DISS. ETH NO. 21732

Adaptive combinatorial *de novo* design of multi-target modulating compounds

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

MICHAEL ROBERT REUTLINGER

Dipl.-Bioinf., Goethe University Frankfurt am Main

born on 07.02.1978

citizen of Germany

accepted on the recommendation of

Prof. Dr. Gisbert Schneider, examiner

Prof. Dr. Gerd Folkers, co-examiner

2014

Table of Contents

Abbreviations

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
5-LO	5-lipoxygenase
ACD	Available Chemicals Directory
ACE	Angiotensin converting enzyme
ACO	Ant Colony Optimization
ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
ANN	Artificial Neural Network
BEDROC	Boltzmann-Enhanced Discrimination of ROC
BIN	Bayesian Interference Network
BKD	Binary Kernel Discrimination
CATS	Chemically Advanced Template Search
CB-1	Cannabonoid-1
CDK2	Cycline-dependent kinase 2
ChemGPS	Chemical Global Positioning System
CMM	Correlative Matrix Mapping
COBRA	Collection Of Bioactive Reference Compounds
CoMFA	Comparative Molecular Fields Analysis
EC_{50}	Half maximal Effective Concentration
ECFP	Extended-Connectivity FingerPrints
EF	Enrichment Factor
FDA	Food and Drug Administration
GP	Gaussian Process
GPCR	G-protein-coupled Receptor
GTM	Generative Topographic Mapping
hERG	Human Ether-á-go-go Related Gene
HGTM	Hierarchical Generative Topographic Mapping
HRMS	High Resolution Mass Spectrometry
hSST5R	Human somatostatin receptor subtype 5 receptor
HTS	High-Throughput Screening
IC_{50}	Half maximal Inhibitory Concentration
ISPE	Isometric Stochastic Proximity Embedding
IUPAC	International Union of Pure and Applied Chemistry
KL	Kullback-Leibler
kNN	k-Nearest Neighbor
LCMC	Local Continuity Meta-Criterion
LiSARD	Ligand-induced Structure-Activity Relationship Display
LLE	Local Linear Embedding
M1	Muscarinic receptor 1
MAE	Mean Absolute Error
MAntA	Molecular Ant Algorithm

MAPK	Mitogen-activated protein kinase
MCMC	Markov Chain Monte Carlo
MCS	Maximum Common Substructure
MD	Molecular Dynamics
MhD	Mahalanobis Distance
MDDR	MDL Drug Data Report
MDS	Multi Dimensional Scaling
mGluR	Metabotropic glutamate receptor
MHC	Major histocompatibility complex
MMAS	MAX-MIN Ant System
MRRE	Mean Relative Rank Error
MTNR	Melatonin receptor
NCE	New Chemical Entity
NEXOM	Neighbor Embedding eXploratory Observation Machine
NLM	Nonlinear Mapping
NQO2	NRH:quinone oxidoreductase 2
NSG	Network-like Similarity Graph
PAINS	Pan Assay Interference Compounds
PCA	Principal Component Analysis
PDB	Protein Data Bank
PI3K	Phosphoinositide 3-kinase
PLS	Projection to Latent Structures
PPAR	Proliferator-activated receptor
PPP	Potential Pharmacophoric Points
QED	Quantitative Estimate of Drug-Likeness
QM/MM	Quantum Mechanics / Molecular Mechanics
QSAR	Quantitative Structure-Activity Relationship
RBF	Radial Basis Function
RECAP	REtrosynthetic Combinatorial Analysis Procedure
REOS	Rapid Elimination Of Swill
ROC	Receiver Operating Characteristic
SAR	Structure-Activity Relationship
SAS	Structure-Activity Similarity
SC_{50}	Half maximal Stabilizing Concentration
SEA	Similarity Ensemble Approach
sEH	Soluble epoxide hydrolase
SHBG	Sex hormone binding globulin
SNE	Stochastic Neighbor Embedding
SOM	Self-Organizing Map
SPE	Stochastic Proximity Embedding
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbor Embedding
TSP	Traveling Salesman Problem
VC	Vapnik-Chervonenkis
VDW	Van der Waals
WDI	World Drug Index

Summary

The goal of computer-aided drug discovery is the identification of *New Chemical Entities* (NCEs) that exhibit a desired therapeutic effect. Selectively inhibiting a single disease-related target (most often a protein) using small molecules has been in focus of drug discovery for the last decades in order to maximize efficacy while minimizing undesired side effects through off-target modulation. However, with the increasing awareness that many diseases are polygenic, often involving cross-connected signaling cascades, the necessity for drugs that simultaneously modulate multiple macromolecular targets has become evident. Carefully balancing the effects on multiple therapeutic targets while avoiding side effect-related targets is generally anticipated to lead to improved drug efficacy and increased safety. Another opportunity closely related to polypharmacology is the repurposing of approved drugs for new therapeutic indications. Computational methods are routinely used in drug discovery projects to identify selective hit- and lead-compounds with optimized pharmacokinetic and safety profiles. In this context, machine-learning models have demonstrated their possession of high prediction quality and allowance for an enrichment of promising compounds. However, they have yet to demonstrate their broad applicability to polypharmacology.

In this thesis, a new method for computer-based *de novo* design of drug candidates with desired multi-target profiles is proposed. The *Molecular Ant Algorithm* (MAntA) features a fragment-based strategy to suggest innovative molecular structures. The quality of these candidates is assessed using Gaussian process regression models built for 640 macromolecular drug targets based on a curated subset of the ChEMBL database. Two molecular representations, topological CATS2 pharmacophore descriptors and ECFP-like circular Morgan fingerprints, are fused in a single kernel function to capture complementary aspects of molecular representation. A molecule construction procedure incorporating a library of readily available building blocks and economically sustainable combinatorial reactions enhances the synthetically accessibility of the proposed designs. To navigate the potentially huge combinatorial space, a nature-inspired *Ant Colony Optimization* (ACO) algorithm has been adapted to combinatorial small-molecule design. Using the Gaussian process models as fitness functions, ACO adaptively proposes candidate compounds according to the desired

objectives and is able to enumerate an objective-specific focused library of readily synthetically accessible compounds. To visually assist medicinal chemists in the molecular design process, MAntA has been equipped with an intuitive chemical space visualization method. Each of the four underlying algorithmic components was individually assessed before ultimately combining all parts of the approach and applying MAntA prospectively in a comprehensive molecular design study.

1) Three-dimensional (3D) landscape visualization (LiSARD), utilizing a neighborhood preserving dimensionality reduction algorithm and an adaptive Gaussian kernel smoother, is introduced as an intuitive method to visually analyze large sets of heterogeneous compound data. In a first application, the progress of a "real world" drug discovery project dataset was visualized for several project stages. Project-relevant areas in chemical space were already identified in the first stage and confirmed in later stages, emphasizing the benefits for hit prioritization and progress monitoring. Additionally, multi-objective landscapes were introduced as visual aid for multi-objective compound design.
2) An extension of the topological pharmacophore descriptor (CATS) with aromatic features (CATS2) was introduced. Retrospective analysis confirmed improved enrichment of bioactive compounds for the CATS2 descriptor while retaining scaffold-hopping potential. In a preliminary prospective study, the viability of CATS2 for target prediction was successfully confirmed in combination with self-organizing maps. Two hitherto unknown targets were discovered for compounds taken from a focused combinatorial library.
3) By utilizing a robust cross-validation scheme, multi-target Gaussian process regression models were evaluated for their ability to accurately predict ligand binding affinity and their discriminative power to separate binders and non-binders. The approach was prospectively applied in combination with on-chip chemical synthesis to generate a focused combinatorial library containing a privileged GPCR scaffold. Four new GPCR targets for this molecular framework were revealed, and 71% of the tested compounds were found to be active as predicted.

4) The potential use of ACO for combinatorial compound construction was evaluated in two scenarios. Its ability to efficiently identify "activity islands" in vast combinatorial space was demonstrated by sampling clusters of major histocompatibility complex class I (MHC-1) binding peptides from the complete octapeptide space. In this study, ACO was coupled to an ensemble machine-learning approach for predicting peptide binding to MHC-1. The ACO technique was further extended to arbitrary combinatorial reactions and retrospectively evaluated for adaptive combinatorial library design employing a virtual assay system as fitness function. With the constrain of limited numbers of tests cycles, ACO was able to efficiently enrich desired compounds and generally outperformed other optimization methods previously studied in the same context.

Finally, the MAntA concept was experimentally validated in two prospective *de novo* design projects with a focus on high-profile drug targets involved in neuropsychiatric disorders. In these proof-of-concept studies, the reaction scheme used to construct new candidate compounds was restricted to the reductive amination reaction. MAntA was applied to generate sigma-1 receptor selective ligands and multi-target modulating dopamine D_4 antagonists. For each scenario, compounds were selected for synthesis according to their predicted potency, selectivity, and exploration of chemical space. LiSARD multi-target landscape visualization was employed to depict the preferred design areas and localize the designed compounds in chemical space. The 16 selected designs were readily synthesizable, and the quantitative affinity predictions were biochemically confirmed. Overall, a success rate of 90% was achieved, and the designed compounds possess lead-like properties. This result demonstrates the potential of the MAntA concept for the design of synthetically accessible compounds that accurately match a predicted multi-target activity profile.

