHs. The larger the number of NBHs becomes for a given v, the more densely the virtual analog is surrounded by actives. The average number d_{mean} of overlapping NBHs containing a virtual analog in V_N can be determined by summing all NBHs and dividing by the number of virtual analogs in V_N

$$d_{mean} = \frac{1}{n_N}\sum_{i=1}^{n_A} |V_i|$$

This parameter is normalized to the range $[0, 1]$, yielding a density score

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

The extent and density of chemical space coverage around a given AS indicate its chemical saturation. Accordingly, the coverage and density scores can be combined to yield a single saturation score S, which is defined as the harmonic mean of C and D.

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

SAR Progression. If a virtual analog is present in overlapping NBHs of multiple actives, the magnitude of potency variations among active analogs indicates the degree of SAR discontinuity in this chemical sub-space. The parameter $\overline{\Delta}_i$ quantifies the potency range for multiple actives, with NBHs containing the same virtual analog. For $v_i \in V_N$, present in NBHs of m_i actives $\{a_1, ..., a_{m_i}\}$, $\overline{\Delta}_i$ is calculated as the average potency difference over all pairs of active analogs.

$$\overline{\Delta}_i = \frac{2}{m_i(m_i - 1)} \sum_{\substack{j,k=1 \\ j<k}}^{m_i} |\text{pot}(a_j) - \text{pot}(a_k)|$$

Here, pot(a) denotes the potency of compound a on a logarithmic scale.

Parameter P is then calculated as the mean potency range of NBHs over all virtual analogs using a weighting scheme $w_i = 1/m_i$ if $m_i > 1$ and $w_i = 0$ if $m_i = 1$.

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

This score accounts for SAR discontinuity in NBHs of active analogs. The weighting scheme emphasizes SAR discontinuity in smaller numbers of overlapping NBHs associated with a given virtual compound, that is, less-explored regions of chemical space around a given AS where highly potent analogs might preferentially be identified.

Score Combinations. Considering S and P scores in combination makes it possible to estimate the potential of an AS for LO. An AS displaying a high degree of chemical saturation and little SAR progression is thought to have low LO potential, corresponding to a large S and small P score. By contrast, an AS with small S and large P score has high LO potential. A continuum of score combinations can be considered to distinguish between different levels of LO potential.

NBH Radius. A threshold distance t determines the NBH radius of an active analog. It is derived from distance distributions between virtual analogs of a given AS. Virtual analogs outnumber active analogs and serve as an indicator of chemical space coverage. With increasing NBH radii, the likelihood increases that NBHs of active analogs are overlapping and populated with virtual analogs. Using 10 000 virtual analogs per AS, we calculated scores for increasing NBH radii. The dependence of coverage C on NBH radii is monitored in Figure 2, which shows the distribution of C values for all ASs at a given radius. As expected, C increased with increasing NBH radii and the score distributions notably widened. Very similar observations were made for the D and P scores. Initially, we profiled ASs by comparing two NBH radii with percentile values of 1.0 and 5.0, respectively, and obtained similar results. Taking into consideration that publicly available ASs were confined in size, with five of 34 series consisting of more than 100 analogs, we consistently set the threshold of the NBH radius for subsequent calculations to the 1st percentile of the distribution of pairwise distances between virtual analogs.

Virtual Analog Samples. Next, C, D, and P scores were calculated by systematically increasing the number of virtual analogs per series from 500 to 13 000 (in 10 increments of gradually increasing size). Each calculation was repeated 10 times by randomly selecting the respective number of virtual analogs from the source set of each AS. C, D, and P scores remained essentially constant for all ASs when virtual analogs at least on the order of 5000 were used. For consistency with calculations varying NBH radii, the results discussed in the following were obtained for a constant number of 10 000 virtual analogs per AS.

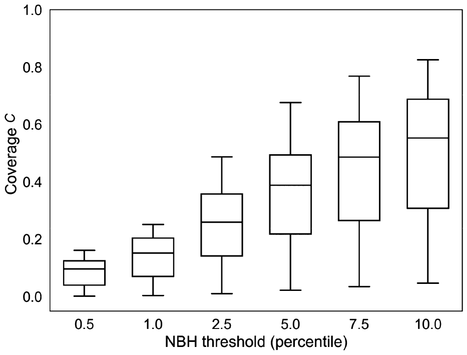


Figure 2. Score dependence on NBH radii. Boxplots report the distribution of coverage C for NBHs with increasing percentile threshold (radius). For each AS, the mean of 10 independent calculations was used. Boxplots report the smallest value (bottom line), first quartile (lower boundary of the box), median value (thick line), third quartile (upper boundary of the box), and largest value (top line).

Comparison of Coverage and Density. Figure 3 compares the distribution of C and D scores for all ASs.

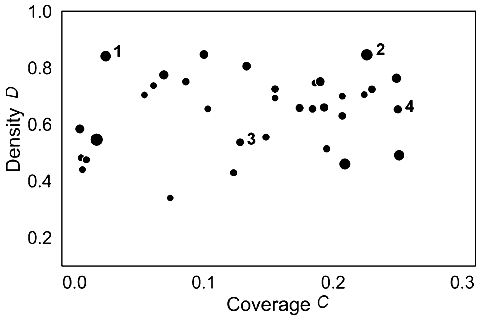


Figure 3. Density vs coverage. The scatter plot compares D and C scores for all ASs. Each AS is represented by a dot that is scaled in size according to the number of analogs it contains (smallest dot, 51 analogs; largest, 166). Exemplary ASs are numbered according to Figure 1.

Given the confined NBH radius used for moderately sized ASs, C scores were distributed over the range [0, 0.25] (see also Figure 2). D scores were widely distributed, mostly over the range [0.4, 0.8]. Two important observations were made. First, C and D scores were uncorrelated and ASs with small C and large D scores and vice versa were detected. Second, C and D scores did not correlate with the size of ASs; both smaller and larger series were found to yield a variety of C or D scores within the observed ranges. Taken together, these observations indicated that score distributions were influenced by chemical characteristics of ASs and that C and D scores captured different aspects of chemical saturation, consistent with their design. Accordingly, the complementary nature of C and D supported the generation of a composite chemical saturation score S.

Profiling of Analog Series. As a central part of our analysis, chemical saturation and SAR progression were explored in context. Figure 4 shows the distribution of progression score P and saturation score S for all ASs. Both scores were broadly distributed. For ASs with S scores greater

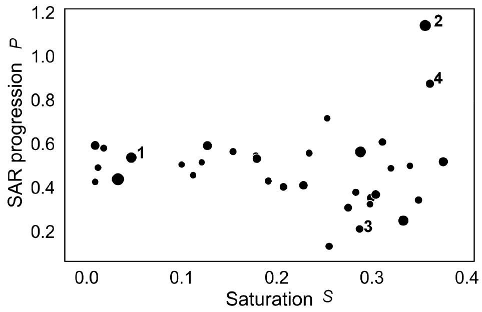


Figure 4. SAR progression vs chemical saturation. The scatter plot compares P and S scores for all ASs. Each AS is represented by a dot that is scaled in size according to the number of analogs. Exemplary ASs are numbered according to Figure 1.

than 0.2, the distribution of P scores widened, covered a large range [0.2, 1.2], and separated ASs from each other. For example, AS 2 in Figure 4 (phosphodiesterase 10A inhibitors, 146 analogs; see also Figure 1) was characterized by one of the highest degrees of chemical saturation in our set of ASs and the overall highest level of SAR progression. S scores of available ASs were generally smaller than 0.4. Therefore, AS 2 was considered to have further LO potential, given its intermediate S and high P score. Accordingly, the generation of additional analogs would be expected to yield potent compounds. In addition, AS 4 (kinase JAK-1 inhibitors, 81 analogs), which was smaller in size than AS 2, had an S score comparable to AS 2 and the overall second highest P score. Therefore, similar conclusions would be drawn for AS 2 and AS 4 and generating additional analogs for AS 4 would be expected to yield further LO progress. However, other ASs with more than 100 analogs yielded some of the smallest S and intermediate P scores. Hence, score differences were not resulting from differences in size, but chemical composition. For example, this was the case for AS 1 (acetyl-CoA carboxylase 2 inhibitors, 126 analogs), which represented one of the series with lowest degree of chemical saturation, but detectable SAR progression. In this case, the generation of additional analogs should be carefully monitored for potential increases in SAR discontinuity. The observed continuum of S/P score combinations made it possible to assign different levels of LO potential to ASs. Notably, ASs with largest S scores within our set and smallest P were expected to have lowest LO potential among the series we compared. An example for such ASs was provided by AS 3 (dopamine D2 receptor antagonists, 69 analogs), which had low priority for further exploration. By contrast, other ASs with S scores around 0.2 and P scores of 0.6 or greater had higher LO potential. Overall, a variety of score combinations were observed for ASs of varying size, which clearly differentiated ASs and made it possible to prioritize series for further optimization efforts. For example, Figure 5 shows exemplary analogs of a series of ATPase inhibitors and virtual analogs falling into their NBHs. This AS was among the smallest we profiled (53 analogs) and had one of the lowest S scores (0.11). However, despite low chemical saturation, the AS reached an intermediate P score (0.46) within our set and should thus be considered to have further optimization potential.

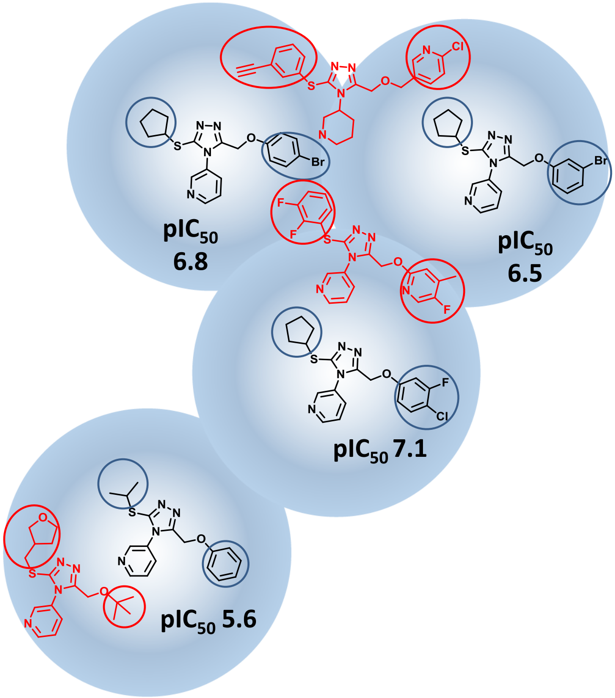


Figure 5. Progressing AS. Shown are exemplary analogs belonging to a series of endoplasmic reticulum ATPase inhibitors (black) and their NBHs (light blue) and virtual analogs (red) falling into the NBHs. Substitution sites are encircled.

DISCUSSION AND CONCLUSIONS

In medicinal chemistry, LO is a largely subjective process that is difficult to rationalize and formalize. Generally applicable LO criteria or protocols are not available, and it is very difficult to understand when optimization efforts become unlikely to further advance an AS. Consequently, approaches that provide guidance during LO and decision support are highly desirable. We have introduced a computational method to evaluate LO progress. The approach combines the assessment of chemical saturation and SAR progression and makes it possible to characterize and differentiate ASs. Sets of virtual analogs generated for individual ASs aid in defining NBHs of active analogs and serve as an indicator of chemical space coverage. Furthermore, detectable SAR discontinuity within overlapping NBHs is regarded as a prerequisite for obtaining increasingly potent compounds. The combined analysis of chemical saturation and SAR progression is translated into a scoring scheme for profiling and prioritizing ASs. High S scores approaching 1 and low P score approaching 0 indicate that ASs are exhausted and provide a criterion for discontinuation beyond subjective assessment. On the other hand, opposite scores are indicative of optimization potential.

The approach is expected to be influenced by the choice of descriptors and chemical reference spaces as well as virtual compound populations. A previous study investigated metrics for the assessment of chemical saturation of ASs under varying conditions.[6] This included investigating the influence of chemical space representations based on different descriptor sets and alternative approaches to virtual compound generation for sampling chemical space. The saturation assessment for different sets of ASs remained essentially stable under these varying conditions. For example, when a 14-

dimensional reference space was used instead of the seven chemically intuitive descriptors used herein, comparable results were obtained.[6] Similarly, only little changes in chemical saturation were observed when virtual analog populations were generated using different R-group sets and design strategies. In this study, we used R-group sets that were systematically extracted from bioactive compounds in ChEMBL to enumerate virtual analogs on the basis of ASB scaffolds representing series, without additional filtering of analogs. Importantly, as introduced herein, the methodology can be tested with many different descriptors and candidate compounds, depending on preferences and project specifics. There are no intrinsic limitations. This also applies to distance measures for defining NBHs in chemical reference space. After initially exploring various alternatives, including Tanimoto distances in fingerprint spaces, our preference is the straightforward calculation of Euclidian distances on the basis of numerical descriptors, but more complex measures can certainly be explored. Essentially, calculation parameters can be adjusted at will and virtual compounds from different sources be used. We also note that the approach is applicable to ASs of any composition and not intrinsically limited by the number of substitution sites. For our study, ASs have been systematically extracted from public domain compounds originating from the medicinal chemistry literature. The ASs studied herein represented the largest we have been able to identify via automated AS extraction, requiring the availability of reliable activity measurements, a prerequisite for meaningful profiling analysis. These ASs contained from one to three substitution sites. In the practice of medicinal chemistry, ASs with more than four or five site substitution sites are rare. However, ASs with three or more sites can be readily profiled.

Given the complexity of LO, it is anticipated that the new methodology will be of interest to practicing chemists. For computational medicinal chemists, the scoring scheme is straightforward to implement and scores are easy to calculate.

Although we are currently limited to studying the relatively small number of sizable ASs that are available in the public domain, it will be exciting to see applications of the methodology to mid- and late-stage series from drug discovery projects.

EXPERIMENTAL SECTION

Chemical Reference Space. For profiling of ASs, a 7-dimensional chemical reference space was generated using chemically intuitive descriptors accounting for molecular properties known to be relevant for ligand−target interactions, as previously used for assessing chemical saturation.[5] These descriptors included the number of hydrogen bond donors, acceptors, and rotatable bonds; logarithmic octanol/water partition coefficient and aqueous solubility, topological polar surface area, and molecular weight were calculated using RDKit[13] and, in the case of aqueous solubility, using a freely available custom implementation based on the ESOL method.[14]

Analog Series. ASs with single and multiple substitution sites, more than 50 compounds, and available high-confidence activity data were systematically extracted from ChEMBL (release 23)[15] using a computational analog selection method[16] based on the matched molecular pair (MMP) formalism.[17] An MMP is defined as a pair of compounds that are distinguished by a chemical change at a single site.[17]

A total of 34 ASs were obtained, including 15 ASs with single and 19 with multiple (two to three) substitution sites. The 34 ASs consisted of 51−166 compounds (five series contained more than 100 analogs). Each AS was active against a unique target. These compound data sets represented the largest and most diverse collection of ASs from medicinal chemistry sources that we were able to identify in the public domain. Figure 1 shows exemplary ASs. Each AS is represented by its AS-based (ASB) scaffold[18] covering all substitution sites (reminiscent of a Markush structure). For all ASs and virtual analogs of AS 4 in Figure 1, molecular formula strings are provided as Supporting Information.

Virtual Analogs. From each AS, the ASB scaffold[18] was isolated. From ChEMBL (release 23), a total of 14 026 unique R-groups were extracted by systematically calculating MMPs with size-restricted chemical changes[19] from bioactive compounds. On the basis of each ASB scaffold representing an AS, virtual analogs were enumerated by systematically adding R-groups to substitution sites. The calculations were carried out with the aid of the OpenEye toolkit.[20]

For ASs with single substitution sites, 13 965 to 14 012 virtual analogs were generated (i.e., one per unique R-group). The number of virtual analogs per AS was in each case smaller than 14 026 because existing analogs were frequently reproduced using the pool of R-groups extracted from ChEMBL. For each AS with multiple substitution sites, 140 260 distinct virtual analogs were enumerated by randomly selecting R-groups for each substitution site from the ChEMBL pool. Virtual analogs generated for each AS provided source sets for score calculations.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jmedchem.8b01626.

Molecular formula strings of ASs and virtual analogs for an exemplary AS are provided (CSV)

AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de. Phone: +49-228-7369-100. Fax: +49-228-7369-101.

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank OpenEye Scientific Software for an academic license and Pat Walters of Relay Therapeutics for the implementation of the ESOL method made available on GitHub (https://github.com/PatWalters/solubility). D.Y. is a fellow of the Jürgen Manchot Stiftung, Düsseldorf, Germany.

ABBREVIATIONS

AS, analog series; ASB, analog series based; LO, lead optimization; MMP, matched molecular pair; SAR, structure−activity relationship.

REFERENCES

(1) *The Practice of Medicinal Chemistry*, 3rd ed,; Wermuth, C. G., Ed.; Academic Press-Elsevier: Burlington, San Diego, USA, London, U.K., 2008.

(2) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going to? *J. Med. Chem.* **2014**, *57*, 4977−5010.

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

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

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

(6) Yonchev, D.; Vogt, M.; Stumpfe, D.; Kunimoto, R.; Miyao, T.; Bajorath, J. Computational Assessment of Chemical Saturation of Analog Series under Varying Conditions. *ACS Omega* **2018**, *3*, 15799−15808.

(7) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

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

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

(10) 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.* **2015**, *59*, 4235−4244.

(11) Maynard, A. T.; Roberts, C. D. Quantifying, Visualizing, and Monitoring Lead Optimization. *J. Med. Chem.* **2015**, *59*, 4189−4201.

(12) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.

(13) RDKit: Cheminformatics and Machine Learning Software, 2013. http://www.rdkit.org (accessed July 1, 2018).

(14) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Model.* **2004**, *44*, 1000−1005.

(15) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.

(16) Stumpfe, D.; Dimova, D.; Bajorath, J. Computational Method for Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. *J. Med. Chem.* **2016**, *59*, 7667−7676.

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

(18) Dimova, D.; Bajorath, J. Computational Design of New Molecular Scaffolds for Medicinal Chemistry, Part II: Generalization of Analog Series-Based Scaffolds. *Future Sci. OA* **2018**, *4*, FSO287.

(19) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.

(20) *OEChem TK*, version 1.7.7; OpenEye Scientific Software, Inc.: Santa Fe, NM, USA, 2012.

DOI: 10.1021/acs.jmedchem.8b01626

Summary

Herein, chemical saturation and SAR progression have been combined to a scoring scheme for AS profiling and prioritization. Newly introduced methodological modifications have been extensively tested in order to ensure robustness and identify the optimal parameter setting that provides the best means for differentiating between series. After plotting ASs based on their characteristic score combinations, no correlation has been observed, and thus their LO status has been intuitively evaluated. High chemical saturation is generally an indicator for more developed series but does not necessarily correlate with the number of synthesized analogs. At the same time, high SAR progression (resulting into detectable SAR discontinuity) can be rationalized as promising from a medicinal chemistry point of view, as more pronounced potency fluctuations are more likely to result in faster convergence in the search for a highly potent candidate.

The successful application of the herein developed scoring system served as a proof-of-principle for combination of chemical saturation and SAR progression as diagnostic measures for estimating the LO potential of individual ASs. These findings have been used as a rationale for the development of the COMO methodology presented in the following chapter.

Chapter 5

Compound Optimization Monitor (COMO) Method for Computational Evaluation of Progress in Medicinal Chemistry Projects

Introduction

The method introduced in the previous chapter, albeit robust in its discriminative power, relies entirely on the chemical NBH principle for score calculations. However, in order to gain a more holistic view of AS characteristics, NBH-independent scoring can be considered as an additional complementary component. This may be of particular use when investigating certain score combinations more profoundly. For example, high chemical saturation in combination with low SAR progression may present a potential termination criterion for a series if no sufficient degree of activity is yet present among analogs. At the same time, it may be a desired characteristic in cases where predictable (continuous) SARs are required for maintaining already achieved high potency levels and avoiding potential pitfalls while optimizing other compound properties. In light of such cases, estimating the general trajectory of potency trends within a series is likely to provide an additional layer of information for medicinal chemists. Furthermore, given the fact that biological activity is not the only compound property being ameliorated during LO, progression of physicochemical properties needs to be monitored in addition.

In this chapter, the Compound Optimization MOnitor (COMO) is introduced building upon the efforts and gathered knowledge from the previous approaches. COMO is developed as a diagnostic method for AS categorization by augmenting the already established scoring system with additional scoring components for SAR heterogeneity and multiple physicochemical properties. Furthermore, the quality of the generated VAs is improved by introducing a reaction-based scaffold enumeration strategy and limitations in molecular size to ensure better synthetic tractability and adequate representation of series-relevant chemical space. In this study, a new set of mid- to large-sized ASs is obtained from ChEMBL and subjected to systematic profiling.

My main contribution to this work was the new VA design strategy, the development and benchmarking of the scoring extensions, and the subsequent analysis of the results.

Reproduced from "D. Yonchev, M. Vogt, J. Bajorath. Compound optimization monitor (COMO) method for computational evaluation of progress in medicinal chemistry projects. *Future Drug Discov.* **2019**, *1*, FDD15" with permission of Future Science Group.