Zusammenfassung

Das Ziel des computergestützten *de novo* Designs ist die Identifikation von neuartigen chemischen Substanzen (*New Chemical Entity,* NCE), welche einen gewünschten therapeutischen Effekt erzielen können. In den letzten Jahrzehnten wurden primär niedermolekulare Wirkstoffe entwickelt, die einzelne krankheitsrelevante Zielmoleküle (zumeist ein Protein, *Target*) selektiv inhibieren, um so die Wirksamkeit zu maximieren und gleichzeitig mögliche Nebeneffekte zu minimieren. Es beginnt sich jedoch die Erkenntnis durchzusetzen, dass viele Krankheiten polygener Natur sind und mehrere vernetzte regulatorische Netzwerke an der phänotypischen Ausprägung involviert sind. Daher steigt der Bedarf an Wirkstoffen, die gleichzeitig mehrere makromolekulare Zielmoleküle modulieren. Es ist erforderlich, die Wirkung auf die unterschiedlichen Zielmoleküle genau auszutarieren, um eine hohe Wirksamkeit zu erreichen und gleichzeitig die Sicherheit des Wirkstoffs zu erhöhen. Darüber hinaus ist es möglich, neue therapeutische Indikationen für bereits etablierte Wirkstoffe zu finden. Computergestützte Methoden werden heutzutage routinemäßig eingesetzt, um neue selektive Wirkstoffkandidaten mit optimierter Pharmakokinetik zu entdecken. In diesem Zusammenhang sind insbesondere maschinelle Lernverfahren interessant, da diese bereits gezeigt haben, dass sie eine hohe Vorhersagekraft besitzen und zu einer Anreicherung von Erfolg versprechenden Molekülen beitragen können. Es muss sich jedoch noch erweisen, ob diese Methoden auch geeignet sind, um neuartige Substanzen gemäß eines polypharmakologischen Profils zu entwerfen.

In dieser Arbeit wird eine neue Methode zum computergestützten *de novo* Design von Wirkstoffkandidaten vorgestellt. Die Wirkstoffkandidaten werden so entworfen, dass sie ein gewünschtes polypharmakologisches Profil adressieren. Der molekulare Ameisen Algorithmus (*Molecular Ant Algorithm*, MAntA) verwendet eine fragment-basierte Strategie für die Konstruktion der neuartigen Molekülstrukturen. Die Qualität der Molekülkandidaten wird mit Gauss'schen Prozess Regressions Modellen bewertet. Basierend auf einer Fehler bereinigten Teilmenge der ChEMBL Datenbank wurden Modelle für insgesamt 640 makromolekulare Zielproteine erstellt. Hierbei wurden zwei unterschiedliche molekulare Repräsentationen verwendet: der topologische CATS2 Pharmakophordeskriptor und der ECFP-ähnliche binäre

topologische Morgan Fingerabdruck. Diese wurden in einer Kernfunktion vereinigt, wodurch sich ergänzende molekulare Repräsentationen im Lernprozess simultan berücksichtigt werden konnten. Mittels virtueller Synthesen werden Moleküle aus direkt verfügbaren chemischen Bausteinen zusammengesetzt. Damit erhöht sich die Wahrscheinlichkeit, dass die vorgeschlagenen Moleküle auch synthetisch zugänglich sind. Ein von der Natur inspirierter Ameisenkolonie Optimierungs Algorithmus (*Ant Colony Optimization*, ACO) wurde an die Anforderungen des molekularen Designs angepasst. Der ACO Algorithmus wird verwendet, um in dem potentiell großen kombinatorischen Raum nach geeigneten Lösungen zu suchen. Dabei werden die Gauss'schen Prozess Modelle als Fitness Funktion verwendet. Der ACO Algorithmus schlägt Molekülkandidaten vor, die den angestrebten Kriterien entsprechen, und adaptiert neue Vorschläge jeweils an die ermittelte Fitness. Dieses Vorgehen ermöglicht es, chemisch zugängliche fokussierte Substanzbibliotheken zu entwerfen. MAntA enthält zusätzlich eine Methode zur intuitiven Visualisierung von chemischen Räumen, durch die Medizinchemiker bei der Analyse der molekularen Designs visuell unterstützt werden. Bevor der MAntA Ansatz in einer umfangreichen prospektiven Studie praktisch evaluiert wurde, erfolgte zunächst eine individuelle Bewertung der vier zugrundeliegenden algorithmischen Komponenten.

1) Für die intuitive Visualisierung von großen heterogenen Moleküldatensätzen wurde die dreidimensionale (3D) Landschaftsvisualisierung (LiSARD) vorgestellt. Zur Berechnung der Landschaften verwendet LiSARD eine Nachbarschaft erhaltende Dimensionsreduktionsmethode und einen adaptiven Gauss'schen Kernglätter. In einer ersten Anwendung wurde die zeitliche Entwicklung eines Moleküldatensatzes aus einem realistischen Wirkstoffforschungsprojekt visualisiert. Bereits in der ersten Projektphase konnten für das Projekt wichtige Regionen im chemischen Raum identifiziert werden. Diese Regionen wurden in späteren Projektphasen bestätigt, womit die Vorteile einer frühen Projektunterstützung durch Visualisierungsmethoden verdeutlicht wurden, und zwar sowohl hinsichtlich der Priorisierung von Treffern als auch zum Monitoring des Projektfortschritts. Darüber hinaus wurden Multi-Kriterien Landschaften vorgestellt, welche zur visuellen Unterstützung beim Design von Molekülen für multiple Kriterien eingesetzt werden können.

2) Der etablierte CATS Pharmakophordeskriptor wurde um aromatische Merkmale erweitert (CATS2). Die retrospektive Analyse bestätigte eine verbesserte Anreicherung von bioaktiven Molekülen für den neuen CATS2 Deskriptor bei gleichbleibendem Potential zur Auffindung neuartiger chemischer Grundgerüste ("scaffold-hopping"). Der CATS2 Deskriptor wurde erfolgreich für die Vorhersage von makromolekularen Targets verwendet, indem der Deskriptor mit selbstorganisierenden Karten (*Self-Organizing Map*, SOM) kombiniert wurde. In der prospektiven Studie wurden für eine fokussierte kombinatorische Bibliothek zwei bis dahin unbekannte Targets entdeckt.
3) Die Gauss'schen Prozess Modelle wurden auf ihre Eignung zur akkuraten Vorhersage von Bindungsaffinitäten und der Diskriminierung zwischen Bindern und Nicht-Bindern hin untersucht. Die Methode wurde prospektiv angewendet in Kombination mit einem Mikrofluidiksystem für die On-Chip-Synthese. Dabei wurde eine fokussierte Substanzbibliothek erzeugt, die ein GPCR-privilegiertes chemisches Grundgerüst enthält. Vier neue GPCR Targets wurden für dieses Grundgerüst enthüllt, und 71% der experimentell getesteten Moleküle waren wie vorhergesagt aktiv.
4) Der potentielle Nutzen des ACO Algorithmus für die kombinatorische Erzeugung von Molekülen wurde anhand von zwei Szenarien evaluiert. Am Beispiel von Haupthistokompatibilitätskomplex I (MHC-1) bindenden Peptiden wurde gezeigt, dass der ACO Algorithmus in der Lage ist, effizient "Aktivitätsinseln" im vollständigen Oktapeptidraum zu identifizieren. In dieser Studie wurde der ACO Algorithmus mit einem kaskadierten maschinellen Ensemble-Lernansatz zur Vorhersage der MHC-1 Bindung kombiniert. Zusätzlich wurde die ACO Methode dahingehend erweitert, dass beliebige kombinatorische Reaktionen eingesetzt werden können. In einer retrospektiven Studie wurde untersucht, inwieweit die erweiterte Methode geeignet ist, fokussierte Substanzbibliotheken zusammenzustellen mit der Randbedingung, dass nur eine begrenzte Anzahl von Testzyklen durchgeführt werden kann. Es wurde gezeigt, dass die erweiterte ACO Methode die gewünschten Moleküle in der fokussierten Bibliothek effizient anreichert. Dabei übertraf die Methode andere Optimierungsalgorithmen, die früher bereits in demselben Zusammenhang untersucht wurden.

Schließlich wurde die MAntA Methode in zwei prospektiven *de novo* Design Projekten experimentell validiert. Der Fokus lag dabei auf Wirkstofftargets, die an neuropsychologischen Erkrankungen beteiligt sind. In diesen Machbarkeitsstudien wurde zur Konstruktion von neuen Molekülkandidaten ausschließlich die reduktive Aminierung verwendet. MAntA wurde eingesetzt, um selektive Sigma-1 Rezeptor Liganden sowie multi-target modulierende Dopamin D_4 Antagonisten zu erzeugen. Für jedes Szenario wurden Moleküle ausgewählt gemäß der vorhergesagten Potenz, der Selektivität sowie der Exploration des chemischen Raumes. Präferierte Regionen im chemischen Raum sowie die Lokalisierung der generierten Moleküle wurden mit der LiSARD multi-target Landschaftsvisualisierung dargestellt. 16 ausgewählte Moleküle waren ohne weiteres synthetisierbar, und die quantitative Bindungsaffinität-Vorhersage konnte biochemisch bestätigt werden. Insgesamt wurde eine Erfolgsquote von 90% erzielt. Weiterhin besitzen die erzeugten Moleküle Leitstruktureigenschaften. Dieses Ergebnis veranschaulicht das Potential von MAntA für den Entwurf von synthetisch zugänglichen Molekülen mit einem akkurat vorhergesagten multi-target Aktivitätsprofil.

1 Introduction

Small-molecule drug discovery is an interdisciplinary endeavor linking chemistry, biology and medicine in order to discover *New Chemical Entities* (NCEs) that exhibit a desired therapeutic effect[1]. Historically, the discovery of several important drugs was largely serendipitous, a prominent example being the discovery of penicillin[2]. With major technical advances in genomic sciences, combinatorial chemistry, and biological and biochemical assay technology, the rational design of drugs has now become feasible. The discovery of potential starting points for drug development is mainly facilitated by screening diverse libraries of readily available compounds by means of automated biochemical *High-Throughput Screening* (HTS) and by synthesizing individual molecules "from scratch" according to a target-specific structure-activity hypothesis[3]. A central paradigm has been the "one target, one gene, one disease" hypothesis. Pharmaceutical research has consequently focused on designing highly selective ligands affecting only a single macromolecular target. Today the mechanisms of several diseases are much better understood on a molecular level through advances in molecular and systems biology[4]. Polygenic causes for a disease involving highly cross-connected networks of proteins and signaling cascades are frequently encountered. These findings have resulted in a gradual paradigm shift towards drugs that exhibit a carefully balanced interaction profile with a whole panel of biological targets[5]. With the recently observed decline of efficacy in pharmaceutical R&D, it seems as if traditional drug discovery approaches have reached their limit. New sustainable technologies are required to effectively support the discovery of next-generation drugs[6]. A variety of computational methods have been developed that could help improve the discovery process[7-9]. Virtual screening approaches have been used to identify potential drug candidates in large databases of virtual compounds. Computer-assisted *de novo* design has been established for suggesting innovative compounds. Both computational approaches are complementary to bench experiments.

1.1 Virtual screening

Virtual screening of potentially large collection ("libraries") of molecules, often implemented in chemical database systems, is an important pillar of the computer-aided search for novel lead compounds[9,10]. Various computational methods are employed to prioritize prospective candidate compounds for biological testing and further investigation in lead discovery programs[11-13]. These candidate compounds can then be used as starting points for further development through medicinal chemistry, and provide "surprising" new ideas for compound optimization in later stages of the development process[9,14]. Before applying any target-specific information, compounds with undesired properties or unwanted chemical structures are often excluded from the screening library (also referred to as "negative design[15]"). Lipinski's *Rule-of-Five*[16] for orally bioavailable drug candidates, or the *Rule-of-Three*[17] for fragment-based lead compounds, are well-known examples of property filters that can be employed for tailoring the screening library towards drug- or lead-like properties[18]. More recently, Hopkins and coworkers introduced the *Quantitative Estimate of Drug-likeness* (QED)[19], a method to rank compounds according to their oral bioavailability, in contrast to simply classifying them as favorable or unfavorable. Undesired chemical substructures usually include reactive groups, which can be eliminated using the *Rapid Elimination Of Swill* (REOS) approach[20], or substructures facilitating promiscuous binding ("frequent hitters")[21], which can be identified by *Pan Assay Interference Compounds* (PAINS) substructure filters[22]. There are two principal categories of virtual screening approaches:

1. Ligand-based, if one or multiple reference ligands are known.
2. Receptor-based, if structural information on the macromolecular target is available.

1.1.1 Ligand-based

Many methods employed in ligand-based virtual screening rely on the *Principle of Strong Causality*, a term derived from technical optimization[23], which has been adapted to the field of drug design by Johnson and Maggiora as the *Chemical Similarity Principle*[24]. The latter states that compounds exhibiting high structural similarity should have an increased probability to exhibit similar properties. In other words, small structural changes should only have a small effect on biological function. Several studies have pointed out that strict compliance with the *Chemical Similarity Principle* might not always be effective[25-27]. However, while a perfect correlation between structural similarity and biological function cannot be anticipated, an enrichment of hits can be expected even if individual selected molecules are inactive[9,26]. In this context, Maggiora has highlighted the importance of chemical representation, due to the "lack of invariance of chemical space"[26]. By changing the chemical representation of a compound library, the neighborhood relationship in the chemical space spanned by the respective descriptors is also changed (Figure 1)[28]. Thus, selecting an appropriate context-dependent molecular representation is crucial for a successful similarity application[9]. Visualization aids with the selection of an appropriate representation, and is discussed in detail in Chapter 1.4.

A wide range of chemical descriptors have been proposed to represent chemical molecules, and an extensive overview of more than 2000 descriptors was compiled by Todeshini and Consonni[29]. Descriptors are usually classified in three categories according to the dimensionality of the chemical structure they operate on (Table 1, Figure 2)[30]. One-dimensional (1D) descriptors describe global molecular properties that can be calculated based on the chemical formula. Two-dimensional (2D) descriptors, derived from the connectivity table or the topological molecular graph, are the most frequently used descriptors for virtual screening[30]. Prominent examples are topological indices, single valued descriptors[31-33], topological autocorrelation descriptors represented as real valued vectors[14,34], and topological fingerprints, bit strings decoding the presence or absence of features[35,36]. A conformation of the molecule is required for the calculation of three-dimensional (3D) descriptors. 3D descriptors represent the spatial relationship of molecular features and properties.

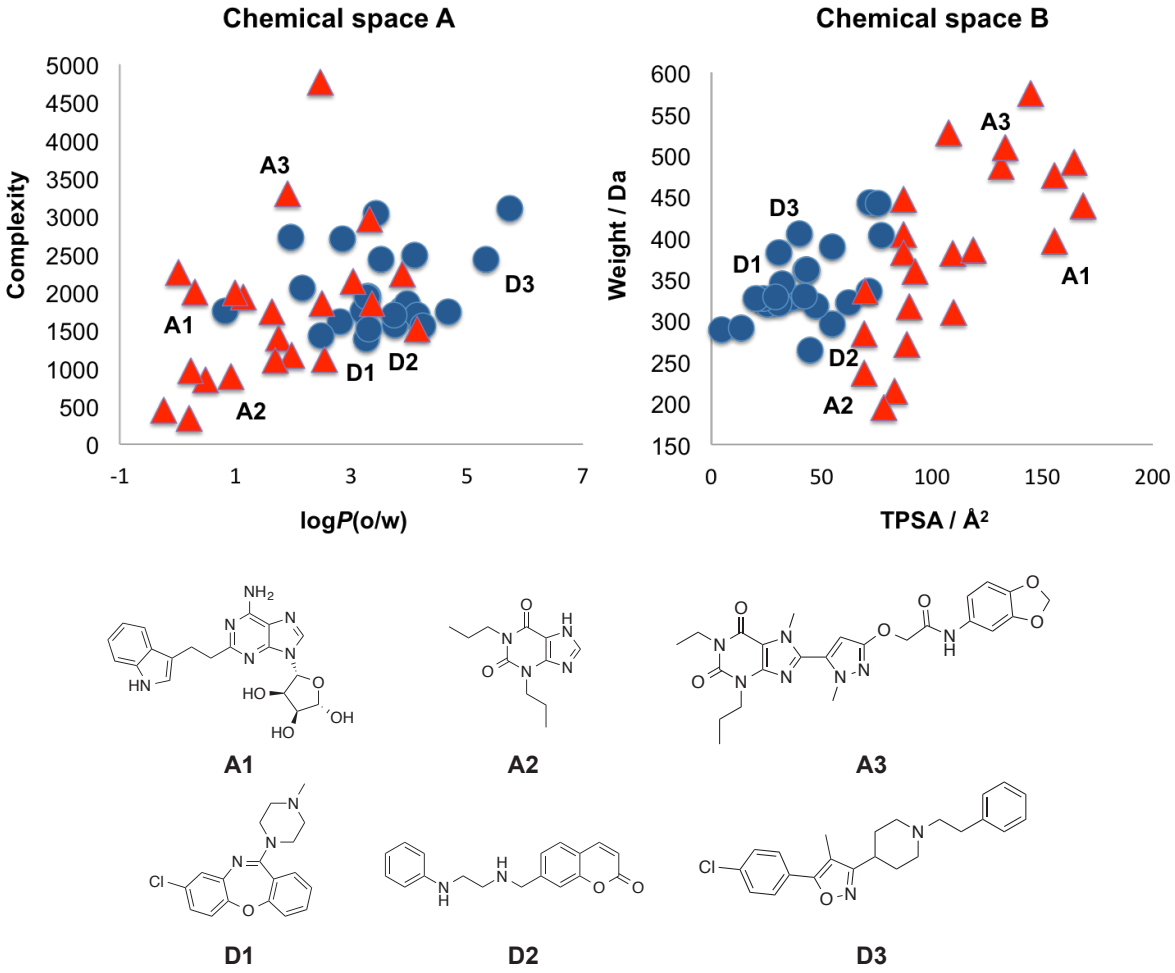


Figure 1. An example illustrating compound distributions according to two different molecular representations by numerical "descriptors". Symbols represent compounds collected from the COBRA database (v12.06)[37]. Red: Adenosine A_{2B} receptor ligands. Blue: Dopamine D_4 receptor ligands. The selection of different molecular descriptors changes the relative distribution of the compounds in the respective chemical spaces **A** and **B**. Descriptors were calculated using MOE 2012.10 (The Chemical Computing Group Inc., Montreal, Canada).

Table 1. Molecular descriptor categories. (Adapted from Ref.[30])

Dimensionality	Type	Examples
One (1D)	Global (whole molecule-based)	Molecular weight, atom and bond counts (*e.g.* number rotatable bonds, number hydrogen-bond donors/acceptors, number of rings), polar surface area, clog*P*
Two (2D)	Topological (molecular graph-based)	Topological and connectivity indices, substructures (*e.g.* maximum common substructures), topological fingerprints (*e.g.* structural keys)
Three (3D)	Conformational	Multi-point pharmacophore, molecular shape, 3D fingerprints

Examples include pharmacophore keys, encoding the distance of pharmacophoric features (*e.g.* hydrogen-bond donor / acceptor) in the Cartesian coordinates as fingerprints[38,39], radial distribution functions derived from the 3D conformation[40], and 3D pharmacophore correlation vectors[41,42]. As the ligand-receptor interaction is an event taking place in 3D space, one would think that 3D descriptors have a general advantage over 1D or 2D descriptors. Surprisingly, several studies indicate that this is not necessarily the case[42-44]. A drawback of current 3D descriptors seems to be the requirement for a suitable 3D molecular conformation (a conformer that is close to the bioactive ligand conformation), which has to be derived from the topological 2D information. Furthermore, the 3D conformation of a ligand and its target is dynamic and time-dependent, which is a property that is insufficiently contained in contemporary 3D descriptor types[11].