Future Drug Discovery

Methodology

For reprint orders, please contact: reprints@future-science.com

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

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

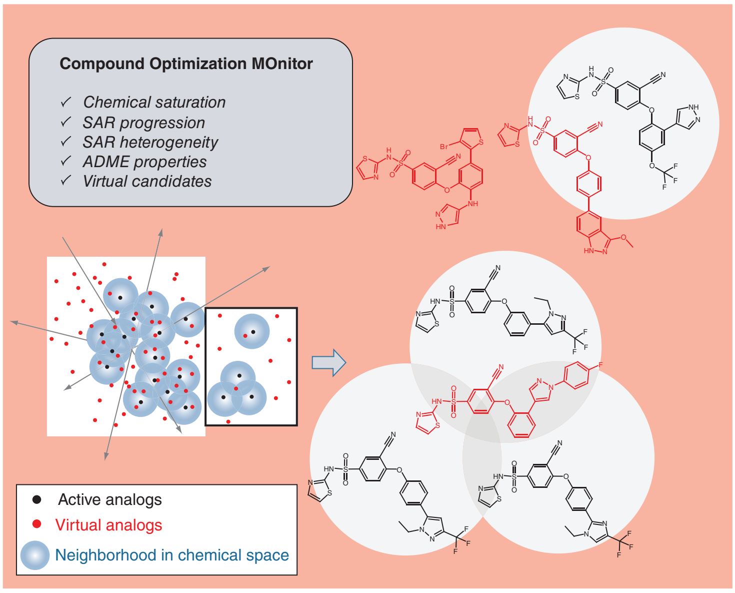
[1]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53113 Bonn, Germany

*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

10.4155/fdd-2019-0016 © 2019 Jürgen Bajorath eISSN 2631-3316

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 ($\mathrm{NBH_{O\_VA}}$) relative to the number of VAs falling into NBHs:

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

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:

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

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 ($\mathrm{pot_j}$ and $\mathrm{pot_k}$ 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 ($\mathrm{pot_i}$ is the log potency of compound i and $\mathrm{pot_{AS}}$ 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

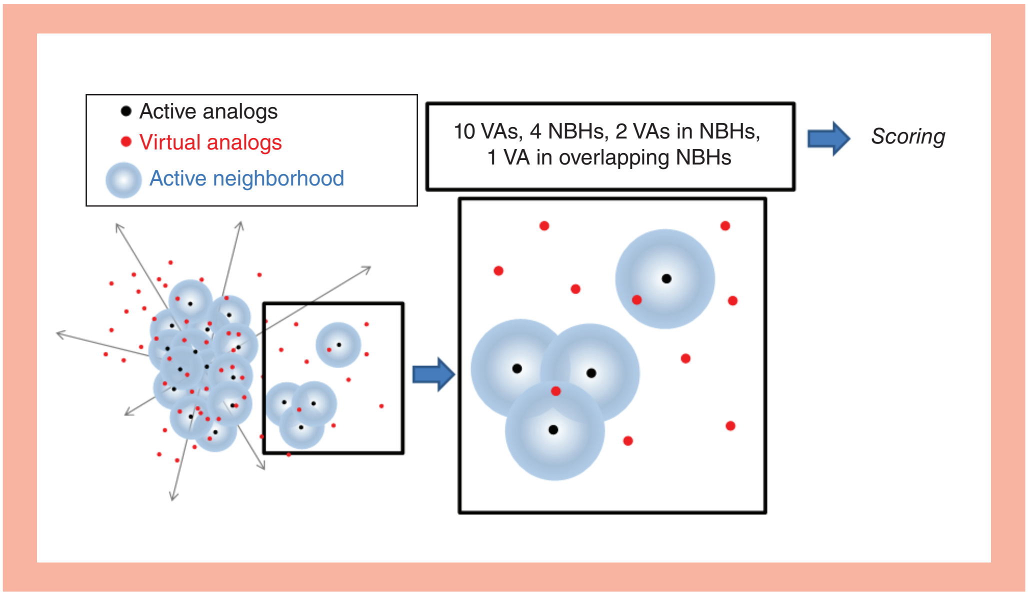


Figure 1. Methodological concept of compound optimization monitor. Shown is an analog series (black dots) with compound NBHs (gray spheres) and VAs (red dots) in n-dimensional chemical reference space. A small section of populated chemical space is enlarged, which contains four active analogs, their 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.

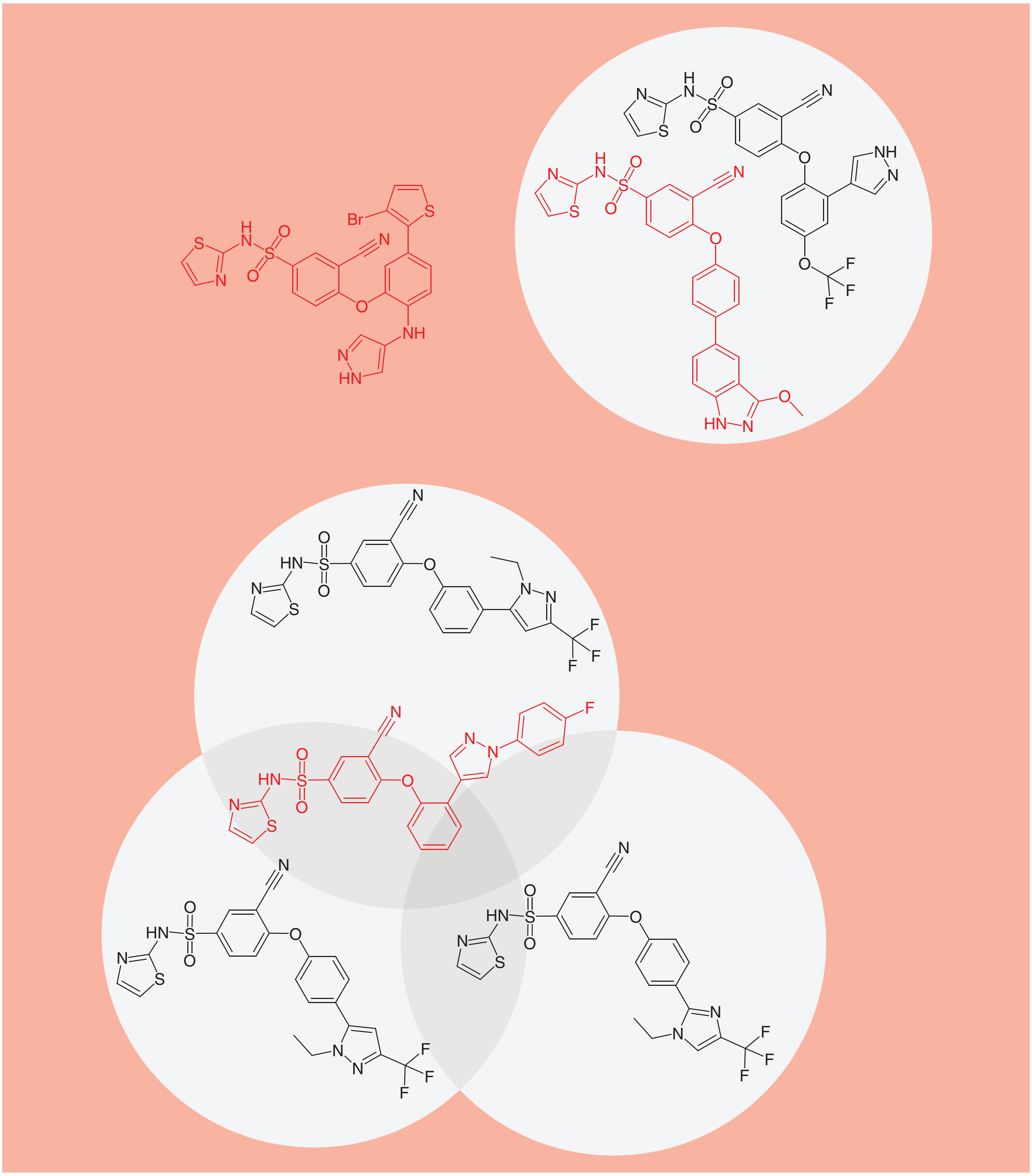


Figure 2. Exemplary active analogs, neighborhoods and virtual analogs. For an analog series of sodium channel protein type IX α subunit ligands, exemplary compounds, their neighborhoods and virtual analogs are shown according to Figure 1.

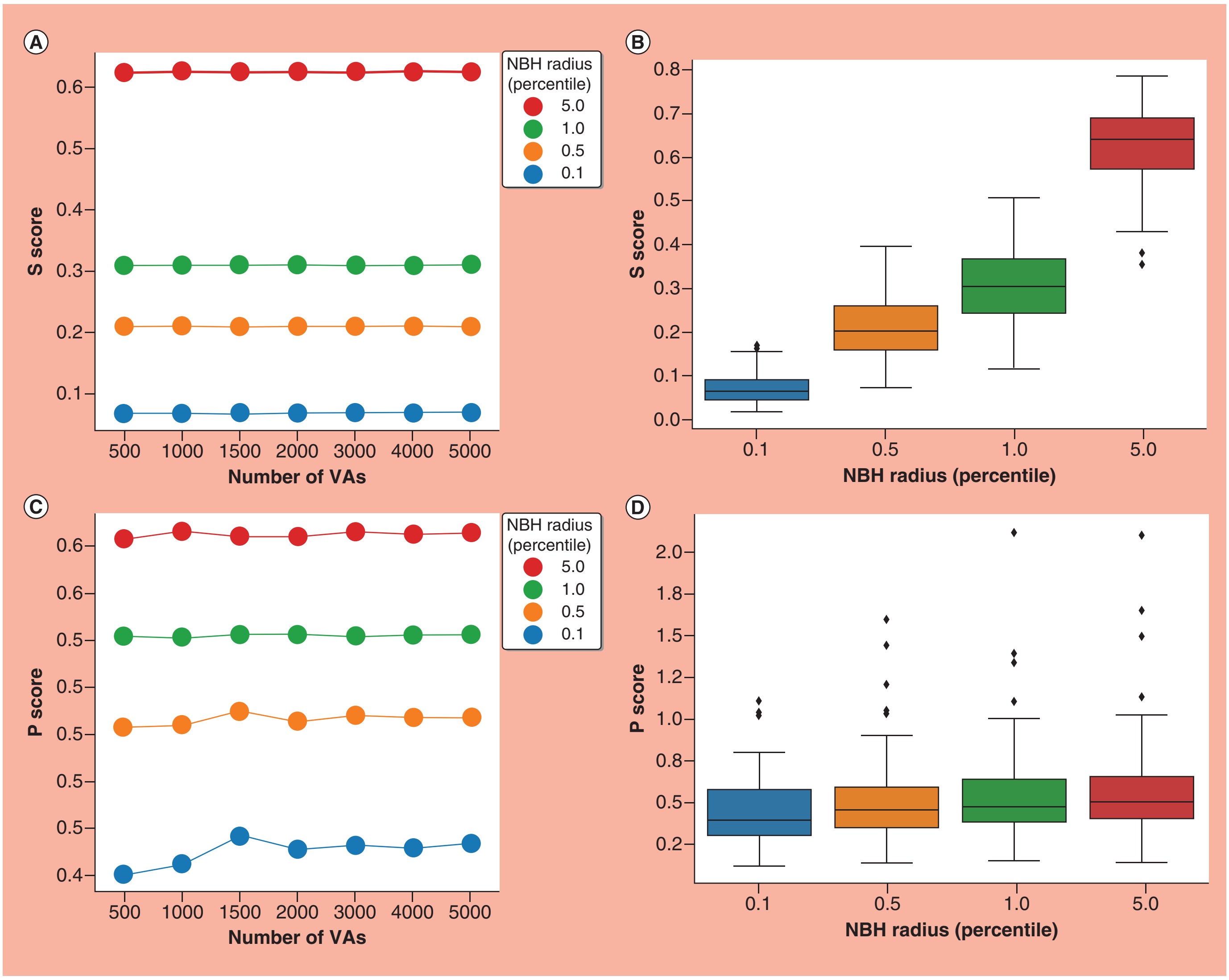


Figure 3. COMO calculation parameters. (A) Reports mean S scores for the set of 72 analog series as a function of virtual analog populations of increasing size over increasing neighborhood radii. **(B)** Shows box plots representing the distribution of S scores across all analog series for increasing 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

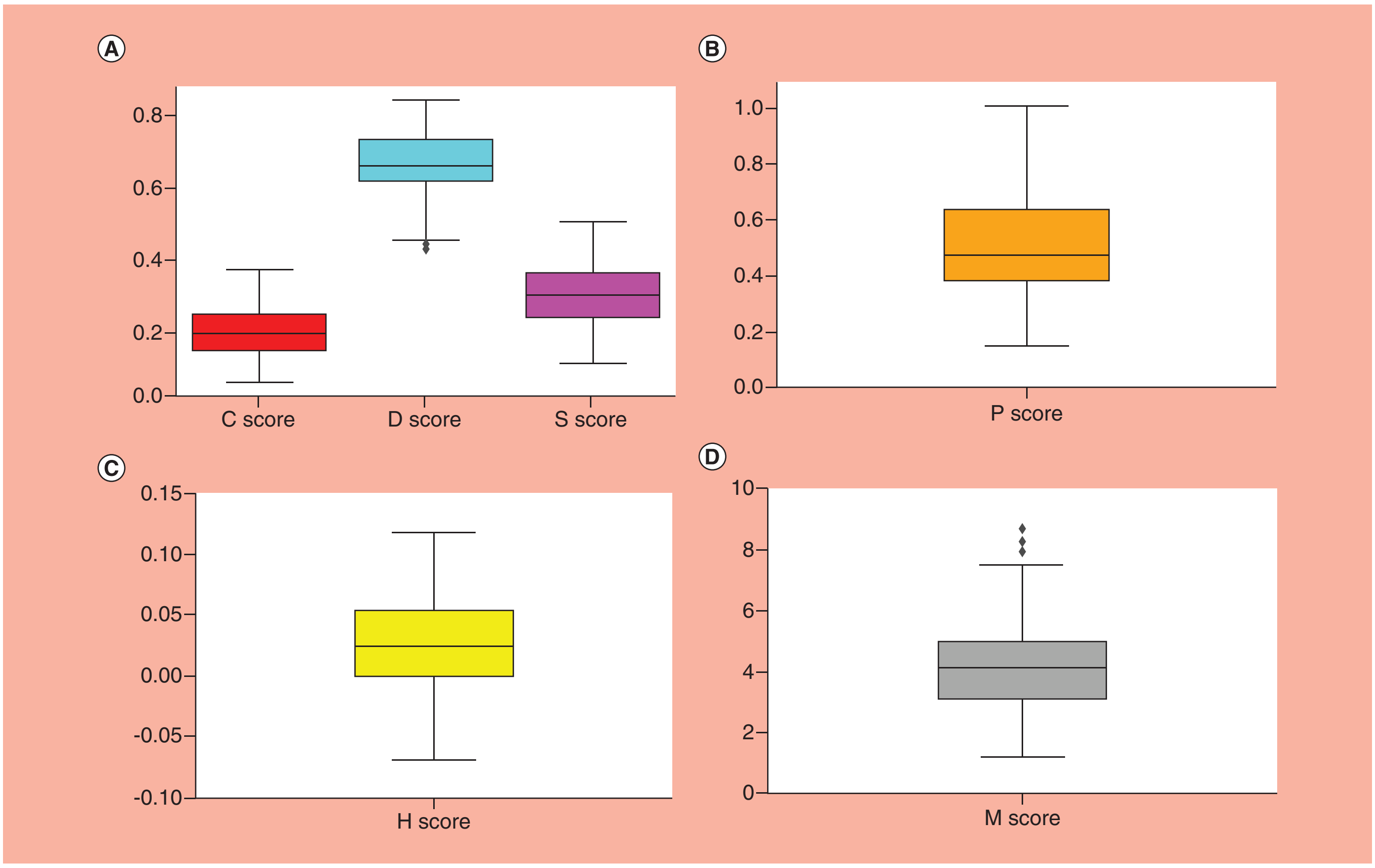


Figure 4. Score distributions. Shown are box plots representing the distributions of the six compound optimization monitor scores for all 72 analog series calculated using a constant neighborhood radius (first percentile) and 3000 virtual analogs.

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.

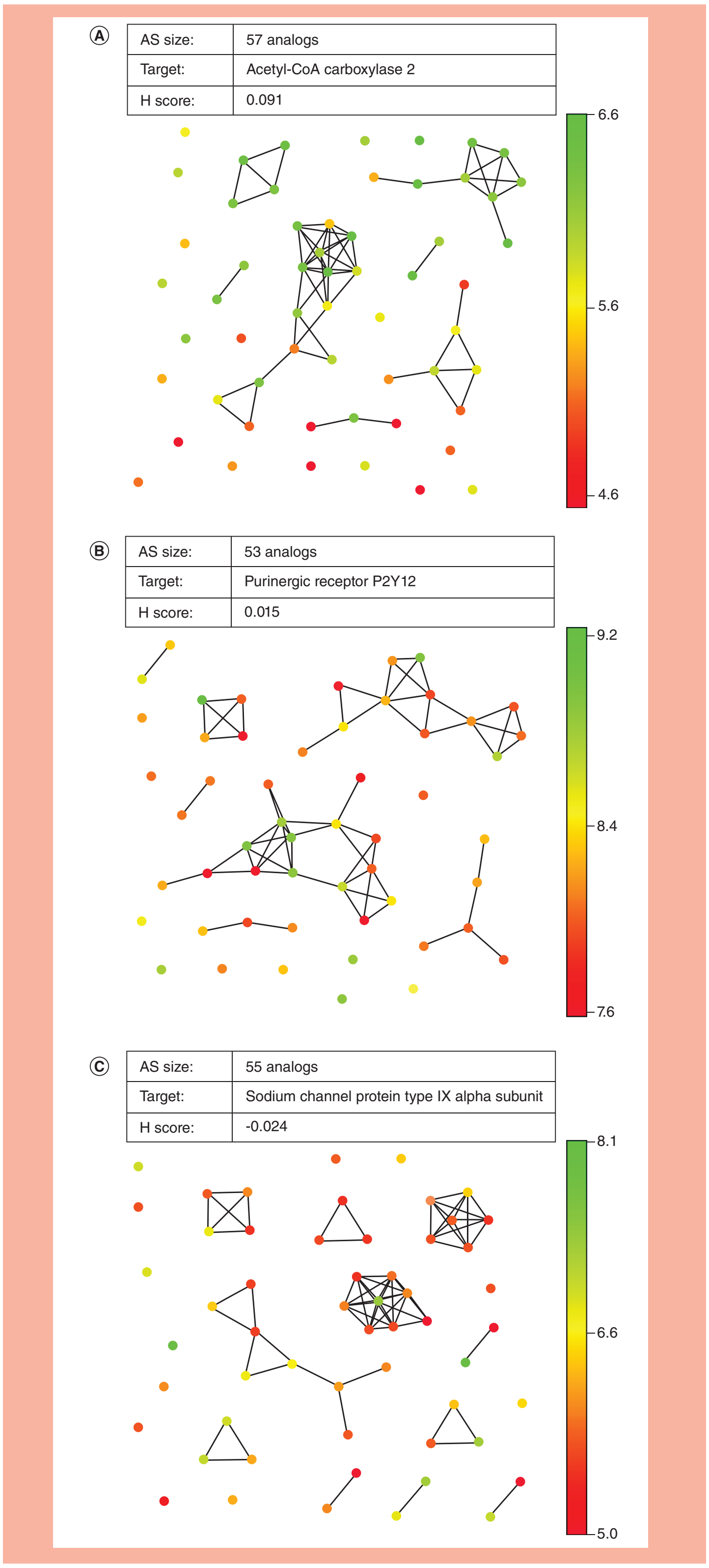


Figure 5. Neighborhood overlap-based analog networks. For different analog series, network representations are shown in which nodes represent compounds. Nodes are color coded by logarithmic compound potency (K_i or IC_{50}) values using a continuous color spectrum as indicated. In addition, edges between 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].

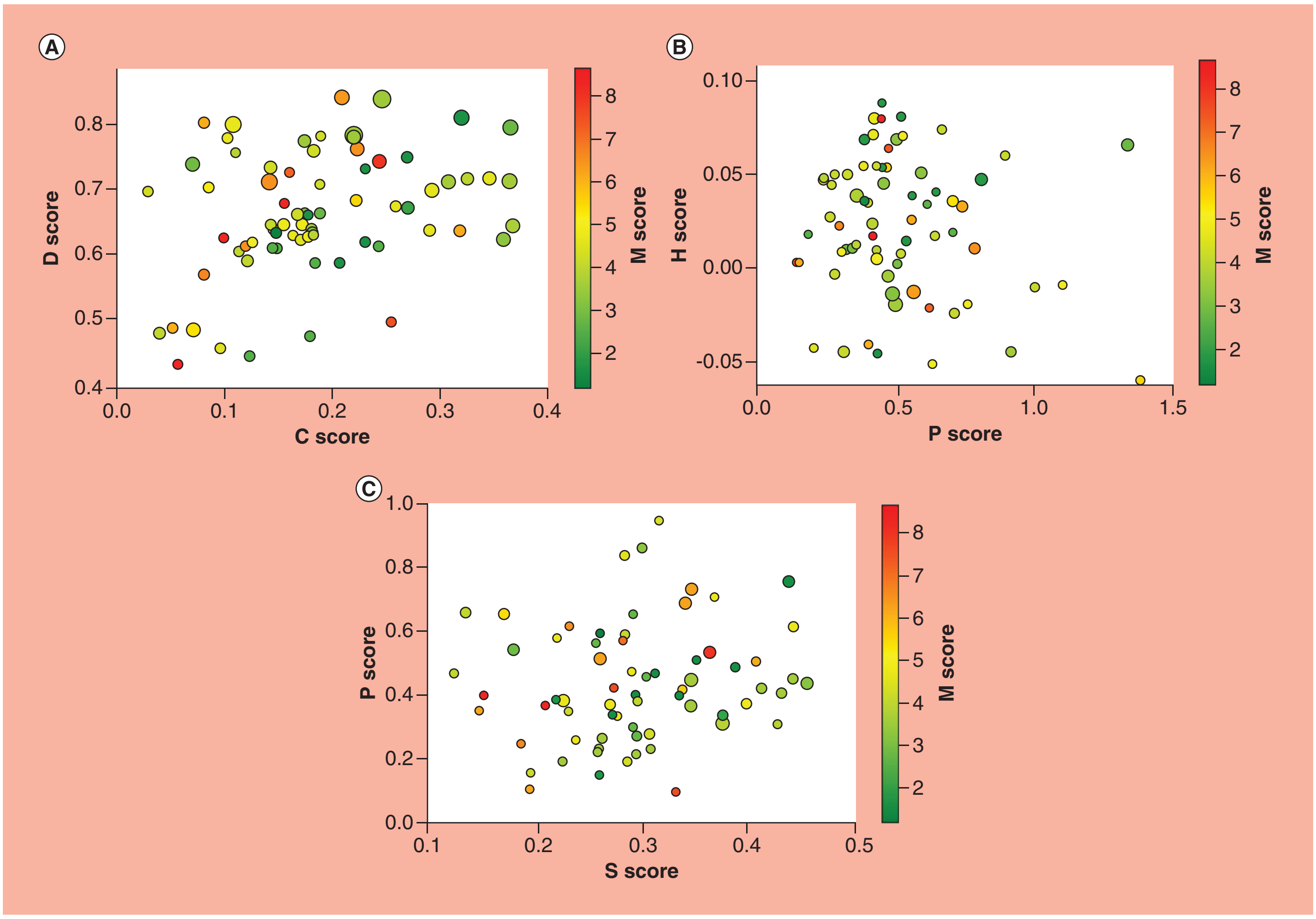


Figure 6. Score comparison. Scatter plots compare the distributions of different COMO scores for the set of 72 AS. Each dot reprscoresesents an AS. Dots are scaled in size according to the number of compounds per AS and color coded according to M scores using a continuous color spectrum as indicated. **(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.