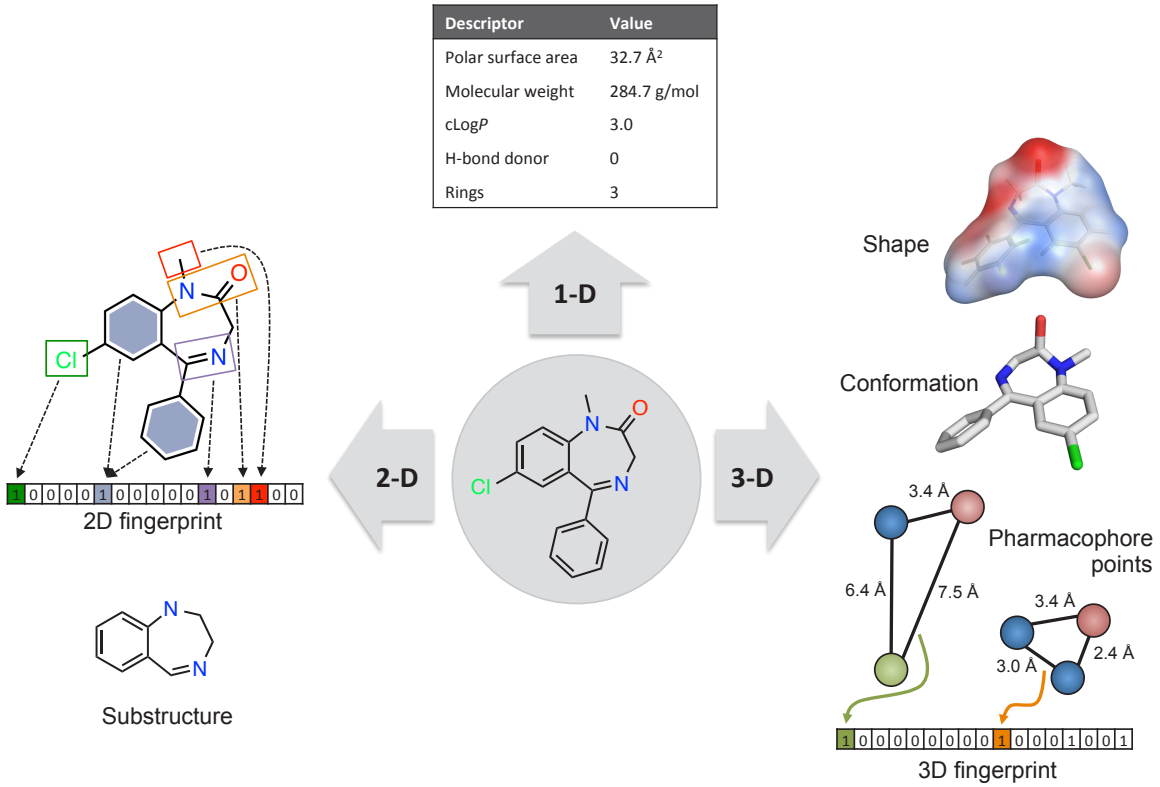


Figure 2. Examples of molecular descriptors for small-molecule ligands. (Adapted from Ref.[30])

To compute the similarity between two molecules, a similarity (or inverted distance) index or metric is required. Numerous concepts of chemical similarity metrics and indices have been described[45]. For real-valued vectors, the Euclidean and Manhattan distances (Minkowski metric, Table 2) are frequently used[42,46,47]. For binary

fingerprint descriptors, several studies have suggested that the Tanimoto coefficient (Table 2) might be a similarity index of choice for similarity searching applications[10,48-50].

Table 2. Equations of metrics and indices discussed in this work. A and B are molecules, $\boldsymbol{x}$ are molecule descriptors (continuous vectors or binary fingerprints), n is the total number of descriptor attributes, and $\boldsymbol{x}_{jA}$ is the value of the jth attribute. $D_{A,B}$ denotes the distance, and $S_{A,B}$ the similarity between molecules A and B. Manhattan and Euclidean distance are instances of the Minkowski distance with p being 1 or 2, respectively. Note that the range of the Tanimoto coefficient is [0,1] if it is used with non-negative attribute values or binary fingerprints.

Name	Equation	Range
Minkowski distance	$D_{A,B} = \left(\sum_{j=1}^{n} \left\lvert \boldsymbol{x}_{jA} - \boldsymbol{x}_{jB} \right\rvert^{p}\right)^{1/p}$, with $p \geq 1$	$[0,\infty]$
Tanimoto coefficient	$S_{A,B} = \frac{\sum_{j=1}^{n} \boldsymbol{x}_{jA}\boldsymbol{x}_{jB}}{\sum_{j=1}^{n} (\boldsymbol{x}_{jA})^2 + \sum_{j=1}^{n} (\boldsymbol{x}_{jB})^2 - \sum_{j=1}^{n} \boldsymbol{x}_{jA}\boldsymbol{x}_{jB}}$	$[-1/3,+1]$

A plethora of performance measures has been proposed to assess a virtual screening methods' prediction quality and predictive power[51-53]. Virtual screening methods usually rank sets of compounds according to some calculated score[54]. In order to be successful, scores assigned to molecules that relevant for the drug discovery project under investigation must be distinguishable from the scores of irrelevant molecules[54]. Therefore, the evaluation of score-based compound ranking performance is of central interest in the scope of virtual screening. Also, a successful method should compute ranked lists of screening compounds in such a way, so that interesting molecules are enriched in the top fraction of the ranked list ("early enrichment"). This requirement is owed to the fact that typically only a small fraction of the screened compounds will actually be tested experimentally[53]. A widely used performance measure in a variety of disciplines, including virtual screening[55], is the area under the *Receiver Operating Characteristic* (ROC) curve[56]. The ROC curve is derived by plotting the fraction of false positives versus the fraction of true positives[57]. However, the area under the ROC curve is not sensitive to early enrichment and therefore might be considered a poor metric for evaluating virtual screening performance[53]. Consequently, additional measures have been suggested that specifically address the early enrichment problem[51,53,54]. A frequently applied measure is the *Enrichment Factor* (EF)[54], which calculates the enrichment of actives in a given fraction of the ranked list (Eq. 1).

$$\mathrm{EF}(\chi) = \frac{s_+}{\chi n} / \frac{n_+}{n}, \tag{1}$$

where $\chi \in [0,1]$ is the considered fraction, n is the number of ranked samples, n_+ is the total number of active samples, and s_+ is the number of actives in the early fraction of the ranked list.While the EF metric is simple to calculate, it has several drawbacks *e.g.* the dependency on the active/inactive ratio, and the lack of including the exact position of actives in the early fraction[54]. An example that circumvent the problems present in the EF measure is the *Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic* (BEDROC)[53] metric (Eq. 2), which has been used in this work to evaluate virtual screening performance. It adapts the ROC AUC to early enrichment by an exponential weighting according to the rank of the actives.

$$BEDROC = RIE \times \frac{\frac{n_+}{n}\sinh\left(\frac{\alpha}{2}\right)}{cosh\left(\frac{\alpha}{2}\right) - cosh\left(\frac{\alpha}{2} - \alpha\frac{n_+}{n}\right)} + \frac{1}{1 - e^{\alpha\left(1 - \frac{n_+}{n}\right)}}, \tag{2}$$

where $RIE = \sum_{i=1}^{n_+} e^{-\alpha x_i} / \frac{n_+}{n}\left(\frac{1-e^{-\alpha}}{e^{\alpha/n}-1}\right)$, x_i is the relative rank of the i^{th} active in the ranked list, and α is the early recognition tuning parameter. Virtual screening methods are best applied once substantial information has accumulated regarding the ligand-receptor interaction under investigation. However, in an early project phase there might not be sufficient information available[58]. The methods discussed in this thesis can be grouped according to the required number of reference/training data: (i) Similarity searching is already applicable if a single reference ligand is available. (ii) Multi-reference methods are able to work with a small set of reference ligands. (iii) Machine-learning approaches that are applicable to large compound databases exceed the simpler similarity metrics and indices by providing an application-specific model.

Similarity searching

The *Chemical Similarity Principle s*tates that compounds similar to a molecule exhibiting an activity of interest, e.g. a marketed drug or clinical candidate, are more likely to also exhibit the activity of this reference compound. By ranking compounds from a screening pool according to their pair-wise similarity to the reference compound, one would expect the top-ranking molecules more likely to be active than

the lower-ranking compounds. Focusing on the top-ranked compounds should therefore yield better results than random hit rates in biochemical screening campaigns. This simple concept of virtual screening was initially published almost three decades ago using binary molecular fingerprint representations[59,60]. It has been successfully used in an overwhelming number of studies ever since, not only retrospectively[10,43,61] but also prospectively, thereby contributing to the finding of novel entry points to drug discovery[62,63]. In early applications, a single molecular representation and similarity coefficient was used. It has been suggested to combine different types of structural representations and similarity metrics in order to improve virtual screening performance[64]. In *similarity fusion*, a single compound is used as reference, and the results obtained with different representations[65] or different similarity metrics are combined[10,65,66].

Multi-reference

When several active molecules are available, an approach closely related to similarity fusion can be used to incorporate information obtained from the individual reference molecules[49]. In data fusion (sometimes also referred to as group fusion[10]) multiple reference compounds are used for the similarity search, while using the identical molecular representation and similarity metric for each search. Analogously to similarity fusion, the results are then combined to yield a consensus score for each screening compound[10]. Multiple studies emphasize the improved performance of multi-reference approaches compared to single reference virtual screening[10,49,67-69]. An example of a simulated similarity search using multiple reference structures is shown in Figure 3. Further improved retrieval performance is possible by not only considering the active molecules, but also molecules with undesired properties that exhibit no activity at a biological target (inactives)[49,69,70]. A prominent example is the B*inary Kernel Discrimination* (BKD, Eq. 3)[70,71] method, a non-parametric machine-learning method that uses kernel density estimation techniques to distinguish between active and inactive molecules (classification).

$$S_{BKD}(j) = \frac{\Sigma_{i \in active} K_\lambda(i,j)}{\Sigma_{i \in inactive} K_\lambda(i,j)}, \quad (3)$$

where

$$K_\lambda(i,j) = (1-\lambda)^\beta \left(\frac{\lambda}{1-\lambda}\right)^{\beta s_{ij}}, \tag{4}$$

with $s_{ij} = T_c(i,j)/n$, T_c is the Tanimoto structural similarity, and n is the length of the binary fingerprint.