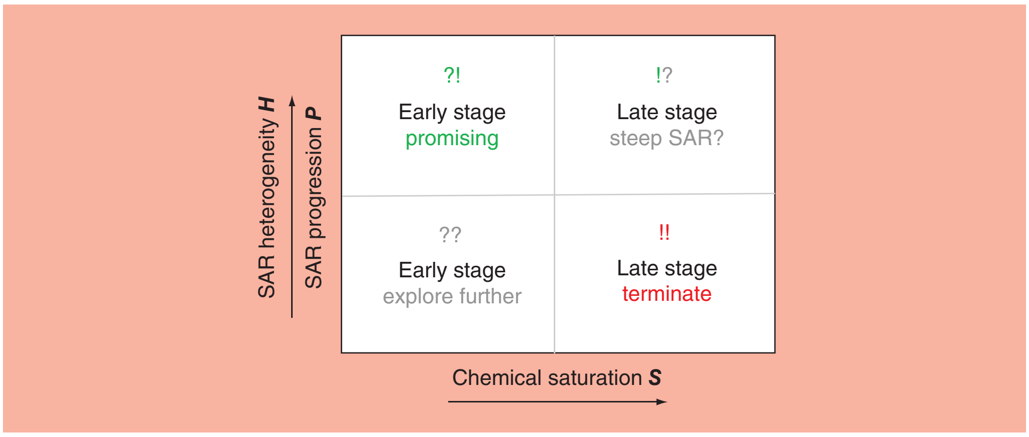


Figure 7. Interpretation of score combinations. The schematic representation illustrates combinations of COMO scores of different magnitude and their interpretation. Characteristic score combinations are used to assign analog series to different lead optimization stages.
COMO: Compound optimization monitor.

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.

Open access

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).

Summary

In this study, the COMO method has been successfully applied for detailed analysis of AS LO profiles. The herein presented combination of four NBH-dependent and two NBH-independent scores is designed to address the multi-dimensional nature of LO and provides complementary points of view on specific AS characteristics. The utility of the newly introduced NBH-independent SAR heterogeneity score for assessing the global direction of potency trends within a series has been corroborated by network analysis and the multi-property score has been shown to provide intuitive orientation in the progression of physicochemical properties. Moreover, it has been shown that all COMO parameters and scores are easily adjustable to the specific needs of individual LO projects and to the endpoints of comparison between ASs (provided that experimentally measured data for this are available). Importantly, synthetic accessibility of the VAs used in this study has been significantly improved, which has laid the foundations for the next step of bridging between their utility as indicators for chemical space sampling and potential synthetic candidates for AS expansion.

The latter aspect is explored in the next chapter, where the diagnostic capabilities of COMO are extended with NBH-oriented prospective VA design.

Chapter 6

Integrating Computational Lead Optimization Diagnostics with Analog Design and Candidate Selectionn

Introduction

The diagnostic capability of COMO enables evaluation of LO progress, however in its support for prospective decision-making, it is limited to indicating general project trajectories (e.g. series continuation/termination). As medicinal chemists face the daily challenge of which compound to synthesize next, an extension of the COMO methodology with a component for candidate prioritization is proposed to complement the discussed scoring scheme. The role of VA populations, as described until now, has been purely diagnostic i.e. they have been solely used as indicators for delineating and sampling AS-centric chemical space. However, they can be also viewed as a pool of potential new molecules for synthesis and testing. Thereby, important prerequisites for selection of suitable candidates are synthetic accessibility and predictable biological activity, which are expected to reduce experimental efforts.

Herein, a procedure for identifying promising candidates is integrated to the COMO methodology. First, standard series-based QSAR models for prospective potency prediction of enumerated synthetically tractable VAs are explored. Furthermore, as an alternative design strategy, VAs are automatically generated and prioritized on the basis of MMP-based NBHs specifically assembled for FW-type analysis. Importantly, these NBHs are not equivalent to the distance-based NBHs utilized in the COMO formalism. Finally, in addition to the existing diagnostic scoring, LO progress is further rationalized based on the extent of exploration of such FW NBHs. The procedure is applied to large ASs comprising more than 100 compounds (newly extracted from ChEMBL) that are likely to exhibit a more advanced LO profile.

Reproduced from "D. Yonchev, J. Bajorath. Integrating computational lead optimization diagnostics with analog design and candidate selection. *Future Sci. OA* **2020**, *6*, FSO451" with permission of Future Science Group.

Future Science OA

Integrating computational lead optimization diagnostics with analog design and candidate selection

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

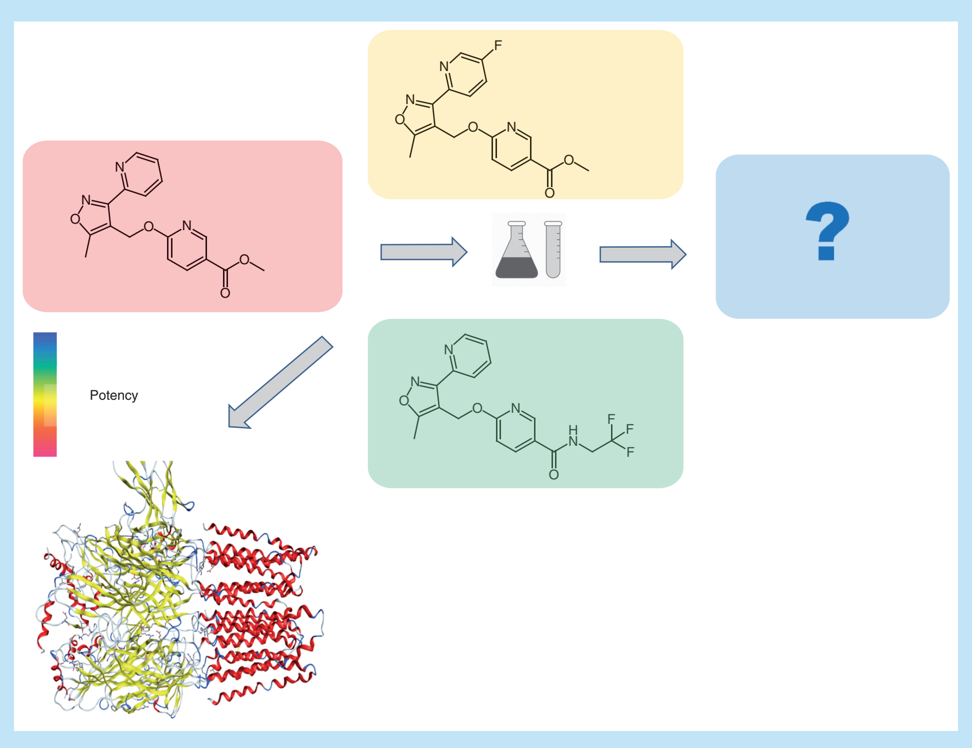
[1]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53113 Bonn, Germany

*Author for correspondence: Tel.: +49 228 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.

Graphical abstract:



First draft submitted: 15 November 2019; Accepted for publication: 20 December 2019; Published online: 24 January 2020

Keywords: analog design • candidate prioritization • computational diagnostics • decision support • Free-Wilson analysis • lead optimization • medicinal chemistry • potency prediction

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 chemcial 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:

$$C = \frac{n_N}{n_V} \qquad \text{(Eq. 1)}$$

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_{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:

$$d_{mean} = \frac{NBH_{O\_VA}}{n_N} \qquad \text{(Eq. 2)}$$

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

$$D = 1 - \frac{1}{d_{mean}} \qquad \text{(Eq. 3)}$$

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:

$$S = \frac{2CD}{C + D} \qquad \text{(Eq. 4)}$$

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:

$$\overline{\Delta_i} = \frac{2}{m_i(m_i-1)} \Sigma^{m_i}_{\substack{j,\,k\,=\,1 \\ j\,<\,k}} |pot_j - pot_k| \qquad \text{(Eq. 5)}$$

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 $m_i = 1$:

$$P = \frac{1}{\Sigma_{i=1}^{n_N} w_i} \Sigma_{i=1}^{n_N} w_i \overline{\Delta}_i \qquad \text{(Eq. 6)}$$

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.

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 kinase 12	1908389	2	111	44	1844
9	5-lipoxygenase activating protein	4550	3	259	162	5204
10	5-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	5-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:

$$min_w ||X_w - y||^2 + \alpha||w||^2 \qquad \text{(Eq. 7)}$$

Here, X_w is the estimated target value, y the true target value, w the weighting coefficient and α 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:

$$f(x) = \Sigma_i(\alpha - \alpha_{i^*})K(x_i, x) + b \qquad \text{(Eq. 8)}$$

Here, α and α_{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 *ϵ-insensitive tube* [20,21]. The ϵ 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 K_i 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 (R_1 and R_2), respectively. By contrast, analogs B and C do not form an MMP because they differ at both substitution sites. Analogs A and B share the same substituent at R_2 while the pyridine ring at R_1 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 R_1 while the methyl ester function at R_2 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 pK_i 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 $pK_i = 9.3$ was only slightly lower than the experimentally observed potency of $pK_i = 9.5$, 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:

Given an AS core structure with indexed substitution sites, all EAs and their site-specific substituents are collected.

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.

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.

For each FW NBH, the direction of the MMP-defining substituent exchanges is determined according to Figure 1 (i.e., A to B, A to C) and the corresponding newly introduced substituents are recorded.

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.

Unique FW VAs associated with one or more FW NBHs are retained.

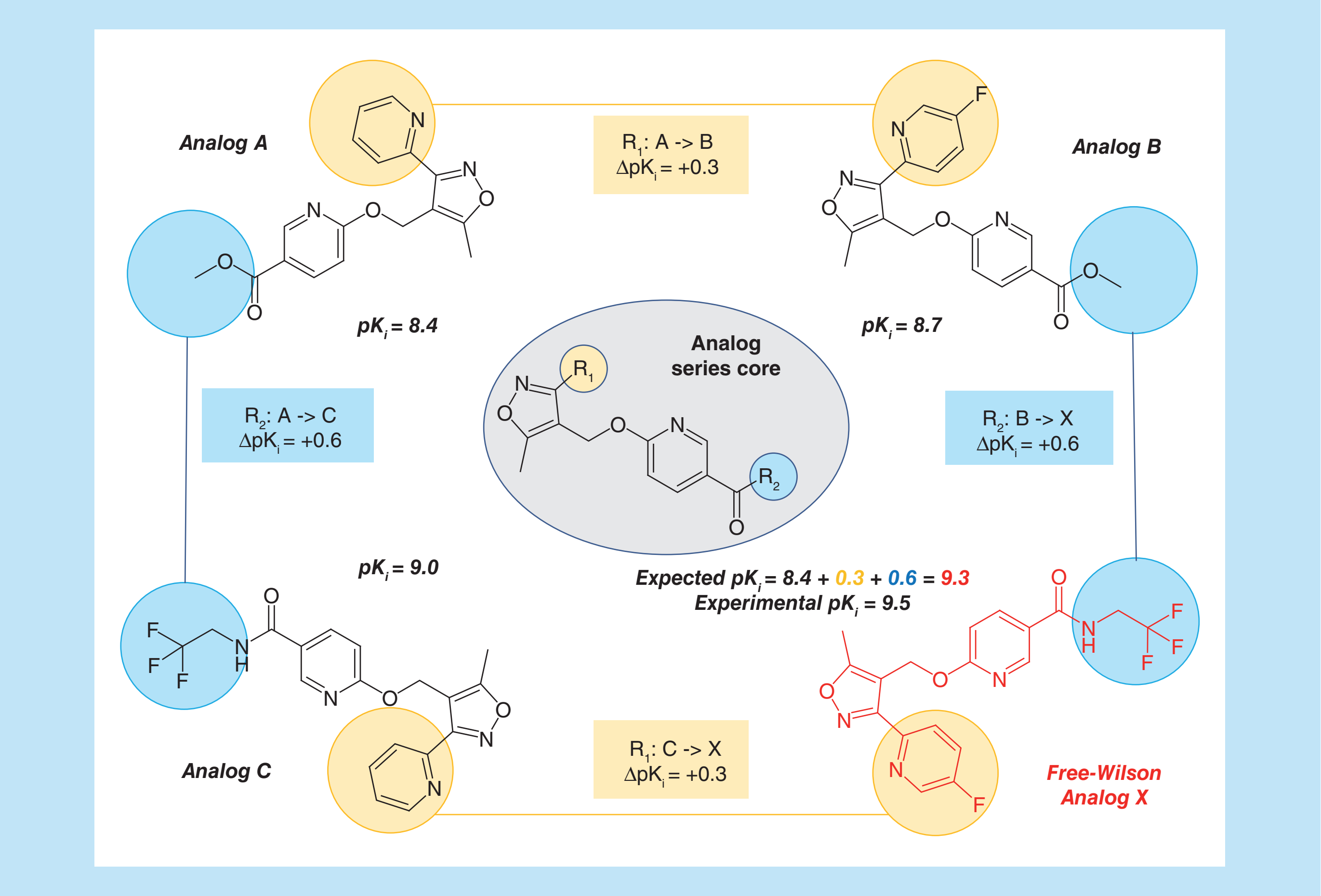


Figure 1. Principles of Free-Wilson analysis. Shown are four analogs from the same AS that are active against GABA receptor alpha-5 subunit (AS 5; ChEMBL target ID 5112). For each compound, its logarithmic experimental potency (pK_i) value is reported. In addition, the core structure of the AS is depicted 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 ΔpK_i 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 ϵ values was optimized in combination with the Tanimoto kernel [30]. For RR, seven different α 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:

$$R^2 = 1 - \frac{\Sigma_i (y_i - f_i)^2}{\Sigma_i (y_i - \overline{y})^2} \qquad \text{(Eq. 9)}$$

Here, y_i is the true value of instance i, f_i the predicted value of instance i, and $\overline{y}$ 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,

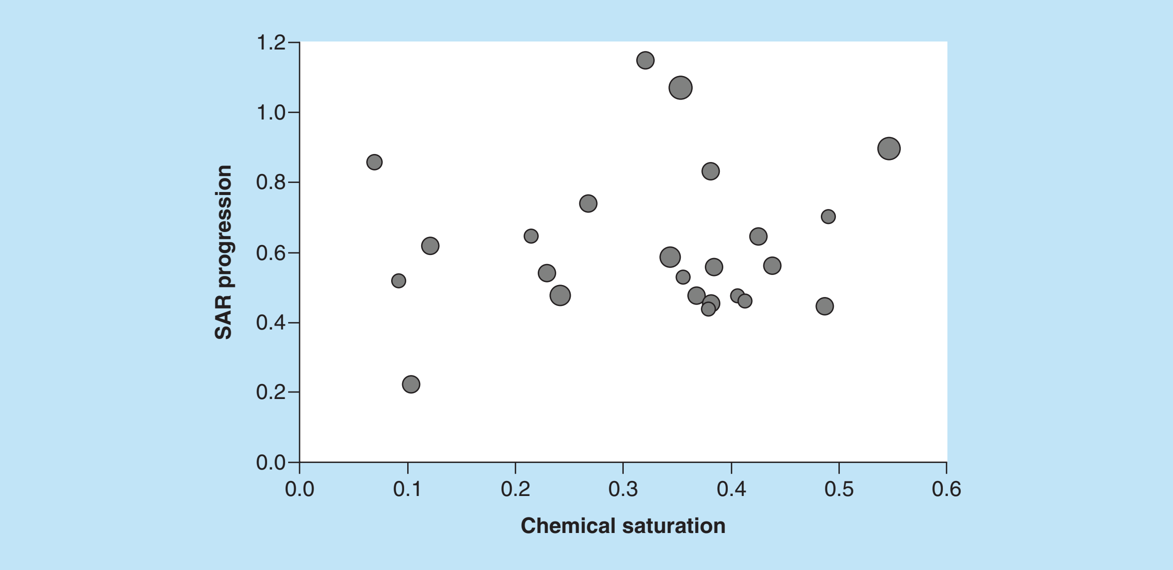


Figure 2. Compound optimization monitor diagnostic scores. The scatter plot compares COMO chemical saturation **(S)** and SAR progression **(P)** scores for ASs from ChEMBL (version 25). Each dot represents a series and is scaled in size according to the number of analogs. Different combinations of S and P scores distinguish ASs at different LO stages.
AS: Analog series; COMO: Compound optimization monitor; LO: Lead optimization; SAR: Structure–activity relationship.

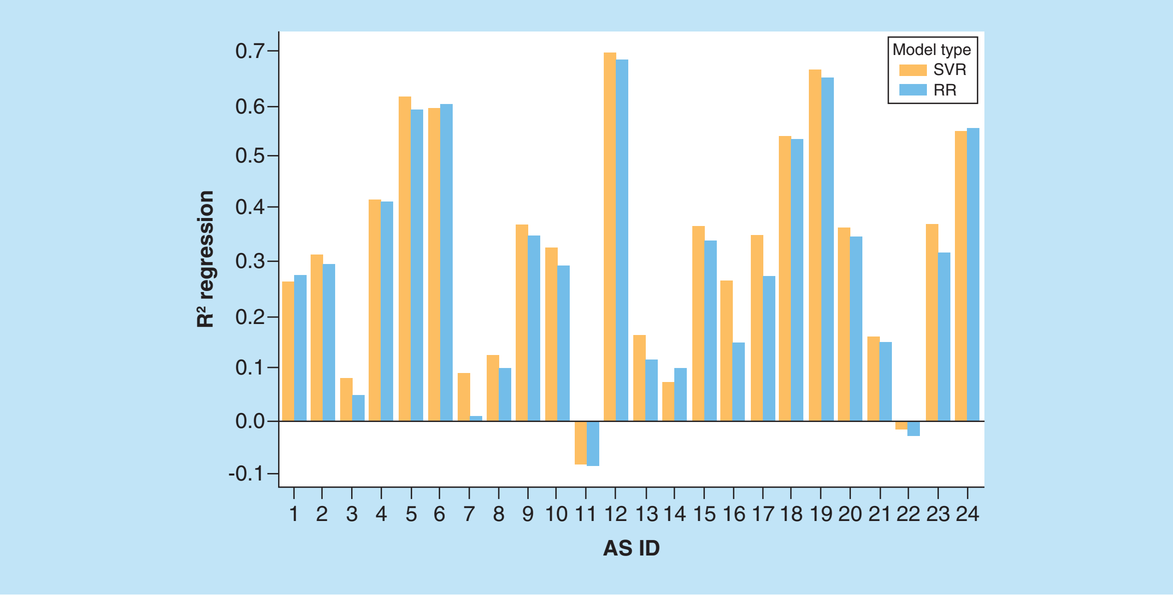


Figure 3. Performance of regression models. On the vertical axis, mean coefficients of determination (R^2) for regression models are reported. The horizontal axis lists ASs using their IDs according to Table 1. For each AS, two R^2 values are given for RR (blue) and SVR (orange) models.
AS: Analog series; ID: Identificaton; 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

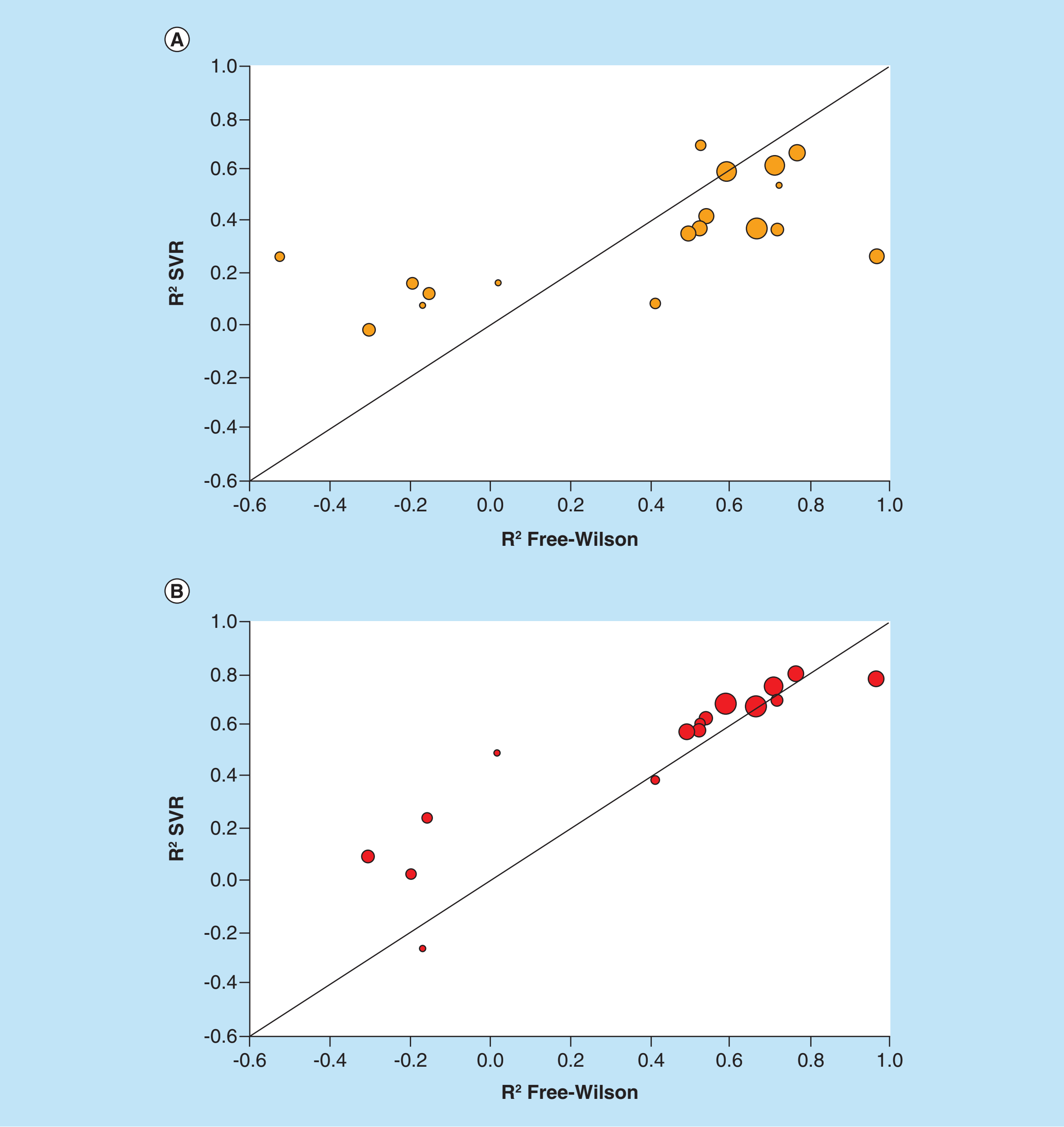


Figure 4. Comparison of Free-Wilson and support vector regression predictions. Scatter plots compare R^2 values of FW and SVR predictions for individual ASs. Each dot represents an AS that is scaled in size according to the number of FW EAs. The diagonal corresponds to perfect correlation between calculated 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.

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 R^2 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 R^2 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

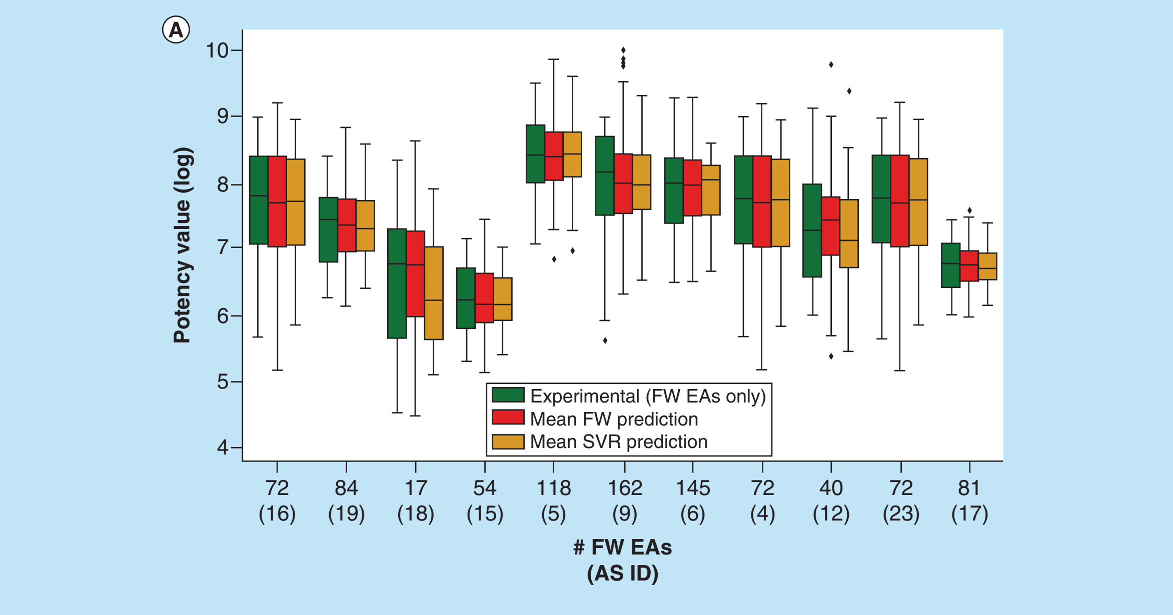


Figure 5. Experimental and predicted potency values. (A) Box plots compare experimental and predicted potency value distributions. Each triplet represents one of 11 ASs yielding predictive FW and SVR models. The y-axis reports log. potency values and the x-axis the number of FW EAs. Numbers in parentheses 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,

(B)

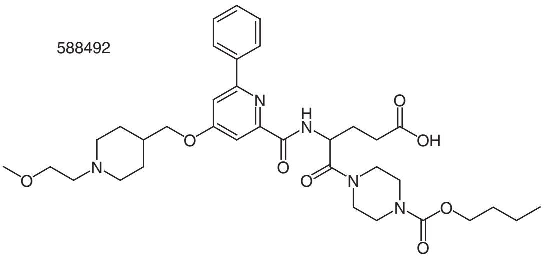
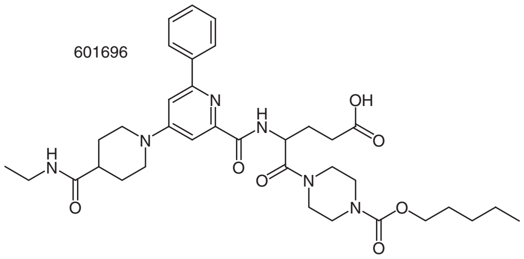
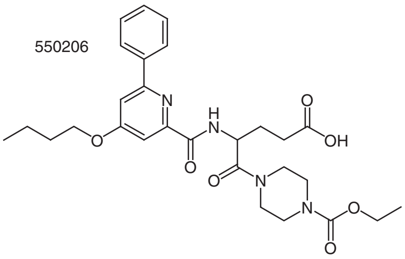
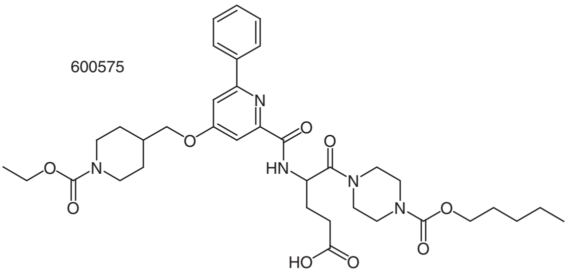
	 588492	 601696
pK_i exp	8.3	9.2
mean pK_i FW	8.3 (36 NBHs)	9.3 (36 NBHs)
mean pK_i SVR	8.3 (11 trials)	8.5 (9 trials)
	 550206	 600575
pK_i exp	7.3	8.9
mean pK_i FW	6.5 (51 NBHs)	8.1 (36 NBHs)
mean pK_i SVR	7.5 (6 trials)	8.1 (13 trials)

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.

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.

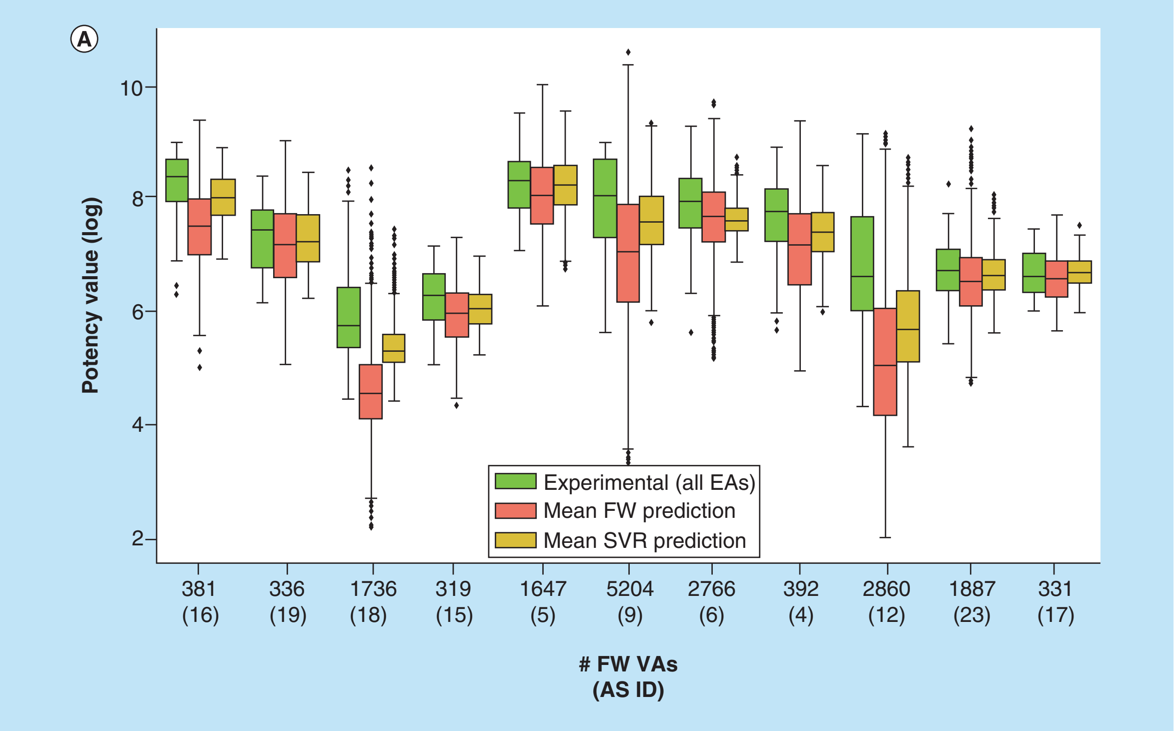


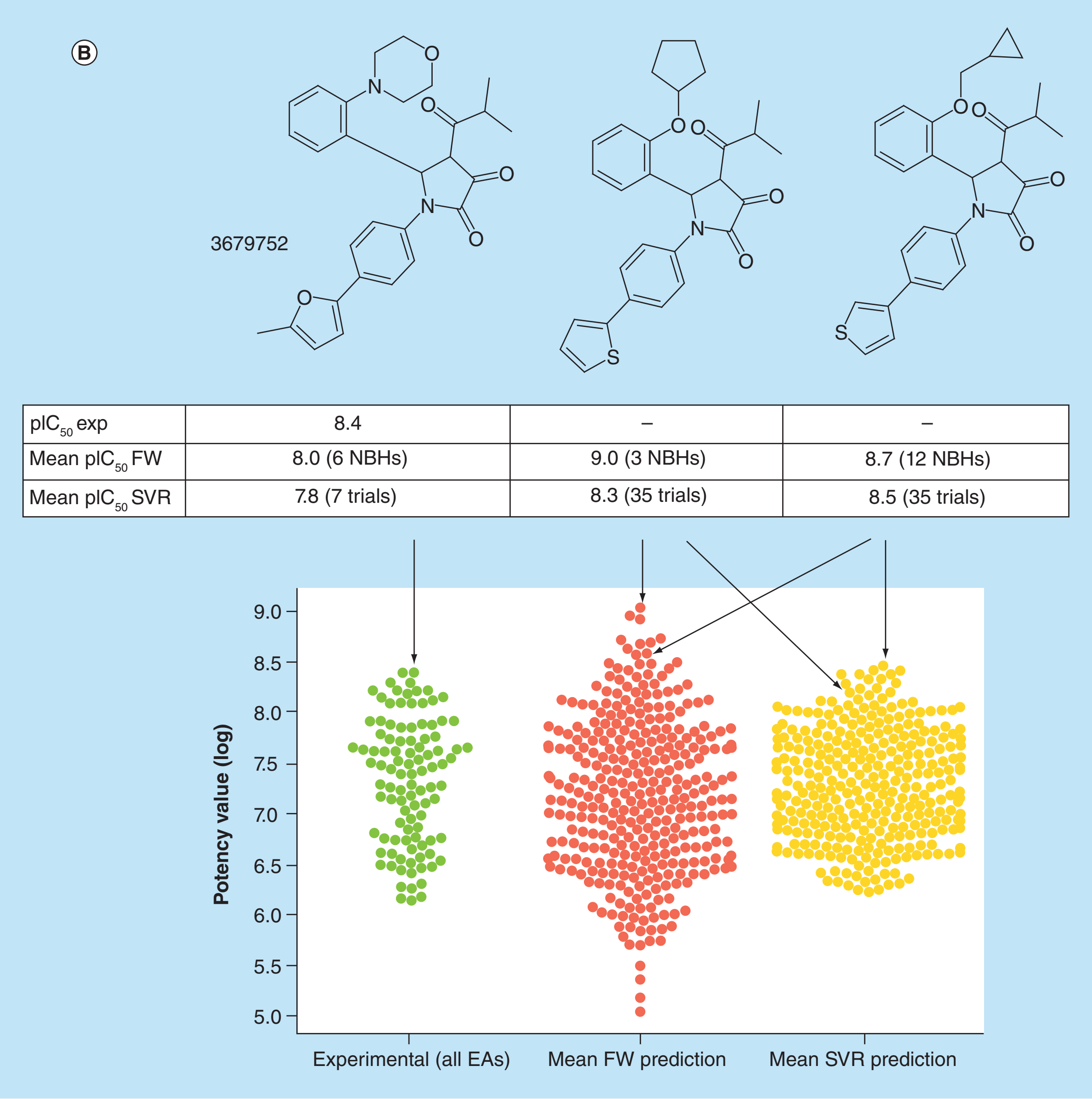
Figure 6. Potency predictions for Free-Wilson virtual analogs. (A) Box plots compare experimental potency value distributions of 11 ASs according to Figure 5A with potency predictions of corresponding FW VA populations. The y-axis reports logarithmic potency values and the x-axis the number of FW VAs 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 cyclopropanyl 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



pIC_{50} exp	8.4	–	–
Mean pIC_{50} FW	8.0 (6 NBHs)	9.0 (3 NBHs)	8.7 (12 NBHs)
Mean pIC_{50} SVR	7.8 (7 trials)	8.3 (35 trials)	8.5 (35 trials)

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.

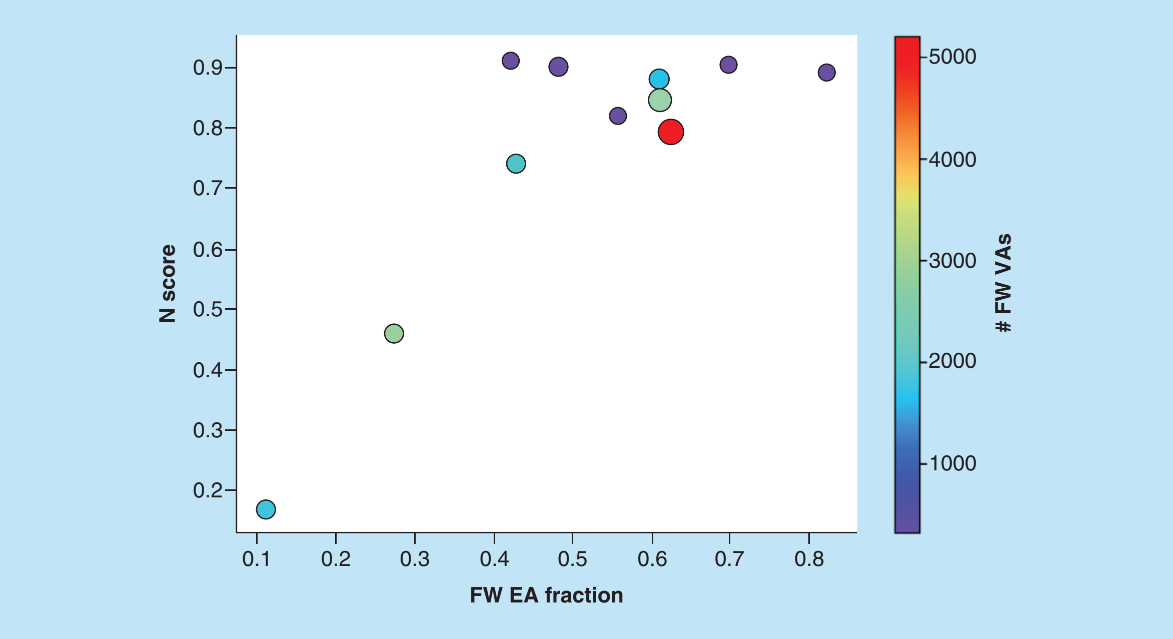


Figure 7. Free-Wilson neighborhood saturation scores. N scores are shown for ASs yielding predictive models as a function of increasing FW EA fraction, defined as the proportion of FW EAs among all EAs. Dots represent ASs that are scaled in size by the number of analogs per series and color-coded according 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:

$$N = 1 - \frac{n_{Fw\ EA}}{n_{Fw\ NBH}} \qquad \text{(Eq. 10)}$$

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

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.
- FW predictions of candidate compounds are explored.
- 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 Leoón 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. Griffen 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).

•• **Introduction of Free-Wilson theory for quantitative structure–activity relationship modeling.**

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/oechem-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).

Summary

In this study, diagnostic assessment of LO progress has been coupled with VA design and candidate selection strategies. Nonlinear QSAR regression models trained on entire ASs have typically outperformed linear ones, however their predictive ability has been shown to be limited for prioritizing highly potent VA candidates. On the other hand, simple local QSAR approximations derived from FW NBHs have consistently yielded VAs that are predicted to be more potent than experimentally measured EAs. Therefore, an algorithmic procedure tailored to identifying such FW VAs in each AS has been developed as an additional VA design strategy to AS core enumeration. Furthermore, comparison of global and local models revealed a general agreement in prediction results, thus lending further credence to their use as a combined predictive tool. Lastly, by quantifying the amount of FW NBHs associated with EAs, those ASs have been identified, which are most likely to benefit from expanding their current chemical space with accurately predicted FW VA candidates.

Based upon the results of this study, the idea of generating VA populations for different stages of LO by exploiting a generative *de novo* design strategy is evaluated in the following chapter.

Chapter 7

DeepCOMO: From Structure-Activity Relationship Diagnostics to Generative Molecular Design Using the Compound Optimization Monitor Methodology

Introduction

As shown in the previous chapter, rule-based generation of VAs presents an attractive opportunity for expanding AS chemical space with potentially highly active analogs. To these ends, two design strategies have been explored. The first one relies on randomly decorating the substitution site(s) of an AS core with R-groups while following retrosynthetic criteria and certain constraints in molecular size. The similarity of these enumerated VAs to EAs depends on the structural diversity of the R-group pool used for enumeration, which can be varied according to the scope of the analysis. The second design strategy identifies distinct MMP-based NBHs complying with the FW additivity principle and generates complementary VAs, the potency of which can be predicted by approximating local R-group contributions. By definition, these FW VAs represent the nearest structural neighbors of EAs and their number is individual for each AS depending on the underlying MMP distribution. While both procedures yield synthetically accessible compounds, they are restricted to predefined substitution sites and retrosynthetic criteria. One way to circumvent these limitations is the utilization of generative deep learning models for *de novo* design.

In this chapter, such models are employed as an additional third strategy for creating a new type of VAs for the COMO methodology. In particular, a transfer learning approach is pursued that focuses a pre-trained recurrent neural network model towards the chemical space of individual ASs, which significantly differ in their LO profiles. VAs sampled by the model are then evaluated with respect to their validity, uniqueness, and similarity to EAs. Finally, they are compared to the previously used enumerated and FW VAs by subjecting all populations to rigorous analysis of chemical space coverage, estimated synthetic tractability, and predicted potency values.

Reprinted with permission from “D. Yonchev, J. Bajorath. DeepCOMO: from structure-activity relationship diagnostics to generative molecular design using the compound optimization monitor methodology. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 1207-1218”. Copyright 2020 Springer Nature.

Journal of Computer-Aided Molecular Design (2020) 34:1207–1218
https://doi.org/10.1007/s10822-020-00349-3

PERSPECTIVE

DeepCOMO: from structure-activity relationship diagnostics to generative molecular design using the compound optimization monitor methodology

Dimitar Yonchev[1] · Jürgen Bajorath[1]

Received: 31 July 2020 / Accepted: 29 September 2020 / Published online: 5 October 2020

Abstract

The compound optimization monitor (COMO) approach was originally developed as a diagnostic approach to aid in evaluating development stages of analog series and progress made during lead optimization. COMO uses virtual analog populations for the assessment of chemical saturation of analog series and has been further developed to bridge between optimization diagnostics and compound design. Herein, we discuss key methodological features of COMO in its scientific context and present a deep learning extension of COMO for generative molecular design, leading to the introduction of DeepCOMO. Applications on exemplary analog series are reported to illustrate the entire DeepCOMO repertoire, ranging from chemical saturation and structure–activity relationship progression diagnostics to the evaluation of different analog design strategies and prioritization of virtual candidates for optimization efforts, taking into account the development stage of individual analog series.

Keywords Analog series · Lead optimization · Chemical saturation · SAR progression · Activity prediction · Generative deep learning

Introduction

The intuition- and experience-driven process of hit-to-lead and lead optimization (LO) presents key challenges for medicinal chemistry. If successful, it ranges from the initial demonstration of sustainable structure–activity relationships (SARs) of selected active compounds and the iterative generation of many analogs to the final stages of confirming pre-clinical candidate status of optimized compound(s). To this date, the LO process is difficult, if not impossible to rationalize. Work on analog series (ASs) continues until multi-property optimization criteria are met or insurmountable roadblocks are hit. This typically is far from being a black-and-white scenario. Partly unclear SAR responses or rather subtle differences between desirable and undesirable compound properties often propagate through optimization efforts until they amplify and result in large-magnitude problems. At such stages, when much work has already been spent on the long road to candidate compounds, it is often difficult to call it a day and discontinue work on advanced series. As a matter of fact, answering the question when sufficient numbers of analogs might have been generated and further progress would be unlikely to expect is at least as critical in the practice of medicinal chemistry as making meaningful initial decisions which compounds or series to advance or not. In light of these caveats looming over optimization efforts, it is self-evident that any approaches providing decision support during LO are more than welcome. However, the problems associated with empirical optimization are conceptually difficult to tackle. Currently, only a limited number of computational approaches are available that are capable of supporting LO efforts. This is the scientific context in which the Compound Optimization MOnitor (COMO) methodology evolved. One of the roots of COMO was the development of a scoring scheme to evaluate chemical saturation of compound series on the basis of biological screening data [1, 2]. Modifying and extending this scoring scheme and combining it with the assessment of SAR progression then gave rise to the introduction of the

✉ Jürgen Bajorath
bajorath@bit.uni-bonn.de

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

COMO approach [3–5], which was originally designed as a diagnostic. On the basis of COMO scoring, ASs can be assigned to different development stages. An integral feature of the COMO approach is the use of virtual analog (VA) populations to aid in the assessment of chemical saturation and SAR progression. By default, these VAs also represent potential candidate compounds for LO. Thus, although COMO was originally devised as a diagnostic/descriptive tool it also had the intrinsic potential to bridge between LO analysis and compound design. Accordingly, different analog design strategies and activity prediction approaches have been implemented in COMO to design and prioritize VAs [5, 6].

Herein, we report a methodological extension of COMO's analog design strategies through deep learning and generative modeling using recurrent neural networks (RNNs). Accordingly, the combined diagnostic scoring and extended analog design approach is termed DeepCOMO. In addition, we discuss current computational approaches having the potential to support different stages of LO efforts. In this context, we also describe key components of the DeepCOMO methodology. Furthermore, we present an application of DeepCOMO on two exemplary ASs, illustrating its entire analysis and design spectrum, as it has evolved since its inception [6]. Here, emphasis is put on the compound design aspect applying the DeepCOMO framework.

The subsequent sections are organized as follows. First, we review computational approaches that are of at least some relevance for chemical optimization (except standard QSAR techniques). Second, we discuss key methodological features of DeepCOMO. Third, exemplary applications are presented.