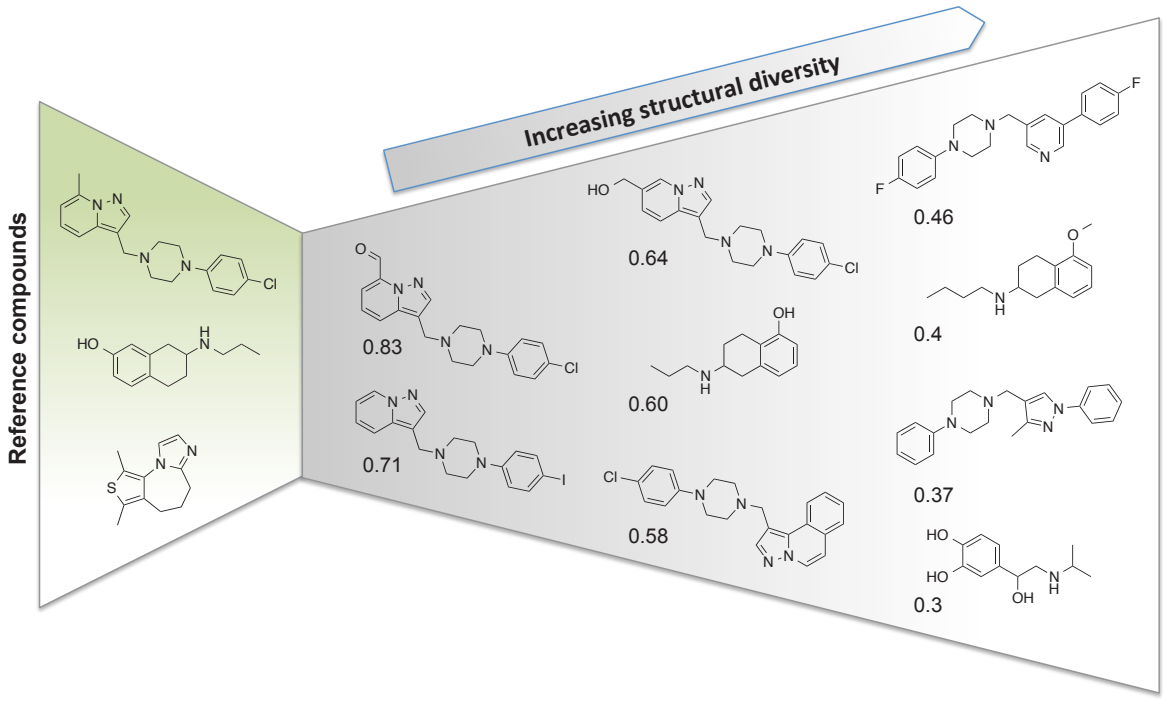


Figure 3. Starting from three diverse dopamine D_4 receptor ligands, a series of hits ranging from close analogs to increasingly diverse structures is identified by simulated similarity searching in a database of annotated binders and non-binders of dopamine D_4 receptor taken from the ChEMBL database[72]. Structural similarity is expressed as the pair-wise Tanimoto coefficient, and reported for each of the hits relative to the most similar reference. Molecules were compared using Morgan structural fingerprints[35].

3D pharmacophore modeling is another type of method that is applicable if multiple reference ligands are available. According to the *International Union of Pure and Applied Chemistry* (IUPAC) a pharmacophore (or pharmacophoric pattern) is defined as the "[...] ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response."[73]. Thus, a 3D pharmacophore model describes the spatial arrangement of key interactions between a receptor (protein, or another macromolecular target) and a ligand. If a crystal structures with co-crystalized ligands is available, the pharmacophore can directly be derived from the receptor-bound ligand conformation (receptor-based case)[74]. In the receptor-free situation, a pharmacophore model is constructed from multiple reference ligands

(known actives and inactives). Here, a crucial step is the computational sampling of a ligand conformation or conformation ensemble, which should ideally mimic the bioactive, receptor-relevant conformation[75]. By detecting spatially conserved potential pharmacophoric interaction patterns, a pharmacophore model is derived from the reference ligand's conformational ensemble. In this context, the ligand-receptor interaction is a function of the individual functional group contributions[9]. The respective groups are also termed *Potential Pharmacophoric Points* (PPP) to reflect the fact that it is not known *a priori* if they actually contribute to the receptor-ligand interaction[9].

To evaluate a screening molecule with a pharmacophore model, a 3D-alignment of the molecular features to the pharmacophore model must be computed. The 3D-alignment often is a time-consuming step, limiting its practicability for large-scale virtual screening experiments[9]. Explicit 3D-alignment can be avoided by using alignment-free representations of the pharmacophore model. Such a representation can be obtained by converting the PPPs of the individual function groups into vector representations (*e.g.* fingerprints, or correlation vectors) that have the identical number of elements. Instead of performing the explicit 3D feature alignment, these reduced molecular representations are then applicable to rapid compound database screening[9].

Machine-learning

Machine-learning techniques are applicable to virtual screening when a set of reference compounds is available (training data). In virtual screening, binary active/inactive labels are used in combination with classification methods. Real-valued labels (*e.g.* compound activity expressed as IC_{50}), on the other hand, are used by regression methods. A focus has been on supervised machine-learning methods. In supervised learning, each training sample is associated with a label, and the task is to infer the label for new samples by means of a statistical mathematical *Structure-Activity Relationship* (SAR) model[76]. Unsupervised methods where the samples are not associated with labels[77] are discussed in Chapter 1.4 for chemical space visualization and dimensionality reduction. The following machine-learning methods are frequently encountered in virtual screening: nearest neighbor analysis, naïve Bayes classifier, *Artificial Neural Network* (ANN), *Support Vector Machine* (SVM),

Gaussian Process (GP) model, and ensemble learning[15,78-81]. These methods exceed linear methods, in which the predictive output is calculated as a linear combination of the descriptor variables, by discovering a nonlinear relationship between the descriptor variables and the associated label. A particular advantage of machine-learning methods is their robustness against noise in the data[82-85]. This advantageous property stems from their ability to generalize from the training data to a model that explains the general trend of the data, whilst neglecting data points whose explanation would increase the model complexity, *e.g.* data points that are subject to noise[86,87]. In the next section, a brief overview of methods frequently encountered in virtual screening is provided.

Nearest Neighbor Analysis

In this approach, the prediction for a query molecule is inferred from the properties (labels) of the most similar compound(s) (*nearest neighbor(s)*) in the training data. *k-Nearest Neighbor* (kNN) methods have demonstrated their effectiveness in several studies[80]. In kNN analysis, the prediction is averaged over the k (k = 1, 2, etc…) nearest neighbors of the query molecule. The predictive kNN model consists of the training data and a similarity metric to calculate the nearest neighbor relationship. This basic kNN approach was further improved by considering the distance between the candidate molecule and its k nearest neighbors for calculating the prediction. This was achieved by weighting the contribution of each neighbor according to its inverse distance[88]. For example, Shen *et al.* used kNN in combination with a variable selection procedure to virtually screen a database of 250,000 molecules for anticonvulsant compounds. 48 reference compounds were used to build the kNN models. The authors synthesized nine compounds, out of which seven displayed appreciable anticonvulsant activity in mice[89]. In the kScore algorithm, Oloff and Mügge combined the epsilon loss function and *Structural Risk Minimization* term from SVM theory with a kNN approach[90,91]. It is used to estimate weights for individual descriptor variables, representing descriptor importance in respect to the training labels. Using a corporate dataset of 10,200 compounds with activity labels for a specific kinase, a 35-fold enrichment over random by selecting the ranked top 1% of a database containing 775,000 molecules was achieved.

Naïve Bayes Classifier

Naïve Bayes classifiers are machine-learning techniques that do not involve complex iterative modeling steps, but rely on observed feature frequencies. The idea behind Bayes classifiers is *Bayes' theorem*[92] (Eq. 5). If we define C as the hypothesis (target class) and X as the data (presence or absence of a feature in the compound to be classified), then $P(C|X)$ is the probability of the hypothesis given the observable features X (*posterior* probability). The conditional posterior probability of an event C, given data X, is:

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)}, \tag{5}$$

where

$$c^* = \arg\max_{c_k} P(C = c_k) \prod_i P(X_i|C = c_k), \tag{6}$$

X_i are individual features, and c_k the hypothesis. The binary *Quantitative Structure-Activity Relationship* (QSAR) method was among the first methods to apply naïve Bayes to virtual screening[93]. Bajorath and coworkers applied the binary QSAR method to a dataset of 463 estrogen receptor ligands with measured affinity, collected from the literature[94]. Affinity values were transformed into binary active/inactive values. A cross-validated classification accuracy of over 90% was reported, which was not greatly affected by induced noise. Such a tolerance is particularly important for HTS data, which is known to often be associated with high noise levels[84,95,96]. The robustness of naïve Bayes classifiers in regard to false-positive prediction was also shown in a comparative study conducted by Glick and coworkers[84]. Further improvements in performance have been reported for naïve Bayes classifiers with feature selection methods[97,98]. In combination with the MOLPRINT2D topological circular fingerprint[50], the naïve Bayes classifier performed favorably to standard similarity searching, data fusion, and BKD[99]. In a recent study, Hopkins and coworkers have successfully applied naïve Bayes classifiers using *Extended-Connectivity Fingerprints* (ECFP)[36] and a genetic algorithm molecular design procedure for the polypharmacological *de novo* design of G-protein-coupled receptor (GPCR) ligands. In this large-scale study, 75% of over 800 ligand-target predictions were experimentally confirmed[100].

In naïve Bayes classifiers, all features are class-conditionally independent; thus, mutual dependencies are not considered when training the classifier. *Bayesian Interference Networks* (BINs) can be used to model the relations between features[101]. BINs have recently been introduced to virtual screening[102,103]. In these two studies, BINs performed favorable in comparison to the Tanimoto similarity searching approach, increasing the relative hit retrieval rate by 8-10%.