Computational approaches supporting compound optimization

Methods specifically developed to aid in different stages of LO are rare. Approaches that have been adopted and applied in the broader context of LO include statistical multi-parameter balancing and optimization of compound sets to suggest candidates for synthesis [7]. Furthermore, statistical attrition analysis of candidate compounds has also been reported to monitor whether compounds synthesized during LO meet pre-defined quality criteria [8]. Other approaches are focused on computational estimation of physicochemical properties [9], taking into consideration the widely applied rule-based oral availability paradigm [10] or ligand efficiency metrics [11]. Attempts have also been made to parameterize drug-likeness as a desirability function, aiming to generate preferred candidates [12]. Furthermore, computational approaches have been devised to elucidate SAR trends in evolving compound data sets [13] and analyze such trends in a qualitative [14] and quantitative [15] manner. Another interesting methodology that is based upon a statistical framework aims to quantify and visualize LO progression and assess the efficiency and tractability of different projects [16]. In addition, computational tools have been introduced to assess synthetic feasibility of candidate compound [17]. Given that compound design is of major importance during LO, computational approaches providing guidance for compound synthesis have been applied for experimental design [18]. Recently, artificial intelligence has entered the de novo design arena providing complex generative deep learning architectures that are also employed in support of LO campaigns [19, 20].

Taken together, most of the computational approaches that can be considered in the context of LO focus on compound property analysis, candidate selection, or design. By contrast, only very few methods have been introduced to monitor compound optimization and/or SAR trends in different ways [13, 16]. Hence, from this viewpoint, the diagnostic COMO framework was conceptualized to fill a void. As it has further evolved, a unique feature of the approach has become that it bridges between assessing progress in the optimization of ASs and compound design, as exemplified by the DeepCOMO extension introduced in the following.

Methodology: from COMO to DeepCOMO

Main principles and diagnostic scoring

COMO combines different scoring schemes including chemical saturation, multi-property, SAR progression, and SAR heterogeneity scores [3–6]. For AS diagnostics, the chemical saturation (S score) and SAR progression (P score) are primary measures for assigning ASs to different development stages. For the calculation of these diagnostic scores, VA populations play a central role because they serve as a representative sample of series-centric chemical space. In addition, the scoring scheme relies on the application of a chemical neighborhood (NBH) principle. Specifically, for chemical saturation and SAR progression diagnostics, the NBH of each existing analog (EA) comprising a series is defined and other compounds falling into the NBH are identified. Accordingly, for a given series, EAs and random samples of a chosen VA population are projected together into a user-defined chemical reference space (typically a vector space formed by numerical chemical descriptors) and the NBH of EAs is defined based on distance relationships between VAs, which determine chemical space coverage, given their large number compared to EAs. In a subsequent step, the proportion of VAs located in NBHs of EAs is calculated, giving rise to the coverage (C) and density (D) scores.

The C score quantifies how extensively EAs cover series-relevant chemical space and is defined as:

$$C = \frac{VA_{NBH}}{VA_{all}} \tag{1}$$

where VA_{NBH} is the number of VAs falling into any NBH of EAs and VA_{all} is the number of all projected VAs.

In addition, the D determines how densely EAs map chemical space by quantifying the overlap of their NBHs:

$$D = 1 - \frac{1}{d_{mean}} \tag{2}$$

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

$$d_{mean} = \frac{NBH_{O\_VA}}{n_{NBH}} \tag{3}$$

Both scores are complementary in their nature and can be summarized into the S score which is a composite metric defined as the harmonic mean of C and D:

$$S = \frac{2CD}{C + D} \tag{4}$$

While the C, D, and S scores are solely devised to quantify chemical saturation, the P score measures the degree of SAR progression as a function of SAR discontinuity in overlapping NBHs of EAs. Hence, for VAs located in overlapping NBHs, the mean pairwise potency range among EAs associated with those NBHs is quantified as the NBH-specific term $\bar{\Delta}_i$:

$$\bar{\Delta}_i = \frac{2}{m_i(m_i - 1)} \sum_{\substack{j,k=1 \\ j<k}}^{m_i} |\text{potj} \ \ - \text{potk}| \tag{5}$$

Here, m_i is the number of EAs with overlapping NBHs containing VA(s), potj and potk represent the logarithmic (log) potency of EA j and k, respectively. The P score for the entire AS represents the mean over Δ_i for all n VAs in overlapping NBHs of EAs applying a weighting scheme $w_i = \frac{1}{m_i}$ if $m_i > 1$ and $w_i = 0$ if $m_i = 1$:

$$P = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \bar{\Delta}_i \tag{6}$$

Hence, large potency variations between structurally similar EAs with overlapping NBHs containing VAs correspond to a strong SAR response to small chemical modifications and, accordingly, to high SAR progression within an AS.

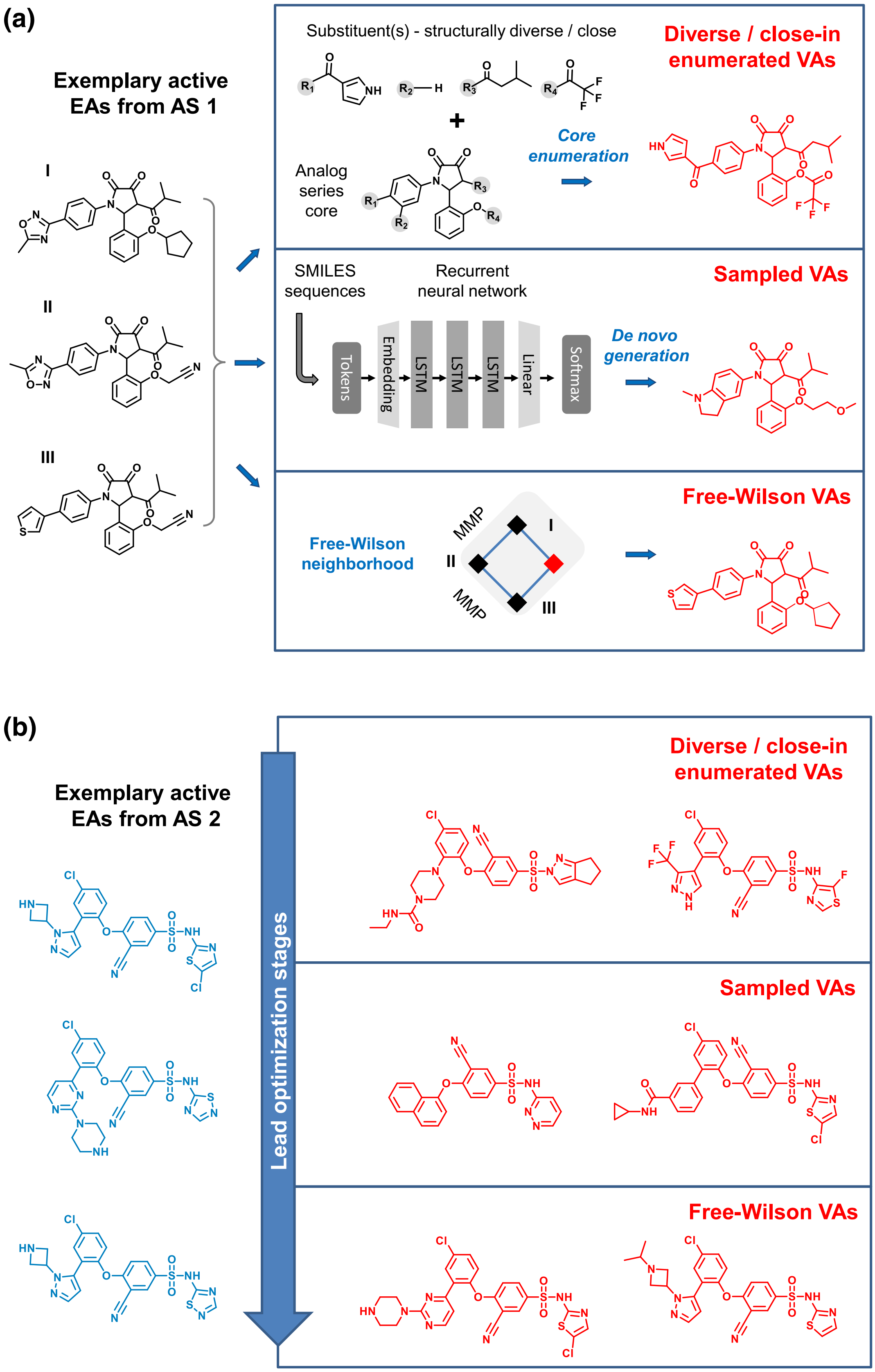
As designed, these scores are robust and practically insensitive to the number of VAs that are used, provided VAs outnumber EAs by at least two to three times [2–4]. Furthermore, the choice of chemical reference spaces for compound distance calculations is variable and can be modified according to the characteristic features and requirements of specific optimization efforts. Herein, a seven-dimensional chemical reference space composed of seven LO-relevant physicochemical descriptors (calculated with RDKit [21]) was used, which was shown to provide sufficient chemical resolution for the characterization of ASs in our previous studies [3–5]. Since the NBH concept plays a central role in the COMO approach, score calculations depend on the definition of a suitable NBH radius that adequately mirrors distances between EAs and VAs. Therefore, this hyper-parameter can be fine-tuned according to different VA populations and/or chemical space representations that might be used [2–4].

Virtual analog design strategies

Different strategies were designed and implemented to generate VA populations as diagnostic tools for COMO scoring and as candidate compounds for optimization efforts [2–4]. These analog design strategies are tailored towards different stages of the LO process (Fig. 1). First, VAs can be generated following a scaffold enumeration procedure. In this case, all substitution sites on the AS core scaffold are decorated with randomly selected terminal fragments according to pre-defined synthetic reactions. For ASs with multiple substitution sites, this procedure can often produce very large and complex VA structures that may not adequately represent AS-specific chemical space. This problem is circumvented by restricting VA size ranges to those of EAs and by randomly decorating one or more substitution sites with a hydrogen atom instead of an organic substituent based on an AS-specific substitution probability [4].

Applying the scaffold-based enumeration approach, two populations of VAs can be generated, termed *diverse* and *close-in VAs*, which differ only in the choice of substituents for enumeration. For diverse VAs, an external pool of R-groups is chosen that have not been used for EAs. For example, such a pool can be extracted from databases of known bioactive compounds. Conversely, for close-in VAs, only substituents obtained from fragmentation of the EAs comprising the AS under study are used for enumeration. Thus, in the case of diverse VAs novel, new chemistry might be introduced, which is more likely to be pursued during early stages of LO. On the other hand, close-in VAs are by design chemically more conservative and should thus be more relevant for mid-stages of LO projects.

In addition to AS scaffold-based enumeration, the Free-Wilson (FW) additivity principle [22, 23] has been adapted and converted into a design strategy for generating VA candidates for late LO stages [5]. Therefore, matched molecular pairs (MMPs) [24] are calculated for EAs of an AS


(a)
Exemplary active EAs from AS 1
I
II
III
Substituent(s) - structurally diverse / close
Analog series core
Core enumeration
Diverse / close-in enumerated VAs
SMILES sequences
Recurrent neural network
Tokens
Embedding
LSTM
LSTM
LSTM
Linear
Softmax
De novo generation
Sampled VAs
Free-Wilson neighborhood
MMP
MMP
Free-Wilson VAs
(b)
Exemplary active EAs from AS 2
Lead optimization stages
Diverse / close-in enumerated VAs
Sampled VAs
Free-Wilson VAs

◂**Fig. 1** Exemplary analogs and design strategies. **a** On the left, three exemplary EAs (black) from AS 1 are displayed (compounds I, II, and III). Sections on the right illustrate different VA design strategies of DeepCOMO, as discussed in the text. For each strategy, exemplary VAs are shown (red). **b** On the left, three EAs from AS 2 are shown (blue). On the right, exemplary VAs (red) are depicted resulting from the different design strategies according to (**a**)

and systematically organized in MMP networks (generated with the NetworkX Python library [25]). Then, analog sets are identified to which FW analysis of substituent contributions is applicable [22]. Such compounds can either be found among EAs (termed *FW EAs*) or they may represent VAs (*FW VAs*) with as of yet unexplored combinations of substituents. FW VAs are designed to become FW prediction targets on the basis of qualifying EAs. By definition, FW VAs can be viewed as a subset of close-in VAs since they contain only R-groups present in the AS. The FW VA population has the advantage of being specifically tailored towards FW potency predictions. Ensuing compound quartets meeting FW requirements consist of three EAs and an FW EA whose putative potency is predicted based upon FW principles. Such quartets represent local mini-QSAR models that have been shown to be surprisingly accurate in many cases and capable of complementing global QSAR strategies for VA prediction and prioritization [5]. Accuracy of FW predictions intrinsically depends on the presence of SAR continuity.

Herein we introduce a strategy for de novo design of VAs (termed *sampled VAs*) using an RNN architecture. This extension of AS-based VA design was inspired by the potential to further extend VA generation by taking information from related compound series or sets into account. Among the many recently introduced approaches for de novo compound design using deep learning, we have given preference to transfer learning (TL) considering the characteristics of the COMO framework.

For COMO-based design, TL [26] is applied to focus a generalized pre-trained generative model by fine-tuning using all EAs of a given AS. The implementation is based upon freely accessible code from the REINVENT 2.0 project [27] as implemented in PyTorch [28], which provides a robust pre-trained generative model (so-called Prior). The model has been trained on more than 1.4 million compounds from ChEMBL (release 25) [29] using tokenized SMILES strings with maximal sequence length of 256 elements [30]. Randomization of SMILES strings was applied as data augmentation technique [30]. As reported, the RNN architecture consists of an embedding layer of size 512, followed by three Long-Short-Term Memory (LSTM) layers of size 512, no drop-out layers, and a linear transformation layer of size 31 (equal to the vocabulary size of the corresponding training data), followed by a softmax function to convert the output into a token probability distribution. Furthermore, adaptive learning rate based on exponential learning rate decay with fixed patience was used [30] and the ADAM optimizer was applied [31]. In addition, a custom Uniformity-Completeness Jensen-Shannon Divergence (UC-JSD) metric [32] was used for estimating model performance. Further details are provided in the source publications [27, 30]. Since typically more than 99% of the compounds sampled using the Prior model have valid SMILES syntax [30] this model can serve as a starting point for TL on the basis of small and structurally confined sets of compounds such as ASs. During multiple epochs, the Prior model is fine-tuned to focus on AS-specific chemical space and generate complementary VAs. As introduced herein, DeepCOMO represents the TL-based extension of COMO's analog design capacity.

Potency prediction

To prioritize VAs for synthesis, potency prediction approaches are applied. In practice, it is hardly possible to systematically generate reliable linear or non-linear machine learning regression models for given ASs [5]. This is often due to their confined size, which limits the applicability of machine learning, and also to the presence of series-specific chemical features and SAR discontinuity in AS, both of which might constrain predictive modeling. Furthermore, regression models predict potency values for all VAs, which is also an approximation at best since VAs might often be inactive. However, compounds predicted to be most potent within VA populations principally represent preferred candidates for further consideration. For large ASs, we generally attempt to build global support vector machine regression (SVR) [33] and linear ridge regression [34] models to prioritize VAs. In addition, for all ASs, local FW predictions are attempted, which are supported by the generation of FW VAs for a given AS [5, 6], as described above.

Exemplary applications

To illustrate the different stages of DeepCOMO analysis, two exemplary ASs were selected as model series mimicking practical LO applications. These two ASs were obtained from our in-house high-confidence activity data version of ChEMBL (release 26) [29]. From this compound database, ASs were extracted using the compound-core-relationship algorithm [35]. Initially, all active compounds were subjected to systematic fragmentation of acyclic single bonds. Subsequently, resulting compound cores were organized into different series. To ensure that algorithmically generated AS cores contained synthetically accessible substitution sites, compound fragmentation was guided by 12 retrosynthetic rules [36, 37] and augmented by nine additional synthetic reactions [38] implemented with the aid of the OpenEye cheminformatics toolkit [39].

Selected analog series

The ASs studied here (termed AS 1 and AS 2) were active against the P2X purinoreceptor 3 (AS 1) and the sodium channel protein type IX alpha subunit (AS 2) and consisted of 219 and 158 analogs, respectively. For all compounds, IC_{50} measurements were available and recorded as negative logarithmic potency values (pIC_{50}). The composition of AS 1 and 2 is summarized in Table 1. These ASs were selected for several reasons. They were among the largest ASs that we algorithmically extracted from public domain data. Furthermore, these ASs were of moderate structural complexity and contained different core structures with four (AS 1) and three (AS 2) substitution sites, hence providing ample opportunities for analog design. Figure 1a and b shows exemplary analogs from AS 1 and 2, respectively.

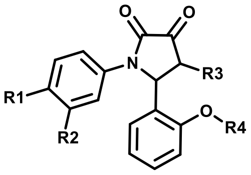
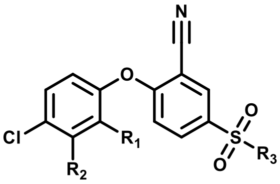
Diverse, close-in, and Free Wilson virtual analogs

Alternative analog design strategies are schematically illustrated in Fig. 1a. As discussed in detail below, TL produced initial sets of 51,200 SMILES representations per AS. For comparison, equally sized sets of diverse and close-in analogs were generated utilizing all substitution sites per AS. Diverse VAs were randomly enumerated using a pool of 44,636 substituent fragments comprising at most 13 atoms that were extracted from bioactive compounds in ChEMBL (release 26). For enumerating close-in VAs, series-based sets of 70 (AS 1) and 133 substituents (AS 2) were used. Different from diverse and close-in VAs, the number of FW VAs per AS is not variable but depends on intra-series structural relationships, the corresponding distribution of MMPs, and the potential to complement FW NBHs formed by EAs with FW VAs (see above). For AS 1 and 2, a total number of 907 and 3167 FW VAs was obtained, respectively. Figure 1a and b show exemplary VAs for AS 1 and 2, respectively.

Diagnostic scoring

Next, chemical saturation and SAR progression scores were calculated for both series using the respective close-in VAs as diagnostic VA populations. Therefore, sets of 1000 VAs were randomly selected for 10 independent score calculations, producing very similar results. Mean scores are reported in Table 1. These scores clearly differentiated between the two ASs. Although AS 1 contained only ~ 25% more compounds than AS 2, it was found to be chemically much more saturated (S score = 0.58) with high substantial coverage of chemical reference space (C score = 0.43) and particularly high density of coverage (D score = 0.90). By contrast, chemical saturation of AS 2 was significantly lower (S score = 0.29), resulting from low coverage (C score = 0.18) of chemical space and more moderate density of coverage (D score = 0.73). However, a different picture emerged when SAR progression scores were compared. Here, AS 2 displayed much stronger SAR responses (P score = 0.95) than AS 1 (P score = 0.55), reflecting the presence of higher SAR discontinuity in overlapping NBHs of EAs. Hence, AS 1 was characterized as a further explored

Table 1 Analog series characteristics

Analog series ID	1	2
Biological target	P2X purinoreceptor 3	Sodium channel protein type IX alpha subunit
ChEMBL target ID	2998	4296
Potency measurement type	IC50	IC50
# EAs	219	158
# EAs in Free-Wilson NBHs	183 (84%)	45 (28%)
# substitution sites	4	3
# unique substituents	70	133
Analog series core		
C score	0.43 (±0.01)	0.18 (±0.02)
D score	0.90 (±0.00)	0.73 (±0.03)
S score	0.58 (±0.01)	0.29 (±0.02)
	0.55 (±0.03)	0.95 (±0.05)

EA existing analog, *NBH* neighborhood compound series with higher series-specific chemical saturation and more balanced potency variations among analogs. On the other hand, the scores indicated that AS 2 still had significantly potential for obtaining analogs with further improved potency. Thus, on the basis of this comparison, AS 1 was categorized as a later-stage series, whereas AS 2 represented an early-/mid-stage series. Notably, conclusions drawn from scoring were fully consistent with the numbers of FW NBHs and participating EAs detected in both ASs. While the subset of FW EAs from AS 1 amounted to 183 (84%) EAs associated with at least one FW NBH, this was the case for only 45 (28%) of the EAs from AS 2, hence reflecting the more advanced development stage of AS 1. Based on these diagnostic findings, one can then decide which VA design strategy would be preferred to generate additional candidate compounds. For instance, AS 1 is likely to benefit from FW VAs as potential candidates (high structural similarity to EAs), given its advanced development stage. On the other hand, for AS 2, a more explorative design strategy would be preferred to further diversify candidate compounds.

Transfer learning

The TL extension included in DeepCOMO was then applied to sample different VAs, aiming to navigate from generalized drug-like space towards narrowly confined series-centric space and further extend VA design.

The generative model was trained for 50 epochs with 1024 sampled VAs per epoch obtained as SMILES strings, which resulted in a total of 51,200 initially sampled strings per AS. Then, the population of sampled VAs was analyzed with respect to model TL performance. Because TL was increasingly focused on a specific AS core structure a well-performing model should be capable of generating many chemically meaningful structures and unique compounds similar to yet chemically distinct from EAs. Figure 2 shows the evolution of the TL model during training and fine-tuning. Beginning with epoch 1, the generalized Prior model produced a uniform random VA sample without compounds containing the AS cores. However, over the course of only few epochs, the model rapidly learned to sample increasing numbers of compounds similar to EAs, as indicated by the steep rise of the curves accounting for the proportion of sampled VAs with AS cores. The models also reproduced EAs from the training sets (Fig. 2), confirming focused sampling of VAs. Furthermore, the apparent focusing effect was accompanied by a similarly steep decrease in the total numbers of unique sampled VAs. By the 50th epoch, less than 50% and 60% of the sampled VAs represented unique compounds for AS 1 and 2, respectively. Around the 30th epoch, the proportion of generated VAs sharing the AS cores or reproduced EAs reached a plateau at which the ratios between the different curves remain relatively constant. By the 50th epoch, approximately 67% of the EAs of both ASs were reproduced within a single epoch run, whereas the fraction of unique sampled VAs containing the AS core was consistently above 80% and 75% for AS 1 and 2, respectively. Taken together, the analysis revealed successful focusing of the TL model for both ASs, with increasing levels of redundancy when sampling VAs.