Artificial Neural Network

Artificial Neural Networks (ANNs) have a long history in the field of chemistry[104] and have attracted much attention towards various tasks in drug design, including their application to virtual screening[105,106]. ANNs are widely used for diverse tasks such as classification, clustering, feature extraction, and function approximation[106]. One of the advantages of neural networks is that nonlinear structure-activity/property relationships can be modeled without prior knowledge about the required functional form of the model. The two types of neural networks most often encountered in drug discovery are three-layered feed-forward neural networks[107] (supervised learning method) and two-dimensional *Self-Organizing Maps* (SOMs)[108,109] (unsupervised learning method). Here, the focus will be on supervised ANNs. For applications of SOMs see Chapter 1.4.1.

Biological neural networks have been an inspiration for the structure and operation of ANNs. Standard multilayer feed-forward ANNs with one layer of hidden neurons, implementing a nonlinear activation function (*e.g. sigmoid* or *tanh*)[110], "[...] are capable of approximating any measurable function to any desired degree of accuracy."[111]. Hence, such networks are universal function approximators[110,111]. Neural networks consist of inter-connected neurons, which take multiple numerical inputs and calculate a transformed weighted sum. In feed-forward networks, descriptors are fed into the initial layer of fan-out neurons, hidden layers take the inputs from the output of the previous layers, and the final layer reports the predictions, usually only a single neuron for the predicted value. The neurons in each layer are fully connected as shown in Figure 6. Training of weights is most often done by *Back-Propagation of Errors*[107], where weights are modified to reduce the observed prediction error. A variety of different training methods exist and are actively used in drug discovery applications[106,112].

ANNs have been used for a variety of objectives in virtual screening and drug design[106,113,114]. One of the early virtual screening applications was to use ANNs as a first-pass filter to distinguish between drug and non-drug molecules[105]. Using the ANN filter, Sadowski and Kubinyi classified 83% of the *Available Chemicals Directory* (ACD)[115] correctly as non-drugs, and 77% of the *World Drug Index* (WDI)[116] as drugs[105]. An external test-set, consisting of a selection of top selling drugs, was correctly classified as drugs by the ANN filter. In a study by Schneider and coworkers at Roche, ANNs were used to identify compounds that show up as hits in assays for many different targets ("frequent hitters") due to assay interference or promiscuous binding[21]. Frequent hitters are in general considered being poor starting points for further drug development. Their nonlinear ANN model, with fragment-based Ghose-Crippen descriptors[117-119] as input, was able to correctly classify over 90% of the compounds, an improvement of about 10% compared to a benchmark linear *Projection to Latent Structures* (PLS) model. Interestingly, only two hidden neurons were sufficient to achieve this result. In a related study, predicting cytochrome P450 3A4 inhibition, an ANN was found to be inferior to linear-PLS[120].

Ajay *et al.* were among the first to apply ANN to virtual screening of a large database with the objective of finding potentially central nervous system active compounds[121]. The authors used a combination of seven 1D and 166 2D ISIS fingerprint descriptors[122] with a Bayesian neural network[123,124] to construct a focused library of central nervous system active compounds, but did not prospectively validate their findings. Comparable work was done for designing kinase and GPCR focused libraries[125,126].

In a genuinely prospective virtual screening study, Derksen *et al.* screened a catalogue of 229,658 molecules for peroxisome proliferator-activated receptor (PPAR) modulators using a probabilistic neural network[127] and the CATS[14] topological pharmacophore descriptor[128]. Nine molecules were selected for testing, based on visual inspection of the top–ranking 20 compounds for PPARα and PPARγ, out of which four showed activity.

At Vanderbilt University, a large-scale virtual screening study aimed at finding novel chemotypes of metabotropic glutamate receptor 5 ($mGluR_5$) NAMs using a HTS screen of 160,000 compounds as basis for ANN modeling[129]. Models were trained on different subsets of the ADRIANA descriptors[130]. The model yielding the highest

retrospective enrichment factor was used to screen a database of 708,416 commercial available compounds. 749 compounds were selected and 88 were experimentally confirmed as $mGluR_5$ ligands. This result corresponds to an enrichment factor of 16 compared to the original HTS approach. The study revealed two potent $mGluR_5$ NAMs, featuring a novel molecular scaffold.

By combining different virtual screening methods, namely pharmacophore modeling and ANN classification using "inductive" QSAR descriptors[131], Cherkasov *et al.* screened a library of 23,836 natural products in order to find novel non-steroidal ligands for the human sex hormone binding globulin (SHBG)[132]. The authors identified 29 potential SHBG ligands in the library and could confirmed eight *in vitro*.

In a comparative study, ANNs and SVMs[133] were applied to the drugs/non-drugs classification problem[134]. While the overall correct prediction was high (> 80%), SVMs had a slight advantage but in general did not outperform the ANNs. The authors pointed out that SVM and ANN complement each other by producing non-identical sets of correctly and misclassified compounds, which could be exploited by ensemble learning methods[134,135].

Support Vector Machine

The *Support Vector Machine* (SVM) is widely used in the machine-learning community and is particular popular for various tasks in chem- and bioinformatics[136,137]. SVMs were originally developed by Vapnik and coworkers[133] in the field of handwritten digit and character recognition and were introduced to chemoinformatics and drug discovery in 2001 by Czermiński *et al*. and Burbidge *et al*.[138,139]. SVMs are based on the *Structural Risk Minimization* principle, which simultaneously takes into account both the capacity of the model, measured by its *Vapnik–Chervonenkis* (VC) dimension[140], and the classification error on the training data[133]. The idea is that while more complex models might fit the training data better, they are prone to only learning the training data by heart without generalizing well to unknown data (overfitting effect). Adding a regularization term that penalizes complex models can be beneficial for generalization abilities of the model by avoiding overfitting via a higher order approximation. An intuitive example of *Structural Risk Minimization* is Occam's rule of simplicity (Occam's razor)[141,142], which would keep the training error equal to zero and minimizes the VC-dimension. By minimizing the VC

dimension, while still allowing a certain degree of errors, one may obtain an even better generalization[133]. Generalization is accomplished by constructing a separating hyperplane, usually in a high-dimensional descriptor space that maximizes the distance between the hyperplane and the nearest training points. The subset of data points defining this boundary is referred to as *support vectors,* and the distance of the closest points of the separating plane as *margin*. Generalization ability is not influenced by the cardinality of the descriptor, but only depends on the number of support vectors required to define the hyperplane[133]. This enables SVMs to be used with high-dimensional descriptors even if only a small number of training data is available[143]. In case of non-separable classes, the soft-margin hyperplane is applicable, which maximizes the margin while keeping the number of misclassified samples minimal. Constructing the optimal hyperplane is a convex quadratic programming problem, which can be globally solved by nonlinear programming. As it is a convex problem, the solution found is always a global optimum[144]. SVMs do not make any assumptions about correlation of descriptor dimensions, in contrast to *e.g.* the naïve Bayes approach that assumes statistical independence of dimensions. Therefore, SVMs are also applicable to problems with highly correlated descriptors[139]. To allow the definition of nonlinear hyperplanes, SVMs utilize kernel functions to implicitly map input data into a potentially indefinite dimensional feature space, without requiring to explicitly calculate this transformation (the so-called *kernel trick*, Figure 4). Kernel functions calculate the similarity between two data points in terms of inner products[144]. Linear separation in high-dimensional space corresponds to a nonlinear separation in the original data space. In combination with real-valued, vectorial molecular property descriptors, the *Gaussian* kernel (also *Radial Basis Function* (RBF) kernel, Eq. 7) is widely used in virtual screening[143].

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right), \tag{7}$$

where σ is the kernel width, $\boldsymbol{x}$ and $\boldsymbol{y}$ are molecular descriptor vectors, and $\|\cdot\|$ is the Euclidean norm. Baldi and coworkers have evaluated kernel functions for a variety of molecular representations[145]. While most kernels rely on explicitly calculated molecular descriptors, molecular graph kernels offer an intriguing alternative by avoiding the step of calculating the descriptors[146]. If the dimensionality of the

descriptor is already very high, as often encountered with substructure fingerprint descriptors, the implicit projection to an even higher dimensional space is not necessarily required and it is sufficient to use linear kernels, which are closely related to fingerprint-based similarity searching[147,148]. Linear SVM kernels are favorable in terms of performance, especially when applied in large-scale settings, while at the same time being able to maintain prediction accuracy in benchmark studies[147].

For virtual screening applications it is desirable to not only classify the compounds but also rank them in order to obtain an enrichment of actives among the top ranking compounds. The signed distance between a candidate compound and the hyperplane can be used for such a ranking[149-151]. When real-valued class labels are available (*e.g.* IC_{50} measurements) another possibility is to use SVM regression and rank according to the prediction[152]. Currently, SVM variants are emerging that directly optimize the ranking performance instead of classification or regression[153-155]. The bipartite RankSVM algorithm[156,157] has shown favorable performance in a retrospective comparison compared to standard SVM and SVM regression in virtual screening applications[153].