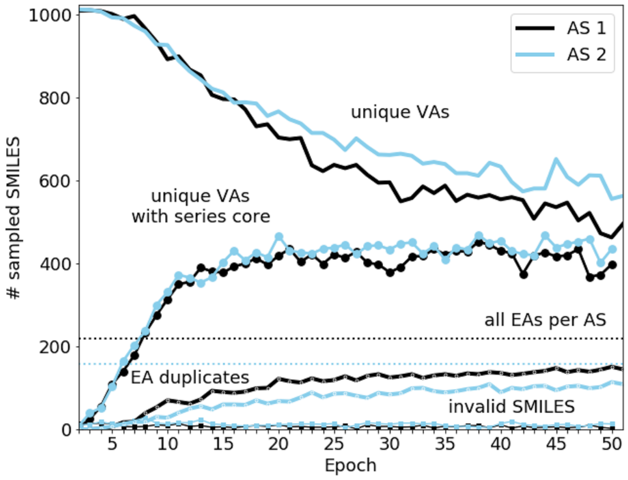


Fig. 2 Design of virtual analogs via transfer learning. Shown is the evolution of multiple parameters across 50 epochs of sampling VAs via transfer learning for AS 1 (black lines) and 2 (blue). The x-axis reports the number of epochs and the y-axis the number of sampled VAs (SMILES strings). Curves with filled circles monitor increasing numbers of sampled VAs containing their AS cores. Dotted horizontal lines indicate the number of EAs for each AS. Curves below these lines record the number of duplicated (reproduced) EAs of each AS. At the bottom, curves with squares monitor the fraction of sampled VAs with invalid SMILES strings

Next, we analyzed how effectively the TL model sampled VAs across different epochs. The 50 training epochs with a SMILES sample size of 1024 produced a total of 26,081 and 28,592 unique VA structures for AS 1 and AS 2, respectively, which corresponded to ~ 51% and ~ 55% of all sampled SMILES strings for AS 1 and AS 2, respectively. These ratios were a consequence of increasing sampling of duplicate structures and reproduced EAs within individual epochs (Fig. 2). Since this effect propagated throughout the fine-tuning phase, some VAs were sampled in multiple epochs, whereas others were obtained in very few or just one. As illustrated in Fig. 3, the majority of sampled VAs was generated during only one of the epochs. In addition, the number of VAs sampled in multiple epochs significantly decreased over increasing number of epochs. In Fig. 3, four exemplary structures of VAs of AS 1 are depicted that were sampled in different numbers of epochs. These VAs were selected from the batch generated during the 40th epoch when the output of the generative model was stable (Fig. 2). The

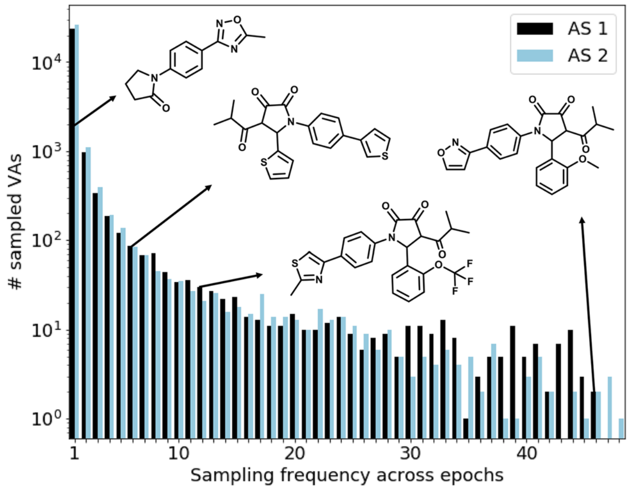


Fig. 3 Sampling frequencies of virtual analogs. The bar plot reports the frequency of occurrence for sampled VAs during TL (AS 1, black; AS 2, blue). The x-axis reports the number of epochs and the y-axis the numbers of VAs falling into each category on a logarithmic (log) scale. For AS 1, exemplary sampled VAs with different sampling frequencies (indicated by the black arrows) are depicted

VA in the upper left corner in Fig. 3 was sampled only once during the 50 generative epochs because it did not contain core of AS 1 but represented a simpler structure, consistent with the initial generalization capability of the Prior model. The next sampled VA to the right contained a substructure of the AS 1 core.

In which the signature *o*-alkoxyphenyl ring at the gamma lactam position was substituted with a thiophene ring. Although this sampled VA not contain the entire AS 1 core, it was sampled six times (in epochs 8, 16, 26, 40, 41, and 48) including the late stages of TL. Thus, the model consistently diversified structural features of sampled VAs including core modifications, even after focusing on the same core over many epochs. These observations mirrored an intrinsic advantage of the deep generative architecture over the simpler VA enumeration strategies based upon a conserved AS core. The third VA from the left in Fig. 3 contained the complete core of AS 1, but was only sampled during 12 of 50 epochs. This is likely due to the varying frequency of occurrence of individual substituents among the EAs used for training. For example, the *o*-methylthiazole and trifluoromethyl groups were only present in two and 15 training instances, respectively. By contrast, the VA on the right with different more frequently occurring substituents was most frequently sampled in 46 epochs. These comparisons illustrate the spectrum of structural modifications of sampled VAs obtained by AS-centric fine-tuning of the model, yielding an expansion of VA space.

Table 2 Virtual analogs statistics

Analog series ID	1	2
# experimental EAs	219	158
# FW VAs	907	3167
# unique sampled VAs # diverse VAs # close-in VAs	26,295	28,748
Sampled VAs & EAs	214	156
Sampled VAs & FW VAs	624	1436
Sampled VAs & close-in VAs	208	1669
Sampled VAs & diverse VAs	0	18
Close-in VAs & diverse VAs	35	37
FW VAs & close-in VAs	53	2909
FW VAs & diverse VAs	0	0

EA existing analog, *VA* virtual analog, *FW* Free-Wilson, *&* intersection

Comparison of virtual analog populations

Next, the coverage of AS-specific chemical space by VA populations produced using the four design strategies of the DeepCOMO framework was analyzed and compared. First, the overlap between differently designed VA populations (and between VAs and EAs) was determined, as reported in Table 2. From the pools of diverse and close-in VAs of AS 1 and 2, subsets were randomly selected to match the number of sampled VAs. For both ASs, nearly all EAs were reproduced by TL. However, the TL model sampled only 69% of the FW VAs of AS 1 and 45% of AS 2. Apart from this, the overlap between different compound populations was generally larger for AS 2 than AS 1. The largest difference was observed between the overlap of close-in VAs with other compound populations. Nonetheless, in both cases, all four VA design strategies produced significant numbers of unique compounds, indicating their principal complementarity in charting analog space. In the next step, VA distributions in series-centric chemical space were compared. Therefore, EAs and equally sized random samples of all VA populations were projected into the descriptor-based seven-dimensional reference space and subjected to dimension reduction using principal component analysis (PCA). Plots were generated using the first two principal components. For both series, equivalent observations were made. For AS 1, pairwise comparisons of the EA distribution and different VA distributions are shown in Fig. 4a–d. As expected, the FW VA population mapped most closely to EAs (Fig. 4a), consistent with the underlying FW NBH-directed design strategy. Close-in VAs were already more widely distributed but mostly covered regions proximal to EA (4b). For diverse VAs, a more extensive spread was observed (4c). For PCA, sampled VAs shown were exclusively selected from the batch obtained for 40th epoch and thus represented a

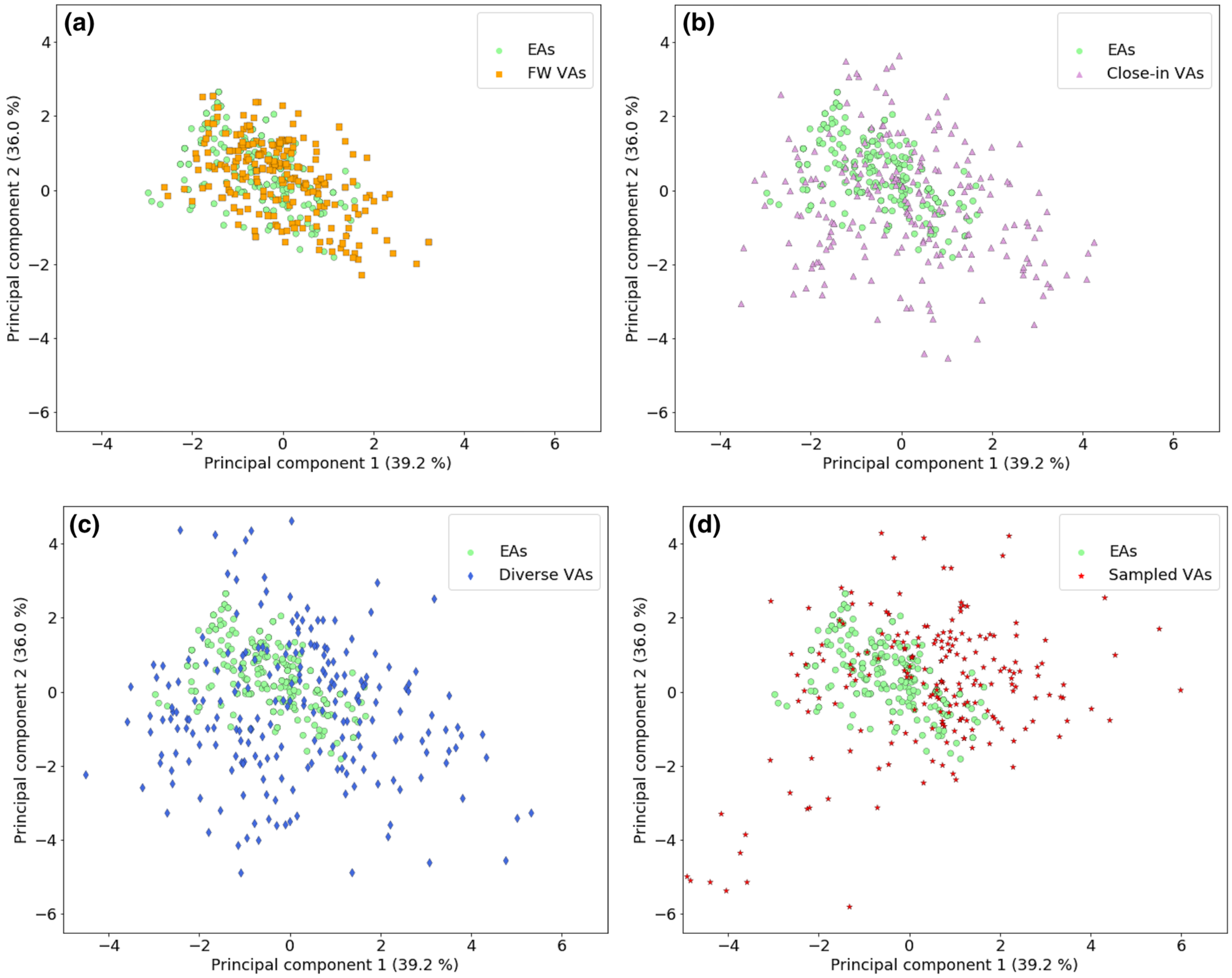


Fig. 4 Chemical space coverage. In **a**–**d**, PC plots compare the coverage of chemical reference space by AS 1 with its four VA populations. Sampled VAs were randomly selected from the batch of the 40th epoch. For each principal component, it is reported for how much of the original data variance it accounts

late-stage "snapshot" of the fine-tuned TL model. Deriving this VA population combined information from compounds with varying structural relationships to EAs and facilitated additional core modifications. Accordingly, the comparison of sampled VAs and EAs in Fig. 4d revealed a combination of different patterns observed for other VAs including strong focusing on subsets of EAs, proximal mapping to many others, but also substantial diversification. Hence, the distribution of sampled VAs combined and further extended characteristics of VA populations obtained with simpler design strategies.

Synthetic accessibility of virtual analogs

Synthetic accessibility of VAs continues to represent a much discussed topic, especially for compounds generated using deep learning architectures. Accordingly, we also calculated and compared synthetic accessibility (SA) scores [17] for our VA populations (using the public RDKit implementation available on GitHub [17]). The SA score ranges from 1 to 10 and accounts for fragment contributions to compounds based upon empirical assessment of synthetic building blocks, stereo chemistry, and non-standard structural features [17]. Increasing scores indicate the presence of chemically complex compounds that are increasingly challenging to synthesize. As shown in Fig. 5, VA populations for AS 2 yielded SA scores that were comparable to or only slightly higher than EA scores, hence indicating general synthetic feasibility. Equivalent observations were made for AS 1. Overall broadest score distributions including subset of higher scoring compounds were observed for diverse VA, which one might expect, as these VAs combine substituent fragments

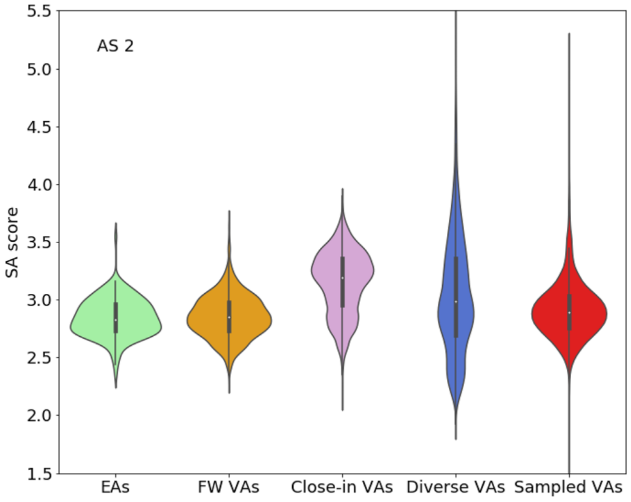


Fig. 5 Synthetic accessibility. Violin plots report SA score distributions for AS 2 and its VA populations

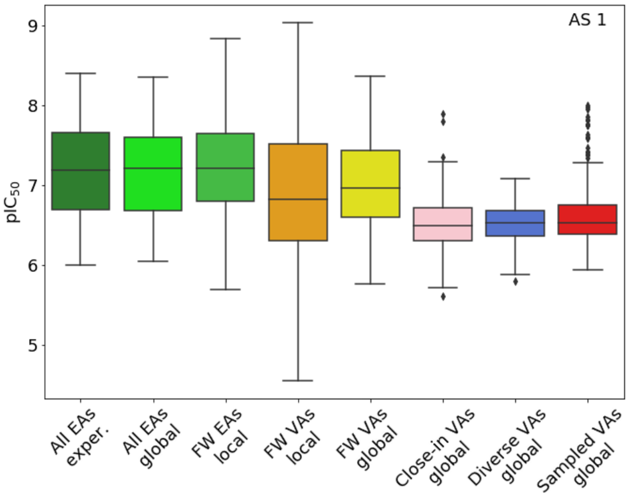


Fig. 6 Compound potency predictions. Box plots report potency predictions for AS 1 and its VA populations using global (SVR) and local (FW) models. The latter models are only applicable to FW EAs and FW VAs

from the entire universe of current bioactive compounds, regardless of their core structures.

Prioritization of virtual analogs

VAs predicted to be most potent represent preferred candidates for further optimization efforts. For AS 1 and AS 2, global series-based and local FW NBH-based prediction models were derived. For global predictions, SVR [33] models were trained via three-fold double cross-validation [40]. For model building, a folded (2048-bit) version of the extended connectivity fingerprint with bond diameter of 4 (ECFP4) [41] was used in combination with the Tanimoto kernel [42] as a similarity function. All calculations were carried out using Python's scikit-learn library [43]. For training, 517 (AS 1) and 1135 (AS 2) compounds with activity against each AS target were collected from ChEMBL that did not belong to the AS (representing structurally diverse active compounds) and combined with 50% of the respective AS. The remaining 50% of the EAs were used as an external validation set. The SVR models were then used to predict the potency of these EAs and of the different VA populations. Furthermore, FW NBH-based potency predictions were carried out for FW VAs and qualifying FW EAs. For AS 1, prediction results are reported in Fig. 6 (comparable observations were made for AS 2). Accurate retrospective potency predictions were obtained for EAs using both local and global models, with R^2 values of 0.84 (±0.0) and 0.81 (±0.03), respectively, and mean absolute errors of 0.18 (±0.0) and 0.2 (±0.03), respectively. For VA populations, global models generally predicted lower potency values than for EAs, as observed previously [5]. Overall highest potency was predicted using local and global models for FW VAs, which most closely resembled EAs. However, for all except diverse VAs, at least few "outlier" compounds were predicted to have higher potency than most of the EAs. These compounds provide focal points for VA prioritization as potential candidates depending on the development stages of an AS, as assessed by COMO scoring.

Conclusions

In medicinal chemistry, LO is still more of an art form than a scientific exercise following firm and generally applicable rules. It is governed by the recurrent need to decide which compounds to make next. These largely experience- and chemical intuition- driven optimization efforts greatly benefit from any approaches that are capable to rationalize at least a part of the proceedings and provide decision support beyond subjective judgment. In principle, computational methods are prime candidates to support LO. However, as discussed herein, only few relevant approaches besides standard QSAR techniques have thus far been introduced to aid in this process. Within this scientific context, COMO was conceptualized, originally as a diagnostic framework, and then further expanded to bridge between chemical/SAR analysis and compound design. Herein, we have discussed key features of the methodology and presented the DeepCOMO extension for further advanced compound design. DeepCOMO provides four design strategies that yield complementary VA populations with varying AS-centric chemical space coverage. It has been applied to two exemplary ASs at different development stages, illustrating the spectrum of its diagnostic and design components and rationalizing how to combine these components for COMO-guided decision making. We hope that our discussion and findings presented herein might catalyze the development of additional computational concepts and methods to aid in compound optimization efforts, which would certainly be beneficial for the practice of medicinal chemistry. Applications of DeepCOMO in practical medicinal projects are underway.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Kunimoto R, Miyao T, Bajorath J (2018) Computational method for estimating progression saturation of analog series. RSC Adv 8:5484–5492
2. Yonchev D, Vogt M, Stumpfe D, Kunimoto R, Miyao T, Bajorath J (2018) Computational assessment of chemical saturation of analog series under varying conditions. ACS Omega 3:15799–15808
3. Vogt M, Yonchev D, Bajorath J (2018) Computational method to evaluate progress in lead optimization. J Med Chem 61:10895–10900
4. Yonchev D, Vogt M, Bajorath J (2019) Compound optimization monitor (COMO) method for computational evaluation of progress in medicinal chemistry projects. Future Drug Discov 1:FDD15
5. Yonchev D, Bajorath J (2020) Integrating computational lead optimization diagnostics with analog design and candidate selection. Future Sci OA 6:FSO451
6. Yonchev D, Vogt M, Bajorath J (2020) From SAR diagnostics to compound design: development chronology of the compound optimization monitor (COMO) method. Mol Inform. https://doi.org/10.1002/minf.202000046
7. Segall M (2014) Advances in multiparameter optimization methods for de novo drug design. Expert Opin Drug Discov 9:803–817
8. Munson M, Lieberman H, Tserlin E, Rocnik J, Ge J, Fitzgerald M, Patel V, Garcia-Echeverria C (2015) Lead optimization attrition analysis (LOAA): a novel and general methodology for medicinal chemistry. Drug Discov Today 20:978–987
9. Lobell M, Hendrix M, Hinzen B, Keldenich J, Meier H, Schmeck C, Schohe-Loop R, Wunberg T, Hillisch A (2006) In silico ADMET traffic lights as a tool for the prioritization of HTS hits. ChemMedChem 1:1229–1236
10. Lipinski CA (2004) Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 1:337–341
11. Cavalluzzi MM, Mangiatordi GF, Nicolotti O, Lentini G (2017) Ligand efficiency metrics in drug discovery: the pros and cons from a practical perspective. Expert Opin Drug Discov 12:1087–1104
12. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopking AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4:90–98
13. Shanmugasundaram V, Zhang L, Kayastha S, de la Vega de Leon A, Dimova D, Bajorath J (2016) Monitoring the progression of structure–activity relationship information during lead optimization. J Med Chem 59:4235–4244
14. Iyer P, Hu Y, Bajorath J (2011) SAR monitoring of evolving compound data sets using activity landscapes. J Chem Inf Model 51:532–540
15. Peltason L, Bajorath J (2007) SAR index: quantifying the nature of structure- activity relationships. J Med Chem 50:5571–5578
16. Maynard AT, Roberts CD (2016) Quantifying, visualizing, and monitoring lead optimization. J Med Chem 59:4189–4201
17. Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminform 1:e8
18. Hoffer L, Voitovich YV, Raux B, Carrasco K, Muller C, Fedorov AY, Derviaux C, Amouric A, Betzi S, Hovath D, Varnek A, Colette Y, Combes S, Roche P, Morelli X (2018) Integrated strategy for lead optimization based on fragment growing: the diversity-oriented-target-focused-synthesis approach. J Med Chem 61:5719–5732
19. Ståhl N, Falkman G, Karlsson A, Mathiason G, Boström J (2019) Deep reinforcement learning for multiparameter optimization in de novo drug design. J Chem Inf Model 59:3166–3176
20. Zhou Z, Kearnes S, Li L, Zare RN, Riley P (2019) Optimization of molecules via deep reinforcement learning. Sci Rep 9:10752
21. RDKit (2013) Cheminformatics and machine learning software. https://www.rdkit.org. Accessed 1 June 2020
22. Free SM, Wilson JW (1964) A mathematical contribution to structure-activity studies. J Med Chem 7:395–399
23. Kubinyi H (1988) Free-Wilson analysis. Theory, application and its relationships to Hansch analysis. Quant Struct Act Relat 7:121–133
24. Griffin E, Leach AG, Robb GR, Warner DJ (2012) Matched molecular pairs as a medicinal chemistry tool. J Med Chem 54:7739–7750
25. Hagberg A, Swart P, Chult DS (2008) Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Laboratory, NM, USA. https://www.osti.gov/biblio/960616
26. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating focused molecule libraries for drug discovery with recurrent neural networks. ASC Cent Sci 4:120–131
27. Blaschke T, Arús-Pous J, Chen H, Margreitter C, Tyrchan C, Engkvist O, Papadopoulos K, Patronov A (2020) REINVENT 2.0—an AI tool for de novo drug design. ChemRxiv. https://doi.org/10.26434/chemrxiv.12058026.v2
28. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in PyTorch. Adv Neural Inf Process Syst 30:1–4
29. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–1107
30. Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tychan C, Reymond JL, Chen H, Engkvist O (2019) Randomized SMILES strings improve the quality of molecular generative models. J Cheminform 11:e71
31. Kingma DP, Ba J (2014) ADAM: A method for stochastic optimization. arXiv:1412.69.80

32. Arús-Pous J, Blaschke T, Ulander S, Reymond JL, Chen H, Engkivst O (2019) Exploring the GDB-13 chemical space using deep generative models. J Cheminform 11:e20
33. Vapnik VN (2000) The nature of statistical learning theory, 2nd edn. Springer, New York
34. Marquardt DW, Snee RD (1975) Ridge regression in practice. Am Stat 29:3–20
35. Naveja JJ, Vogt M, Stumpfe D, Medina-Franco JL, Bajorath J (2019) Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. ACS Omega 4:1027–1032
36. Lewell XQ, Judd DB, Watson SP, Hann MM (1998) 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 38:511–522
37. de la Vega de Leon A, Bajorath J (2014) Matched molecular pairs derived by retrosynthetic fragmentation. Med Chem Commun 5:64–67
38. Hartenfeller M, Eberle M, Meier P, Nieto-Oberhuber C, Altmann KH, Schneider G, Jacoby E, Renner S (2011) A collection of robust organic synthesis reactions for in silico molecule design. J Chem Inf Model 51:3093–3098
39. OEChem TK (2012) OpenEye Scientific Software Inc, NM, USA. https://www.eyesopen.com/oechem-tk. Accessed 1 June 2020
40. Baumann D, Baumann K (2014) Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. J Cheminform 6:e47
41. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754
42. Ralaivola L, Swamidass SJ, Saigo H, Baldi P (2005) Graph kernels for chemical informatics. Neural Netw 18:1093–1110
43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Summary

In this work, the final extension of the COMO methodology has been introduced as an additional approach for generating focused virtual libraries for ASs. An RNN model, initially pre-conditioned to uniformly sample valid compounds covering a large biologically relevant chemical space, has been fine-tuned to sample VAs within AS-relevant chemical space through repeated exposure to a training set of all EAs from a given series. Following this entirely probabilistic approach, a large pool of syntactically correct SMILES corresponding to novel molecules with variable degree of similarity to EAs has been sampled, including virtual compounds that contain the congeneric series scaffold but are substituted at different sites. Comparison of sampled VAs to FW, close-in, and diverse enumerated VAs has revealed similar yet distinct chemical space coverage with very limited structure overlap between the different populations. Sampled VAs have been shown to exhibit synthetic accessibility comparable to that of EAs. Furthermore subsets have been identified as potentially highly active candidates by QSAR models. Despite the challenging setup, which has included two large series inherently different in their LO characteristics, comparable results have been obtained for both of them. Importantly, depending on the LO categorization of ASs, different VA populations can be utilized as candidates given their distinct chemical space coverage. Accordingly, diverse VAs are more likely to be suitable for early-stage, close-in VAs for mid-stage, and FW VAs for late-stage series. Since *de novo* sampled VAs combine characteristic features of all other VA populations but also go beyond deterministic (rule-based) design constraints, they can be viewed as complementary compound pools for potential use in all stages of LO.

Taken together, this study shows that differently designed VA populations have potential utility for different stages of LO. Diagnostic AS categorization with COMO can be employed for rationalizing the choice of VA candidates for series expansion.

Chapter 8

Conclusion

Chemical LO is a central task in medicinal chemistry projects. Thereby, new compounds are iteratively synthesized and tested aiming to produce (pre)clinical candidates with favorable chemical and biological properties. This gives rise to ASs with different potential for further exploration. Since LO campaigns are mainly guided by medicinal chemistry intuition, objective assessment of their progress is challenging and can greatly benefit from data-driven computational methods. While a variety of *in silico* approaches for characterization and property prediction of individual compounds exist, rationalization on the basis of entire ASs has remained largely unexplored.

This thesis reports the evolution of computational methods for evaluation of progress during LO combined with virtual compound design and candidate selection strategies. In the first study (*Chapter 3*), the utility of quantifying chemical space saturation as a novel principle for distinguishing between different AS profiles is explored. Herein, a previously introduced methodology, which relies on diagnostic VA populations for delineating AS-relevant chemical space and deriving distance-based chemical NBHs around EAs, has been further refined. Chemical saturation characteristics of ASs have been thoroughly analyzed under varying conditions via a dual scoring scheme. This has yielded consistent and robust results, which have served as a motivation for further enhancing the methodology. Consequently, in *Chapter 4*, chemical saturation analysis has been extended with quantification of NBH-based SAR progression as an additional component for estimating the LO potential of individual series. Based on the newly introduced intuitive scoring system, a set of mid- to large-sized ASs has been successfully profiled and different LO scenarios have been rationalized. Hence, this study has laid the foundation for the development of COMO as a holistic method for evaluation of progress in LO campaigns presented in *Chapter 5*. Therein, complementary scores accounting for chemical saturation, SAR progression and heterogeneity as well as for multiple physicochemical properties have provided a detailed multi-dimensional view on AS characteristics and have uncovered more subtle features. Importantly, the approach allows for flexibility in fine-tuning critical methodological parameters. In the next study (*Chapter 6*), the utility of COMO has been extended beyond the retrospective evaluation of current LO status by integrating prospective *de novo* design strategies. The motivation has been to augment COMO's purely diagnostic capability with a system for prioritization of potentially active synthetic candidates to support medicinal chemists in decision-making for future series development. Therefore, synthetically accessible diagnostic VA populations and specifically designed FW VAs have been utilized as candidate pools and their potency predicted via ML regression and FW-type QSAR models, respectively. For all studied ASs, highly potent FW VAs have been obtained as prospective candidates. Moreover, extensive neighborhood behaviour among EAs has been rationalized as a suitable criterion for AS expansion. Finally, in *Chapter 7*, an additional VA design strategy based upon a generative deep learning method has been explored, giving rise to the DeepCOMO extension. SMILES-based *de novo* design with RNNs has been exploited in a transfer learning setting by navigating the output domain of a pre-trained model towards the confined chemical space of individual ASs. Hereby sampled VAs have been compared to deterministically generated VAs with respect to chemical space coverage, structure overlap, synthetic feasibility, and candidate potential. As a result, they have been shown to combine characteristic features of both enumerated and FW VAs while going beyond the limitations of explicitly defined rules and scaffold substitution sites. Hence, they have been rationalized as complementary VA groups. Importantly, all differently designed virtual compound populations analyzed in this work have the potential to be utilized as synthetic candidate pools tailored to medicinal chemists' needs in different stages of LO.

In conclusion, this thesis provides an in-depth exploration of novel computational strategies for evaluating progress in LO campaigns based on characterizing entire ASs. This has led to the natural evolution of COMO as a method for diagnostic series profiling augmented with strategies for focused library design and candidate prioritization. The presented cheminformatic framework is thought to provide a practical and intuitive tool to aid medicinal chemists in rationalizing and planning LO efforts.

Bibliography

[1] Drews, J. Drug discovery: a historical perspective. *Science* **2000**, *287*, 1960–1964.

[2] Jones, A. W. Early drug discovery and the rise of pharmaceutical chemistry. *Drug Test. Anal.* **2011**, *3*, 337–344.

[3] Kinch, M. S.; Haynesworth, A.; Kinch, S. L.; Hoyer, D. An overview of FDA-approved new molecular entities: 1827–2013. *Drug Discov. Today* **2014**, *19*, 1033–1039.

[4] DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **2016**, *47*, 20–33.

[5] Morgan, S.; Grootendorst, P.; Lexchin, J.; Cunningham, C.; Greyson, D. The cost of drug development: a systematic review. *Health Policy* **2011**, *100*, 4–17.

[6] Sanders, J. M.; Monogue, M. L.; Jodlowski, T. Z.; Cutrell, J. B. Pharmacologic treatments for coronavirus disease 2019 (COVID-19): a review. *JAMA* **2020**, *323*, 1824–1836.

[7] Liu, C.; Zhou, Q.; Li, Y.; Garner, L. V.; Watkins, S. P.; Carter, L. J.; Smoot, J.; Gregg, A. C.; Daniels, A. D.; Jervey, S.; Albaiu, D. Research and development on therapeutic agents and vaccines for COVID-19 and related human coronavirus diseases. *ACS Cent. Sci.* **2020**, *6*, 315–331.

[8] Le, T. T.; Andreadakis, Z.; Kumar, A.; Roman, R. G.; Tollefsen, S.; Saville, M.; Mayhew, S. The COVID-19 vaccine development landscape. *Nat. Rev. Drug Discov.* **2020**, *19*, 305–306.

[9] Erlanson, D. A. Many small steps towards a COVID-19 drug. *Nat. Commun.* **2020**, *11*, 5048.

[10] Arshadi, A. K.; Webb, J.; Salem, M.; Cruz, E.; Calad-Thomson, S.; Ghadirian, N.; Collins, J.; Diez-Cecilia, E.; Kelly, B.; Goodarzi, H.; Yuan, J. S. Artificial intelligence for COVID-19 drug discovery and vaccine development. *Front. Artif. Intell.* **2020**, *3*, 65.

[11] Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of early drug discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239–1249.

[12] Knowles, J.; Gromo, G. Target selection in drug discovery. *Nat. Rev. Drug Discov.* **2003**, *2*, 63–69.

[13] Schenone, M.; Dančík, V.; Wagner, B. K.; Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.* **2013**, *9*, 232–240.

[14] Kraus, V. B. Biomarkers as drug development tools: discovery, validation, qualification and use. *Nat. Rev. Rheumatol.* **2018**, *14*, 354–362.

[15] Tan, D. S. Diversity-oriented synthesis: exploring the intersections between chemistry and biology. *Nat. Chem. Biol.* **2005**, *1*, 74–84.

[16] Smith, A. Screening for drug discovery: the leading question. *Nature* **2002**, *418*, 453–455.

[17] Walters, W. P.; Patrick Walters, W.; Namchuk, M. Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* **2003**, *2*, 259–266.

[18] Keseru, G. M.; Makara, G. M. Hit discovery and hit-to-lead approaches. *Drug Discov. Today* **2006**, *11*, 741–748.

[19] Lipinski, C. A. Overview of hit to lead: the medicinal chemist's role from HTS retest to lead optimization hand off. *Top. Med. Chem.* **2009**, *5*, 1–24.

[20] Wermuth, C. G. *The practice of medicinal chemistry*; Academic Press, 2011.

[21] Meinert, C. L. *Clinical trials handbook: design and conduct*; John Wiley & Sons, 2012.

[22] Guarino, R. A.; Guarino, R. *New drug approval process*; CRC Press, 2016.

[23] Hay, M.; Thomas, D. W.; Craighead, J. L.; Economides, C.; Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **2014**, *32*, 40–51.

[24] Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **2004**, *3*, 1–5.

[25] Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **2010**, *9*, 203–214.

[26] Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discov.* **2015**, *14*, 475–486.

[27] Moos, W.; Miller, S.; Munk, S.; Munk, B. *Managing the drug discovery process: how to make it more efficient and cost-effective*; Woodhead Publishing, 2016.

[28] Bajorath, J. Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. *Drug Discov. Today* **2001**, *6*, 989–995.

[29] Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813–1818.

[30] Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr. Computational methods in drug discovery. *Pharmacol. Rev.* **2014**, *66*, 334–395.

[31] Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **2018**, *17*, 97–113.

[32] Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

[33] Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.

[34] De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem.* **2016**, *59*, 4035–4061.

[35] Gasteiger, J. Chemoinformatics: a new field with a long tradition. *Anal. Bioanal. Chem.* **2006**, *384*, 57–64.

[36] Engel, T. Basic overview of chemoinformatics. *J. Chem. Inf. Model.* **2006**, *46*, 2267–2277.

[37] Wooller, S. K.; Benstead-Hume, G.; Chen, X.; Ali, Y.; Pearl, F. M. G. Bioinformatics in translational drug discovery. *Biosci. Rep.* **2017**, *37*, BSR20160180.

[38] Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-driven medicinal chemistry in the era of big data. *Drug Discov. Today* **2014**, *19*, 859–868.

[39] Chen, H.; Kogej, T.; Engkvist, O. Cheminformatics in drug discovery, an industrial perspective. *Mol. Inf.* **2018**, *37*, e1800041.

[40] Griffen, E. J.; Dossetter, A. G.; Leach, A. G.; Montague, S. Can we accelerate medicinal chemistry by augmenting the chemist with Big Data and artificial intelligence? *Drug Discov. Today* **2018**, *23*, 1373–1384.

[41] Griffen, E. J.; Dossetter, A. G.; Leach, A. G. Chemists: AI is here; unite to get the benefits. *J. Med. Chem.* **2020**, *63*, 8695–8704.

[42] Kubinyi, H. Similarity and dissimilarity: a medicinal chemist's view. *Perspect. Drug Discov. Des.* **1998**, *9*, 225–252.

[43] Kubinyi, H. A general view on similarity and QSAR studies. *Comp. Assist. Lead Finding and Optim.* **1997**, 9–28.

[44] Lill, M. A. Multi-dimensional QSAR in drug discovery. *Drug Discov. Today* **2007**, *12*, 1013–1017.

[45] Cherkasov, A. et al. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **2014**, *57*, 4977–5010.

[46] Varnek, A.; Baskin, I. Machine learning methods for property prediction in chemoinformatics: quo vadis? *J. Chem. Inf. Model.* **2012**, *52*, 1413–1437.

[47] van Hilten, N.; Chevillard, F.; Kolb, P. Virtual compound libraries in computer-assisted drug discovery. *J. Chem. Inf. Model.* **2019**, *59*, 644–651.

[48] Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882–894.

[49] Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

[50] Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468–481.

[51] Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* **2015**, *20*, 318–331.

[52] Bajorath, J. Data analytics and deep learning in medicinal chemistry. *Future Med. Chem.* **2018**, *10*, 1541–1543.

[53] Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250.

[54] Lavecchia, A. Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug Discov. Today* **2019**, *24*, 2017–2032.

[55] Mak, K.-K.; Pichika, M. R. Artificial intelligence in drug development: present status and future prospects. *Drug Discov. Today* **2019**, *24*, 773–780.

[56] Reker, D.; Schneider, P.; Schneider, G.; Brown, J. B. Active learning for computational chemogenomics. *Future Med. Chem.* **2017**, *9*, 381–402.

[57] Blaschke, T.; Bajorath, J. Compound design using generative neural networks. *Artif. Intell. in Drug Discov.* **2020**, *75*, 217.

[58] Xue, D.; Gong, Y.; Yang, Z.; Chuai, G.; Qu, S.; Shen, A.; Yu, J.; Liu, Q. Advances and challenges in deep generative models for de novo molecule generation. *WIREs Comp. Mol. Sci.* **2019**, *9*, e1395.

[59] Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.

[60] Schwaller, P.; Laino, T. Data-driven learning systems for chemical reaction prediction: an analysis of recent approaches. Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions. 2019; pp 61–79.

[61] Johansson, S.; Thakkar, A.; Kogej, T.; Bjerrum, E.; Genheden, S.; Bastys, T.; Kannas, C.; Schliep, A.; Chen, H.; Engkvist, O. AI-assisted synthesis prediction. *Drug Discov. Today Technol.* **2020**, *32–33*, 65–72.

[62] Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.

[63] Bajorath, J.; Overington, J.; Jenkins, J. L.; Walters, P. Drug discovery and development in the era of Big Data. *Future Med. Chem.* **2016**, *8*, 1807–1813.

[64] Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.

[65] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

[66] The UniProt Consortium, UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.

[67] Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; Hersey, A.; Overington, J. P. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res.* **2016**, *44*, D1220–8.

[68] Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.

[69] Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102–D1109.

[70] Gaulton, A. et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.

[71] Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug–target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* **2006**, *5*, 730–739.

[72] Lu, H.; Tonge, P. J. Drug–target residence time: critical information for lead optimization. *Curr. Opin. Chem. Biol.* **2010**, *14*, 467–474.

[73] Morphy, J. R. The challenges of multi-target lead optimization. *Designing Multi-Target Drugs* **2012**, 141–154.

[74] Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: challenges and opportunities in drug discovery. *J. Med. Chem.* **2014**, *57*, 7874–7887.

[75] Schuster, D.; Laggner, C.; Langer, T. Why drugs fail – a study on side effects in new chemical entities. *Antitargets* **2008**, 1–22.

[76] Aronson, J. K. *Meyler's side effects of drugs: the international encyclopedia of adverse drug reactions and interactions*; Elsevier, 2015.

[77] Lucas, A. J.; Sproston, J. L.; Barton, P.; Riley, R. J. Estimating human ADME properties, pharmacokinetic parameters and likely clinical dose in drug discovery. *Expert Opin. Drug Discov.* **2019**, *14*, 1313–1327.

[78] Korfmacher, W. A. Lead optimization strategies as part of a drug metabolism environment. *Curr. Opin. Drug Discov. Devel.* **2003**, *6*, 481–485.

[79] Baillie, T. A. Metabolism and toxicity of drugs. Two decades of progress in industrial drug metabolism. *Chem. Res. Toxicol.* **2008**, *21*, 129–137.

[80] Korfmacher, W. A. Advances in the integration of drug metabolism into the lead optimization paradigm. *Mini Rev. Med. Chem.* **2009**, *9*, 703–716.

[81] Recanatini, M.; Bottegoni, G.; Cavalli, A. In silico antitarget screening. *Drug Discov. Today Technol.* **2004**, *1*, 209–215.

[82] Hessler, G.; Matter, H.; Schmidt, F.; Giegerich, C.; Wang, L.-H.; Güssregen, S.; Baringhaus, K.-H. Identification and application of antitarget activity hotspots to guide compound optimization. *Mol. Inf.* **2011**, *30*, 996–1008.

[83] Zakharov, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Quantitative prediction of antitarget interaction profiles for chemical compounds. *Chem. Res. Toxicol.* **2012**, *25*, 2378–2385.

[84] van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* **2003**, *2*, 192–204.

[85] Gleeson, M. P.; Paul Gleeson, M.; Hersey, A.; Hannongbua, S. In-silico ADME models: a general assessment of their utility in drug discovery applications. *Curr. Topics Med. Chem.* **2011**, *11*, 358–381.

[86] Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.* **2012**, *52*, 617–648.

[87] Kirchmair, J.; Göller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G. Predicting drug metabolism: experiment and/or computation? *Nat. Rev. Drug Discov.* **2015**, *14*, 387–404.

[88] Nicolotti, O.; Giangreco, I.; Introcaso, A.; Leonetti, F.; Stefanachi, A.; Carotti, A. Strategies of multi-objective optimization in drug discovery and development. *Expert Opin. Drug Discov.* **2011**, *6*, 871–884.

[89] Segall, M. D. Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Curr. Pharm. Des.* **2012**, *18*, 1292–1310.

[90] Nicolaou, C. A.; Brown, N. Multi-objective optimization methods in drug design. *Drug Discov. Today Technol.* **2013**, *10*, e427–e435.

[91] Cheshire, D. R. How well do medicinal chemists learn from experience? *Drug Discov. Today* **2011**, *16*, 817–821.

[92] Olesen, P. H. The use of bioisosteric groups in lead optimization. *Curr. Opin. Drug Discov. Devel.* **2001**, *4*, 471–478.

[93] Gedeck, P.; Lewis, R. A. Exploiting QSAR models in lead optimization. *Curr. Opin. Drug Discov. Devel.* **2008**, *11*, 569–575.

[94] Nicolaou, C. A.; Brown, N.; Pattichis, C. S. Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discov. Devel.* **2007**, *10*, 316–324.

[95] Segall, M. Advances in multiparameter optimization methods for de novo drug design. *Expert Opin. Drug Discov.* **2014**, *9*, 803–817.

[96] Tian, S.; Wang, J.; Li, Y.; Li, D.; Xu, L.; Hou, T. The application of in silico drug-likeness predictions in pharmaceutical research. *Adv. Drug Deliv. Rev.* **2015**, *86*, 2–10.

[97] Choy, Y. B.; Prausnitz, M. R. The rule of five for non-oral routes of drug delivery: ophthalmic, inhalation and transdermal. *Pharm. Res.* **2011**, *28*, 943–948.

[98] Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **2007**, *6*, 881–890.

[99] Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* **2004**, *1*, 337–341.

[100] Mignani, S.; Rodrigues, J.; Tomas, H.; Jalal, R.; Singh, P. P.; Majoral, J.-P.; Vishwakarma, R. A. Present drug-likeness filters in medicinal chemistry during the hit and lead optimization process: how far can they be simplified? *Drug Discov. Today* **2018**, *23*, 605–615.

[101] Nissink, J. W. M.; Degorce, S. Analyzing compound and project progress through multi-objective-based compound quality assessment. *Future Med. Chem.* **2013**, *5*, 753–767.

[102] Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. *ACS Chem. Neurosci.* **2010**, *1*, 435–449.

[103] Kruisselbrink, J. W.; Emmerich, M. T. M.; Bäck, T.; Bender, A.; IJzerman, A. P.; van der Horst, E. Combining aggregation with Pareto optimization: a case study in evolutionary molecular design. Evolutionary Multi-Criterion Optimization. 2009; pp 453–467.

[104] Segall, M. D.; Beresford, A. P.; Gola, J. M. R.; Hawksley, D.; Tarbit, M. H. Focus on success: using a probabilistic approach to achieve an optimal balance of compound properties in drug discovery. *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*, 325–337.

[105] Debe, D. A.; Mamidipaka, R. B.; Gregg, R. J.; Metz, J. T.; Gupta, R. R.; Muchmore, S. W. ALOHA: a novel probability fusion approach for scoring multi-parameter drug-likeness during the lead optimization stage of drug discovery. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 771–782.

[106] Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* **2004**, *9*, 430–431.

[107] Shultz, M. D. Setting expectations in molecular optimizations: Strengths and limitations of commonly used composite parameters. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 5980–5991.

[108] Hopkins, A. L.; Keserü, G. M.; Leeson, P. D.; Rees, D. C.; Reynolds, C. H. The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discov.* **2014**, *13*, 105–121.

[109] Cavalluzzi, M. M.; Mangiatordi, G. F.; Nicolotti, O.; Lentini, G. Ligand efficiency metrics in drug discovery: the pros and cons from a practical perspective. *Expert Opin. Drug Discov.* **2017**, *12*, 1087–1104.

[110] Keserü, G. M.; Makara, G. M. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov.* **2009**, *8*, 203–212.

[111] Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90–98.

[112] Yusof, I.; Segall, M. D. Considering the impact drug-like properties have on the chance of success. *Drug Discov. Today* **2013**, *18*, 659–666.

[113] Shultz, M. D. Two decades under the influence of the rule of five and the changing properties of approved oral drugs. *J. Med. Chem.* **2018**, *62*, 1701–1714.

[114] Kenny, P. W.; Montanari, C. A. Inflation of correlation in the pursuit of drug-likeness. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 1–13.

[115] Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. *Chem. Biol.* **2014**, *21*, 1115–1142.

[116] Doak, B. C.; Kihlberg, J. Drug discovery beyond the rule of 5 - opportunities and challenges. *Expert Opin. Drug Discov.* **2017**, *12*, 115–119.

[117] DeGoey, D. A.; Chen, H.-J.; Cox, P. B.; Wendt, M. D. Beyond the rule of 5: lessons learned from AbbVie's drugs and compound collection. *J. Med. Chem.* **2018**, *61*, 2636–2651.

[118] Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861.

[119] Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824–828.

[120] Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823–823.

[121] Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

[122] Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675–679.

[123] Maggiora, G. M.; Bajorath, J. Chemical space networks: a powerful new paradigm for the description of chemical space. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 795–802.