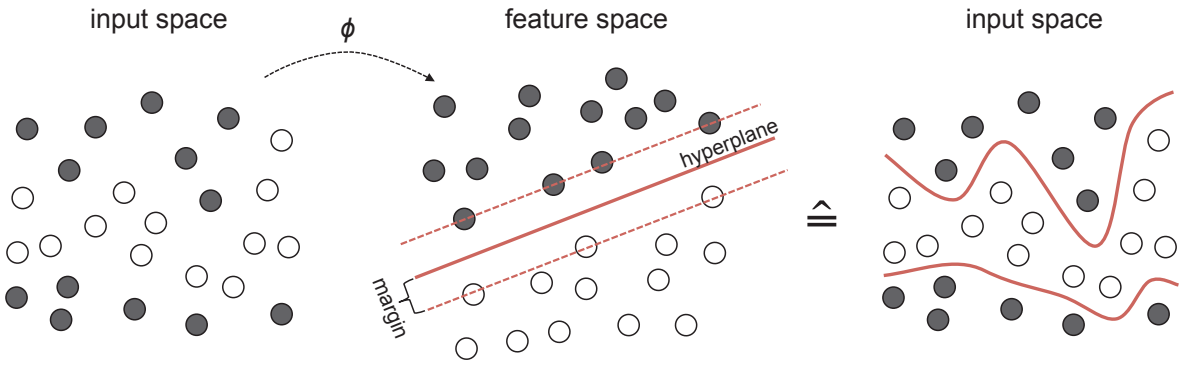


Figure 4. The kernel approach. Black and white dots represent two distinct classes (active / inactive) that are not linearly separable in input space. The mapping ϕ transforms the data in a feature space in which the data is linearly separable. There are many hyperplanes that could separate the data, and the one maximizing the margin is chosen by SVM. Dots intercepted by the dotted line are support vectors. The hyperplane in feature space defines a complex nonlinear surface in input space. (Adapted from Ref.[158])

Schneider and coworkers were the first to prospectively apply SVM methods for the screening of COX-2 inhibitors[150]. The authors screened a vendor library of 2.7 million substances with SVM models trained on subsets of their COBRA collection of pharmacologically active compounds. From 13 cherry-picked candidates with novel structural features, three compounds showed an inhibitory effect with one compound exhibiting even greater activity than the references celecoxib and rofecoxib. Machine-

learning methods are usually considered to be "black box" methods without an immediate interpretation. In the same study, Schneider *et al*. demonstrated the possibility to derive important pharmacophore points from the SVM model, which can then be visualized for intuitive interpretation[150]. The same group also applied the concept of semi-supervised learning (termed "active learning" in these publications) to increase the enrichment by retraining the SVM with the top ranked fraction of the training data[150,159]. The prospective applicability of SVM regression was shown in a study searching for selective dopamine D_3 receptor ligands[159]. The authors combined a binary SVM for discrimination of actives/inactives with SVM regression for predicting selectivity ratios between dopamine D_2 and D_3 receptors. 11 compounds were experimentally tested, and six of them were confirmed as D_3 receptor ligands with low micromolar binding affinities. Three compounds exhibited the desired selective for the D_3 receptor. Using similarity searching, one initial hit was successfully optimized to nanomolar potency, while maintaining the desired receptor selectivity.

By combining kNN regression[160], and SVM regression in a consensus approach, Tropsha and coworkers identified novel human histone deacetylase inhibitors[161]. The best performing models were selected for prospective virtual screening of 9.5 million compounds. From 45 consensus hits, predicted by both methods, the authors selected four compounds and confirmed three of them experimentally as inhibitors with micromolar activity.

In a study mimicking the traditional iterative drug discovery process, Warmuth *et al*. investigated the benefits of applying the active learning paradigm from machine-learning theory in combination with SVM methods to virtual screening[148]. They proposed different selection strategies and demonstrated that active learning is advantageous in comparison to passive learning.

For orphan GPCR-receptors, *i.e.* GPCRs that bind unknown ligands to modulate their function, a SVM regression screening method has been proposed to identify potential small-molecule ligands[162]. A target-ligand kernel combined information about targets as well as corresponding ligands. In a thorough retrospective evaluation, the authors showed the potential applicability to orphan GPCR screening. An experimental validation is, however, still outstanding.

Gaussian Process Model

Gaussian Processes (GPs) are a popular machine-learning technique in a variety of fields including computer science, bioinformatics, environmental sciences, and robotics, but are a relative new addition to the cheminformatics toolbox[163,164]. The usage of GP to define prior distributions over functions dates back to the work of O'Hagan in 1978 with the application to one-dimensional curve-fitting[165]. In cheminformatics, GPs were initially used for prediction of physicochemical properties like solubility[166] or lipophilicity[163] and ADMET properties[167,168]. Most recently, GPs have also been employed for virtual screening[169].

The idea of GP modeling is to place a prior directly on the space of model functions without parameterizing the functions. In a sense this is a generalization of a Gaussian distribution over a finite vector space to an infinite dimensional function space. A Gaussian process is fully defined by its mean and covariance function. Usually, the mean function is the zero function. The covariance function expresses the expected covariance between the function values at two points, comparable to the SVM kernel function[170]. Importantly, the prior does not depend on the data but defines general properties of the functions by specifying the mean and covariance functions with corresponding hyperparameters. The prior is used for Bayesian inference in feature space. To yield the posterior distribution, the probability distribution is updated in the light of the observed data. For unobserved data points, the mean and variance of the posterior distribution is used as the prediction and confidence estimate, respectively. In Figure 5a, four random functions are drawn at random from the prior distribution, only constrained by the covariance function and the zero mean function. In Figure 5b, observations are added, and random samples are drawn from the posterior distribution. Only functions going through or passing the data closely are likely to be sampled. The shaded area indicates the estimated uncertainty of each prediction, which is low in close proximity to the training data and large in greater distance. Consequently, the predictive variance depends only on distribution of data points and not on their actually observed values[171].

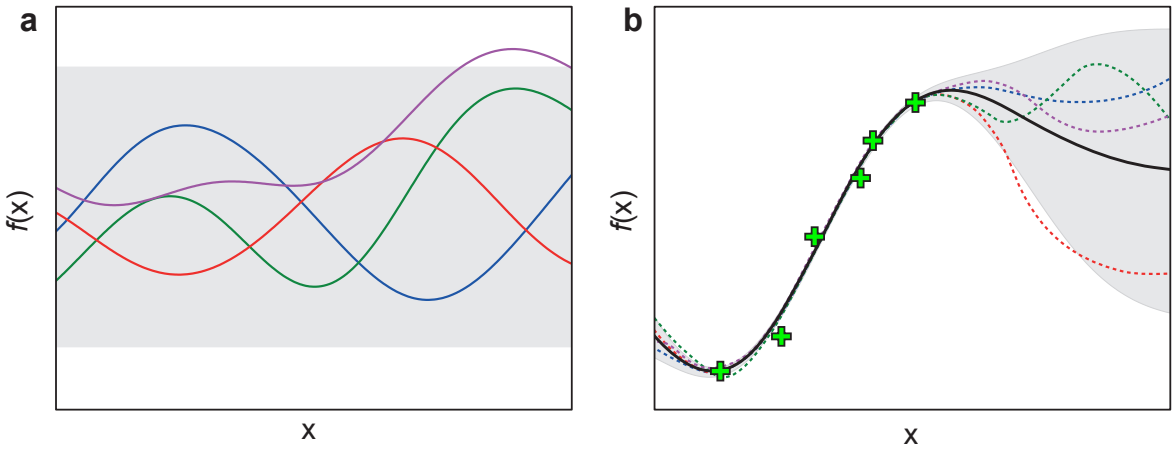


Figure 5. Illustration of *Gaussian Process* (GP) inference. (**a**) Four sample function (colored lines) are drawn from a GP prior with zero mean and isotropic squared exponential covariance function. The shaded area indicates the confidence interval (**b**) Four sample functions drawn from the posterior distribution (dashed lines) and the predictive mean function (solid black line) after introducing six observations (green symbols).

Training a GP model corresponds to finding hyperparameters suitable for the problem under investigation. Due to the underlying Bayesian framework, it is possible to directly infer hyperparameter values from the training data. One intriguing option is to choose those hyperparameter values that maximize the *log marginal likelihood*, which is the probability of the data given the hyperparameters. Particularly beneficial is the fact that, by using the *log marginal likelihood*, the complexity of the model and the data-fit are automatically balanced without relying on external methods such as cross-validation[171].

Several retrospective studies have been published showing the general usefulness of GPs for early drug discovery and the state-of-the-art performance that is comparable to SVM models or random forests[172,173]. In a pioneering prospective application, Rupp *et al.* used GP modeling to identify a natural product derivative that selectively activates PPARγ[169]. Different GP models with varying descriptors and kernels were trained on a dataset of 144 published PPARγ ligands. With the three best performing models a collection of 360,000 compounds was screened. 15 candidate compounds were manually selected from the obtained ranked lists, out of which eight were experimentally validated as PPARγ agonists. Interestingly, the most active compound was a derivative of the natural product truxilic acid and presented a scaffold-hop from synthetic compounds to a natural product.

The statistically well-founded estimate of confidence in a prediction provided by the GPs predictive variance can be further exploited. It has been shown to be useful for defining the model's domain of applicability[174] and could potentially be used in virtual screening to improve enrichment by focusing on the domain of applicability. It can also be used as an exploration-exploitation trade-off in active learning[175]. The theory for GP-based active learning is well established due to the work in closely related fields, including experimental design[176] and global optimization[177]. Initial work on GP active learning with a focus on drug discovery was done by De Raedt and coworkers who investigated different strategies to select next candidates based on the predictive variance in an iterative screening application[178]. Recently, Rupp *et al*. demonstrated the applicability of GP-based active learning to accurately estimate natural product conformational energies at the density functional level of theory[179]. Conformations were obtained from *Molecular Dynamics* (MD) simulations of the natural product archazolid A.

Ensemble Methods

Traditional modeling approaches consist of one single predictive model. Recently, methods have been proposed which aim at improving prediction accuracy and robustness by combining an ensemble of models. Such ensembles can consist of homogeneous classifier, trained with varying descriptors or different training data, or even heterogeneous classifiers. Examples include jury classifiers[180], bagging[181], boosting[182], and random forests[183].

In jury approaches, several (possibly weak) base classifiers are combined in a voting or cascaded learning scheme to obtain strong ensemble classifiers[180,184]. An illustration of a cascaded learning model combining two neural networks trained on different molecular representations is shown in Figure 6. For finding selective metabotropic glutamate receptor 5 ($mGluR_5$) NAMs, Renner *et al*. combined 10 ANNs trained on varying training data and descriptors in an averaged jury prediction[185]. From over one million molecules, a focused library of 8403 was selected according to the jury prediction. To assess the diversity of the library, SOM projections were calculated. From each SOM cluster, a representative was picked and 33 compounds were experimentally tested. Seven molecules exhibited activity for either $mGluR_1$ or $mGluR_5$ in the EC_{50} range 9-50 μM.