[124] Oprea, T. I.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3*, 157–166.

[125] Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J.-L.; Chen, H.; Engkvist, O. Exploring the GDB-13 chemical space using deep generative models. *J. Cheminf.* **2019**, *11*, 20.

[126] Reymond, J.-L.; van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm.* **2010**, *1*, 30–38.

[127] Reymond, J.-L. The chemical space project. *Acc. Chem. Res.* **2015**, *48*, 722–730.

[128] Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. The proximal Lilly collection: mapping, exploring and exploiting feasible chemical space. *J. Chem. Inf. Model.* **2016**, *56*, 1253–1266.

[129] Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to explore chemical space using algorithms and automation. *Nat. Rev. Chem.* **2019**, *3*, 119–128.

[130] Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269–1275.

[131] Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* **2004**, *47*, 4891–4896.

[132] Bonnet, P. Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? A comparative assessment between medicinal and computational chemists. *Eur. J. Med. Chem.* **2012**, *54*, 679–689.

[133] Baber, J. C.; Feher, M. Predicting synthetic accessibility: application in drug discovery and development. *Mini Rev. Med. Chem.* **2004**, *4*, 681–692.

[134] Boda, K.; Seidel, T.; Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 311–325.

[135] Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.

[136] Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.* **2005**, *4*, 649–663.

[137] Schneider, G. De novo design – hop(p)ing against hope. *Drug Discov. Today Technol.* **2013**, *10*, e453–e460.

[138] Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.-H.; Schneider, G.; Jacoby, E.; Renner, S. A collection of robust organic synthesis reactions for in silico molecule design. *J. Chem. Inf. Model.* **2011**, *51*, 3093–3098.

[139] Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: reaction-driven

de novo design of bioactive compounds. *PLoS Comput. Biol.* **2012**, *8*, e1002380.

[140] Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.-H.; Schneider, G.; Jacoby, E.; Renner, S. Probing the bioactivity-relevant chemical space of robust reactions and common molecular building blocks. *J. Chem. Inf. Model.* **2012**, *52*, 1167–1178.

[141] Ujváry, I.; Hayward, J. Bioster: a database of bioisosteres and bioanalogues. Bioisosteres in medicinal chemistry. 2012; pp 53–74.

[142] Stewart, K. D.; Shiroda, M.; James, C. A. Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.* **2006**, *14*, 7011–7022.

[143] Raymond, J. W.; Watson, I. A.; Mahoui, A. Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *J. Chem. Inf. Model.* **2009**, *49*, 1952–1962.

[144] Nicolaou, C. A.; Apostolakis, J.; Pattichis, C. S. De novo drug design using multiobjective evolutionary graphs. *J. Chem. Inf. Model.* **2009**, *49*, 295–307.

[145] Schneider, G.; Hartenfeller, M.; Reutlinger, M.; Tanrikulu, Y.; Proschak, E.; Schneider, P. Voyages to the (un)known: adaptive design of bioactive compounds. *Trends Biotechnol.* **2009**, *27*, 18–26.

[146] Nicolaou, C.; Kannas, C.; Loizidou, E. Multi-objective optimization methods in de novo drug design. *Mini Rev. Med. Chem.* **2012**, *12*, 979–987.

[147] Yoshikawa, N.; Terayama, K.; Sumita, M.; Homma, T.; Oono, K.; Tsuda, K. Population-based de novo molecule generation, using grammatical evolution. *Chem. Lett.* **2018**, *47*, 1431–1434.

[148] van der Horst, E.; Marqués-Gallego, P.; Mulder-Krieger, T.; van Veldhoven, J.; Kruisselbrink, J.; Aleman, A.; Emmerich, M. T. M.; Brussee, J.; Bender, A.; Ijzerman, A. P. Multi-objective evolutionary design of adenosine receptor ligands. *J. Chem. Inf. Model.* **2012**, *52*, 1713–1721.

[149] Besnard, J. et al. Automated design of ligands to polypharmacological profiles. *Nature* **2012**, *492*, 215–220.

[150] Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR analysis for chemical structure generation (from y to x). *J. Chem. Inf. Model.* **2016**, *56*, 286–299.

[151] Miyao, T.; Arakawa, M.; Funatsu, K. Exhaustive structure generation for inverse-QSPR/QSAR. *Mol. Inf.* **2010**, *29*, 111–125.

[152] Schneider, P.; Schneider, G. De novo design at the edge of chaos. *J. Med. Chem.* **2016**, *59*, 4077–4086.

[153] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

[154] Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screen.* **2000**, *3*, 363–372.

[155] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

[156] Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7*, 23.

[157] O'Boyle, N. M. Towards a universal SMILES representation - a standard method to generate canonical SMILES based on the InChI. *J. Cheminf.* **2012**, *4*, 22.

[158] Schneider, N.; Sayle, R. A.; Landrum, G. A. Get your atoms in order – an open-source implementation of a novel and robust molecular canonicalization algorithm. *J. Chem. Inf. Model.* **2015**, *55*, 2111–2120.

[159] David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminf.* **2020**, *12*, 56.

[160] Langer, T.; Hoffmann, R. D. Pharmacophore modelling: applications in drug discovery. *Expert Opin. Drug Discov.* **2006**, *1*, 261–267.

[161] Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.

[162] Brugger, W. E.; Stuper, A. J.; Jurs, P. C. Generation of descriptors from molecular structures. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 105–110.

[163] Glen, R. C.; Rose, V. S. Computer program suite for the calculation, storage and manipulation of molecular property and activity descriptors. *J. Mol. Graph.* **1987**, *5*, 79–86.

[164] Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. A neural device for searching direct correlations between structures and properties of chemical compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 715–721.

[165] Merkwirth, C.; Lengauer, T. Automatic generation of complementary descriptors with molecular graph networks. *J. Chem. Inf. Model.* **2005**, *45*, 1159–1168.

[166] Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; John Wiley & Sons, 2008.

[167] Grisoni, F.; Consonni, V.; Todeschini, R. Impact of molecular descriptors on computational models. *Methods Mol. Biol.* **2018**, *1825*, 171–209.

[168] Prasanna, S.; Doerksen, R. J. Topological polar surface area: a useful descriptor in 2D-QSAR. *Curr. Med. Chem.* **2009**, *16*, 21–41.

[169] Leo, A.; Jow, P. Y. C.; Silipo, C.; Hansch, C. Calculation of hydrophobic constant (log P) from π and f constants. *J. Med. Chem.* **1975**, *18*, 865–868.

[170] Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006**, *11*, 1046–1053.

[171] Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminf.* **2013**, *5*, 26.

[172] Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.* **2009**, *49*, 108–119.

[173] Maccs, K. MDL Information Systems. *Inc. : San Leandro, CA* **1984**,

[174] Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

[175] Heikamp, K.; Bajorath, J. Fingerprint design and engineering strategies: rationalizing and improving similarity search performance. *Future Med. Chem.* **2012**, *4*, 1945–1959.

[176] Bajorath, J. Improving the utility of molecular scaffolds for medicinal and computational chemistry. *Future Med. Chem.* **2018**, *10*, 1645–1648.

[177] Hu, Y.; Stumpfe, D.; Bajorath, J. Recent advances in scaffold hopping. *J. Med. Chem.* **2017**, *60*, 1238–1246.

[178] Zhao, H.; Dietrich, J. Privileged scaffolds in lead generation. *Expert Opin. Drug Discov.* **2015**, *10*, 781–790.

[179] Müller, G. Medicinal chemistry of target family-directed masterkeys. *Drug Discov. Today* **2003**, *8*, 681–691.

[180] Bajorath, J. Large-scale SAR analysis. *Drug Discov. Today: Technol.* **2013**, *10*, e419–e426.

[181] Peltason, L.; Bajorath, J. Systematic computational analysis of structure–activity relationships: concepts, challenges and recent advances. *Future Med. Chem.* **2009**, *1*, 451–466.

[182] Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discov. Today* **2010**, *15*, 630–639.

[183] Stumpfe, D.; Bajorath, J. Recent developments in SAR visualization. *MedChemComm.* **2016**, *7*, 1045–1055.

[184] Peltason, L.; Bajorath, J. SAR index: quantifying the nature of structure-activity relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.

[185] Bajorath, J. Evolution of the activity cliff concept for structure–activity relationship analysis and drug discovery. *Future Med. Chem.* **2014**, *6*, 1545–1549.

[186] Stumpfe, D.; Hu, H.; Bajorath, J. Advances in exploring activity cliffs. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 929–942.

[187] Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

[188] Bajorath, J. Molecular similarity concepts for informatics applications. *Methods Mol. Biol.* **2017**, *1526*, 231–245.

[189] Stumpfe, D.; Bajorath, J. Similarity searching. *WIREs Comput. Mol. Sci.* **2011**, *1*, 260–282.

[190] Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* **2014**, *57*, 3186–3204.

[191] Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *J. Chem. Inf. Model.* **2012**, *52*, 2884–2901.

[192] Rogers, D. J.; Tanimoto, T. T. A computer program for classifying plants. *Science* **1960**, *132*, 1115–1118.

[193] Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*, 20.

[194] Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **2007**, *12*, 225–233.

[195] Wawer, M.; Bajorath, J. Local structural changes, global data views: graphical substructure-activity relationship trailing. *J. Med. Chem.* **2011**, *54*, 2944–2951.

[196] Kenny, P. W.; Sadowski, J. Structure modification in chemical databases. *Chemoinform. in Drug Discov.* **2005**, *23*, 271–285.

[197] Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* **2011**, *54*, 7739–7750.

[198] Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched molecular pair analysis in drug discovery. *Drug Discov. Today* **2013**, *18*, 724–731.

[199] Dalke, A.; Hert, J.; Kramer, C. mmpdb: An open-source matched molecular pair platform for large multiproperty data sets. *J. Chem. Inf. Model.* **2018**, *58*, 902–910.

[200] Hussain, J.; Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348.

[201] de la Vega de León, A.; Bajorath, J. Matched molecular pairs derived by retrosynthetic fragmentation. *MedChemComm* **2014**, *5*, 64–67.

[202] Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR maps: a new SAR visualization technique for medicinal chemists. *J. Med. Chem.* **2007**, *50*, 5926–5937.

[203] Wassermann, A. M.; Bajorath, J. Directed R-group combination graph: a methodology to uncover structure–activity relationship patterns in a series of analogues. *J. Med. Chem.* **2012**, *55*, 1215–1226.

[204] Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR matrices: automated extraction of information-rich SAR tables from large compound data sets. *J. Chem. Inf. Model.* **2012**, *52*, 1769–1776.

[205] Gupta-Ostermann, D.; Bajorath, J. The 'SAR Matrix' method and its extensions for applications in medicinal chemistry and chemogenomics. *F1000 Res.* **2014**, *3*, 113.

[206] Yoshimori, A.; Bajorath, J. The SAR matrix method and an artificially intelligent variant for the identification and structural organization of analog series, SAR analysis, and compound design. *Mol. Inf.* **2020**, *39*, e2000045.

[207] Hu, Y.; Stumpfe, D.; Bajorath, J. Computational exploration of molecular scaffolds in medicinal chemistry. *J. Med. Chem.* **2016**, *59*, 4062–4076.

[208] Stumpfe, D.; Dimova, D.; Bajorath, J. Computational method for the systematic identification of analog series and key compounds representing series and their biological activity profiles. *J. Med. Chem.* **2016**, *59*, 7667–7676.

[209] Dimova, D.; Stumpfe, D.; Hu, Y.; Bajorath, J. Analog series-based scaffolds: computational design and exploration of a new type of molecular scaffolds for medicinal chemistry. *Future Sci. OA* **2016**, *2*, FSO149.

[210] Dimova, D.; Stumpfe, D.; Bajorath, J. Computational design of new molecular scaffolds for medicinal chemistry, part II: generalization of analog series-based scaffolds. *Future Sci. OA* **2018**, *4*, FSO267.

[211] Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-Franco, J. L.; Bajorath, J. Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS Omega* **2019**, *4*, 1027–1032.

[212] Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. 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.* **1998**, *38*, 511–522.

[213] Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.

[214] Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-activity relationship anatomy by network-like similarity graphs

and local structure-activity relationship indices. *J. Med. Chem.* **2008**, *51*, 6075–6084.

[215] Free, S. M., Jr; Wilson, J. W. A mathematical contribution to structure-activity studies. *J. Med. Chem.* **1964**, *7*, 395–399.

[216] Kubinyi, H. Free Wilson analysis. Theory, applications and its relationship to Hansch analysis. *Quant. Struct.-Act. Relat.* **1988**, *7*, 121–133.

[217] Gupta-Ostermann, D.; Shanmugasundaram, V.; Bajorath, J. Neighborhood-based prediction of novel active compounds from SAR matrices. *J. Chem. Inf. Model.* **2014**, *54*, 801–809.

[218] Trevor, H.; Robert, T.; Jerome, F. *The elements of statistical learning*; Springer Science+ Business Media, LLC, 2009.

[219] Baumann, D.; Baumann, K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J. Cheminf.* **2014**, *6*, 47.

[220] Krstajic, D.; Buturovic, L. J.; Leahy, D. E.; Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminf.* **2014**, *6*, 10.

[221] Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.

[222] Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

[223] Vapnik, V. *The nature of statistical learning theory*; Springer Science & Business Media, 2013.

[224] Maltarollo, V. G.; Kronenberger, T.; Espinoza, G. Z.; Oliveira, P. R.; Honorio, K. M. Advances with support vector machines for novel drug discovery. *Expert Opin. Drug Discov.* **2019**, *14*, 23–33.

[225] LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.

[226] Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263–274.

[227] Marquardt, D. W.; Snee, R. D. Ridge regression in practice. *Am. Stat.* **1975**, *29*, 3–20.

[228] Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.

[229] Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Netw.* **2005**, *18*, 1093–1110.

[230] Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. In *Advances in Neural Information Processing Systems*; Mozer, M. C., Jordan, M., Petsche, T., Eds.; MIT Press, 1997; Vol. 9; pp 155–161.

[231] Smola, A. J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222.

[232] Heikamp, K.; Bajorath, J. Support vector machines for drug discovery. *Expert Opin. Drug Discov.* **2014**, *9*, 93–104.

[233] Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **2015**, *61*, 85–117.

[234] Xu, Y.; Lin, K.; Wang, S.; Wang, L.; Cai, C.; Song, C.; Lai, L.; Pei, J. Deep learning for molecular generation. *Future Med. Chem.* **2019**, *11*, 567–597.

[235] Segler, M. H. S.; Waller, M. P. Modelling chemical reasoning to predict and invent reactions. *Chemistry* **2017**, *23*, 6118–6128.

[236] Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **2019**, *9*, 10752.

[237] Ståhl, N.; Falkman, G.; Karlsson, A.; Mathiason, G.; Boström, J. Deep reinforcement learning for multiparameter optimization in de novo drug design. *J. Chem. Inf. Model.* **2019**, *59*, 3166–3176.

[238] Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? Advances in Neural Information Processing Systems. 2014; pp 3320–3328.

[239] Peters, M. E.; Ruder, S.; Smith, N. A. To tune or not to tune? Adapting pretrained representations to diverse tasks. *arXiv* **2019**, 1903.05987.

[240] Yuan, W.; Jiang, D.; Nambiar, D. K.; Liew, L. P.; Hay, M. P.; Bloomstein, J.; Lu, P.; Turner, B.; Le, Q.-T.; Tibshirani, R.; Khatri, P.; Moloney, M. G.; Koong, A. C. Chemical space mimicry for drug discovery. *J. Chem. Inf. Model.* **2017**, *57*, 875–882.

[241] Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.

[242] Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* **2018**, *37*.

[243] Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of generative autoencoder in de novo molecular design. *Mol. Inf.* **2018**, *37*, 1700123.

[244] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Centr. Sci.* **2018**, *4*, 268–276.

[245] Jin, W.; Barzilay, R.; Jaakkola, T. Chapter 11. junction tree variational autoencoder for molecular graph generation. *Artif. Intell. in Drug Discov.* **2020**, 228–249.

[246] Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. A de novo molecular generation method using latent vector based generative adversarial network. *J. Cheminf.* **2019**, *11*, 74.

[247] Vanhaelen, Q.; Lin, Y.-C.; Zhavoronkov, A. The advent of generative chemistry. *ACS Med. Chem. Lett.* **2020**, *11*, 1496–1505.

[248] Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative recurrent networks for de novo drug design. *Mol. Inf.* **2018**, *37*, 1700111.

[249] Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: an AI tool for de novo drug design. *J. Chem. Inf. Model.* **2020**, *60*, 5918–5922.

[250] Mercado, R.; Rastemo, T.; Lindelöf, E.; Klambauer, G.; Engkvist, O.; Chen, H.; Bjerrum, E. J. Graph networks for molecular design. *Mach. Learn. Sci. Technol.* **2020**,

[251] Xia, X.; Hu, J.; Wang, Y.; Zhang, L.; Liu, Z. Graph-based generative models for de novo drug design. *Drug Discov. Today Technol.* **2020**, *32–33*, 45–53.

[252] Arús-Pous, J.; Patronov, A.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminf.* **2020**, *12*, 38.

[253] Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, eaap7885.

[254] Nielsen, M. A. *Neural networks and deep learning*; Determination press San Francisco, CA, 2015; Vol. 2018.

[255] Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep learning*; MIT press Cambridge, 2016; Vol. 1.

[256] Dreyfus, S. The computational solution of optimal control problems with time lag. *IEEE Trans. Automat. Contr.* **1973**, *18*, 383–385.

[257] Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.

[258] Jozefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; Wu, Y. Exploring the limits of language modeling. *arXiv* **2016**, 1602.02410.

[259] Graves, A.; Eck, D.; Beringer, N.; Schmidhuber, J. Biologically plausible speech recognition with LSTM neural nets. Biologically Inspired Approaches to Advanced Information Technology. 2004; pp 127–136.

[260] Gers, F. A.; Schmidhuber, E. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Trans. Neural Netw.* **2001**, *12*, 1333–1340.

[261] Bhoopchand, A.; Rocktäschel, T.; Barr, E.; Riedel, S. Learning Python code suggestion with a sparse pointer network. *arXiv* **2016**, 1611.08307.

[262] Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.

[263] Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* **2014**,

[264] Blaschke, T.; Engkvist, O.; Bajorath, J.; Chen, H. Memory-assisted reinforcement learning for diverse molecular de novo design. *J. Cheminf.* **2020**, *12*, 68.

[265] Yoshimori, A.; Bajorath, J. Deep SAR matrix: SAR matrix expansion for advanced analog design using deep learning architectures. *Future Drug Discov.* **2020**, *2*, FDD36.

[266] Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminf.* **2017**, *9*, 48.

[267] Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminf.* **2019**, *11*, 71.

[268] Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2020**, *2*, 171–180.

[269] Kotsias, P.-C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E. J. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* **2020**, *2*, 254–265.

[270] Williams, R. J.; Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1989**, *1*, 270–280.

[271] Goldberg, Y. A primer on neural network models for natural language processing. *J. Artif. Intell. Res.* **2016**, *57*, 345–420.

[272] Munson, M.; Lieberman, H.; Tserlin, E.; Rocnik, J.; Ge, J.; Fitzgerald, M.; Patel, V.; Garcia-Echeverria, C. Lead optimization attrition analysis (LOAA): a novel and general methodology for medicinal chemistry. *Drug Discov. Today* **2015**, *20*, 978–987.

[273] Maynard, A. T.; Roberts, C. D. Quantifying, visualizing, and monitoring lead optimization. *J. Med. Chem.* **2016**, *59*, 4189–4201.

[274] Shanmugasundaram, V.; Zhang, L.; Kayastha, S.; de la Vega de León, A.; Dimova, D.; Bajorath, J. Monitoring the progression of structure–activity relationship information during lead optimization. *J. Med. Chem.* **2016**, *59*, 4235–4244.

[275] Kunimoto, R.; Miyao, T.; Bajorath, J. Computational method for estimating progression saturation of analog series. *RSC Adv.* **2018**, *8*, 5484–5492.

Additional publications

Yonchev, D.; Vogt, M.; Bajorath, J. From SAR diagnostics to compound design: development chronology of the compound optimization monitor (COMO) method. *Mol. Inf.* **2020**, *39*, e2000046.

Yonchev, D.; Bajorath, J. Inhibitor bias in luciferase-based luminescence assays. *Future Sci. OA* **2020**, FSO594.

Yonchev, D.; Dimova, D.; Stumpfe, D.; Vogt, M.; Bajorath, J. Redundancy in two major compound databases. *Drug Discov. Today* **2018**, *23*, 1183-1186.

Galati, S.; Yonchev, D.; Rodríguez Pérez R.; Vogt, M.; Tuccinardi, T.; Bajorath, J. Predicting isoform-selective carbonic anhydrase inhibitors via machine learning and rationalizing structural features important for selectivity. *ACS Omega* **2021**, *6*, 4080-4089.

Feldmann C.; Yonchev D.; Bajorath J. Structured data sets of compounds with multi-target and corresponding single-target activity from biological assays. *Future Sci. OA*, in press

Feldmann, C.; Yonchev, D.; Bajorath, J. Analysis of biological screening compounds with single- or multi-target activity via diagnostic machine learning. *Biomolecules* **2020**, *10*, e1605.

Feldmann, C.; Yonchev, D.; Stumpfe, D.; Bajorath, J. Systematic data analysis and diagnostic machine learning reveal differences between compounds with single- and multi-target activity. *Mol. Pharmaceutics* **2020**, *17*, 4652-4666.

Feldmann, C.; Miljković, F.; Yonchev, D.; Bajorath, J. Identifying promiscuous compounds with activity against different target classes. *Molecules* **2019**, *24*, e4185.

Curriculum Vitae

Eidesstattliche Erklärung

An Eides statt versichere ich hiermit, dass ich die Dissertation “Development of Computational Methods for Rationalizing Chemical Lead Optimization and Compound Design” selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch an keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nachstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist:

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* **2018**, *3*, 15799-15808.

Vogt, M.; Yonchev D.; Bajorath, J. Computational method to evaluate progress in lead optimization. *J. Med. Chem.* **2018**, *61*, 10895-10900.

Yonchev D.; Vogt, M.; Bajorath, J. Compound optimization monitor (COMO) method for computational evaluation of progress in medicinal chemistry projects. *Future Drug Discov.* **2019**, *1*, FDD15.

Yonchev D.; Bajorath, J. Integrating computational lead optimization diagnostics with analog design and candidate selection. *Future Sci. OA* **2020**, *6*, FSO451.

Yonchev D.; Bajorath, J. DeepCOMO: from structure-activity relationship diagnostics to generative molecular design using the compound optimization monitor methodology. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 1207-1218.

Dimitar Yonchev
März 2021
Bonn