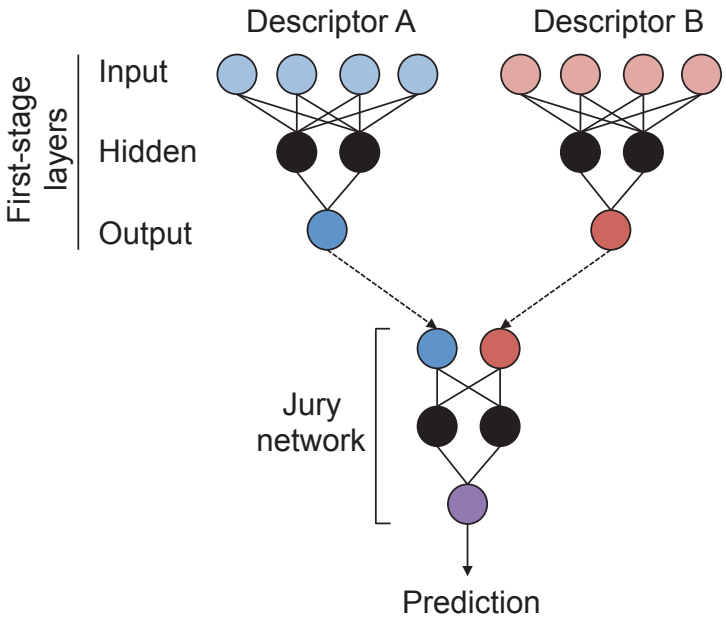


Figure 6. Cascaded neural network model. Individual first-stage feed-forward neural network models, trained on different molecular representations (descriptor A and B), are combined in one jury network. The final prediction is a weighted model of the two different descriptor perspectives. (Adapted from Ref.[15])

The bagging scheme employs bootstrapping[186] to train one type of classifier on varying subsets of training data[181]. By allowing the classifier to learn from different samples of the original distribution, the classifiers predictive performance is stabilized. In boosting, a set of classifiers is constructed by giving poorly predicted samples additional weight in the subsequent classifier, thus focusing on the observations that are difficult to predict[187]. In both methods the final prediction is found by majority voting. Agrafiotis and coworkers used bagging with ANNs to construct predictive models for human ether-á-go-go related gene (hERG) binding[188]. Svetnik and coworkers investigated the performance of boosting for a set of ten cheminformatics datasets[182]. Two datasets were representative of virtual screening applications; one consisted of 15,440 compounds screened against cycline-dependent kinase 2 (CDK2), and the other of 23,102 compounds with IC_{50} inhibition measurements for an unspecified channel protein. The authors concluded that the performance of their stochastic gradient boosting method is comparable to that of random forest or SVM techniques.

Random forests are a relative new addition to ensemble learning[183]. They consist of a forest of decision trees, each built from a randomly sampled subset of descriptors and training data. Trees are grown to the maximum size, and generalization is achieved by constructing several trees without pruning[183]. Random forests are easy

to construct, inherently provide measures of variable importance, and have only one adjustable parameter; still, they achieve performance levels in various benchmark studies that are comparable to other machine-learning methods[147,172,182,189]. For example, a random forest trained on 2597 bioactive plant compounds was used to screen a Chinese herbs dataset consisting of 8264 compounds[190]. 83 herb-target predictions were confirmed by literature reference.

1.1.2 Receptor-based

While in ligand-based virtual screening information about the receptor-ligand interaction is only implicitly included by structural preferences in the reference ligands, it is explicitly exploited in receptor-based (structure-based) virtual screening. When a receptor protein structure is known, typically an X-ray structure or a carefully designed comparative protein structural model, it can be used for virtual screening by deriving 3D pharmacophores or shape information from the ligand-binding pocket, or by applying automated ligand-receptor docking[7,191]. Note that even with the availability of a structure of the biological target, there often lies value in applying ligand-based approaches in parallel or combining them in hybrid methods[192,193].

Receptor-derived pharmacophore screening

With the known macromolecular target structure, information about *e.g.* pocket shape or excluded volumes[194] can be added to a pharmacophore model, or the pharmacophore model can be directly derived from the receptor without considering any ligand. A common concept is to identify pharmacophore features in the binding site. Two seminal approaches are GRID[195] and LUDI[196]. GRID calculates the interaction energy of representative small-molecule probes (hydrophobic, hydrogen-bond donor / acceptor) in the binding site and identifies regions of favored or disfavored interactions for the respective probe[195]. In LUDI, potential hydrogen-bonding and hydrophobic interaction sites are identified by a set of rules that were derived from interactions observed in known crystal structures[196]. PPPs can be identified by analysis of the detected interaction sites. The resulting 3D pharmacophore model can then be applied to compound database screening using

pharmacophore fingerprints, autocorrelation vectors, or pharmacophore alignment[197].

In a pioneering study at Roche Pharmaceuticals, Basel, pharmacophores derived from LUDI interaction sites were used to identify DNA gyrase inhibitors[198]. A library of 350,000 compounds was screened and 3000 selected compounds were tested experimentally, revealing 150 hits. Further hit optimization yielded several potent DNA gyrase inhibitors. Schneider and coworkers used clustering of LUDI-derived feature maps to define a "fuzzy" pharmacophore representation[199]. The idealized receptor-derived ligand pharmacophores (termed "virtual ligand") were translated into a feature correlation vector, which was used for alignment-free similarity searching. In a prospective study aiming at finding inhibitors of *Helicobacter pylori* protease HtrA, the authors extracted pharmacophore models based on the predicted binding pocket of a homology model. They screened 556,763 compounds, selected 26 hits, and experimentally confirmed six as HtrA inhibitors, which is notable considering the absence of a *H. pylori* HtrA crystal structure[199]. LigandScout was used in a receptor-based study to automatically derive pharmacophore models for angiotensin converting enzyme 2 (ACE2) and also accounting for ACE-subtype selectivity[200]. This model identified 26 ACE2 inhibitors from a library of 3.8 million compounds. For GPCR proteins, Sanders *et al.* developed the pharmacophore-screening tool Snooker that combines comparative protein structural modeling, interaction feature extraction, pharmacophore modeling, and screening all in one process[201]. The authors reported enrichment factors exceeding tenfold for eight out of 15 GPCR targets in a retrospective virtual screening evaluation. An in-depth review of pharmacophore-based methods and applications with a special focus on virtual screening is given by A. R. Leach[191] and D. Horvath[202].

Molecular docking

If the macromolecular target structure is known, another strategy commonly used in drug discovery projects is molecular docking[7,62,192,203-205]. Molecular docking aims at "predicting the structure and binding free energy of a ligand-receptor complex given only the structures of the free ligand and receptor"[206]. Many docking tools have been described. In 2008, Moitessier *et al*. referenced more than 60 different programs with over 30 individual scoring functions[207], and new tools are continuously reported[208].

Table 3 gives an overview of the most actively applied docking programs based on their citation in the recent literature[208]. In the following, a brief overview of the various concepts used in docking algorithms is provided. Further approaches can be found in the literature[207-210].

Docking methods have two integral components: An efficient search procedure to generate a plausible conformation, as well as the position and orientation of a ligand in the protein binding site ("pose"), and a scoring scheme ("fitness function") to evaluate the quality of the ligand-protein interaction. The latter is closely related to predicting ligand binding affinity. In most of the major docking programs, the ligand is treated as flexible and the protein is kept fixed in its crystal structure conformation, or only allowing limited flexibility of the amino acid side chains. Docking can be achieved by iteratively sampling ligand conformations in the binding site and subsequently assessing the quality using a scoring function. The goal is to find conformations closely resembling the ligand conformation observed in the crystal, and to assign higher scores to high affinity ligands than to non-binders. The latter is crucial for virtual screening applications. While both aspects can be governed by a single scoring function, it is possible to use separate scoring functions for docking and final pose scoring. In fact, several studies suggest that rescoring of docked poses is favorable for virtual screening applications[211-213].

Table 3. Examples of commonly used docking programs.

Software	**Sampling strategy**[a]	**Scoring function**[b]
AutoDock[214]	GA	FF/E
GOLD[215]	GA	FF/ E
Glide[213]	Hierarchical filters and MC	E
Surflex[216]	IC with MA	E
FlexX[217]	IC	E
DOCK[218]	IC	FF
ICM[219]	MC	FF
MOE[220]	MA	FF
CDOCKER[221]	MD-SA	FF
eHiTS[222]	IC	E

[a] GA, *Genetic Algorithm*; IC, *Incremental Construction*; MA, *Matching Algorithm*; MC, *Monte Carlo*; MD, *Molecular Dynamics*; SA, *Simulated Annealing*

[b] FF, *Force Field*; E, *Empirical*; KB, *Knowledge Based*

Ligand conformation sampling and placement strategies

Essentially, docking is an optimization problem, and many different algorithms are applicable for efficiently identifying the optimum pose that maximizes the scoring function[223]. Within rigid docking methods, pre-calculated conformation libraries are docked using shape and interaction complementary of the ligand and the protein active site[224]. Matching algorithms place the rigid ligand by matching the ligands' shape, geometrical features, or PPPs with the receptor binding pocket[224-227]. Matching algorithms are very fast, but the fixed conformation ensemble prevents the method from tailoring conformations to a specific binding pocket[202,228]. Other search algorithms treat the conformational exploration as an integral part of the process. Incremental construction methods[217,222] divide the ligand in several fragments and re-assemble the ligand on-the-fly in the binding site. One fragment is used as anchor and the remaining fragments are incrementally connected using a library of preferred geometries. Fragments are then placed using rigid-docking.

A second class of sampling methods uses stochastic optimization to simultaneously search for both ligand conformation and position by randomly modifying the ligand pose. Monte Carlo sampling[229] and genetic algorithms[230] are examples of such strategies. Monte Carlo methods modify the pose through random bond-rotation, translation, or rigid body rotation. The candidate pose is evaluated, and if its score is an improvement over the previous pose, it is immediately accepted. For higher energy conformations the Metropolis criterion is used to decide if they are accepted or rejected[219]. Genetic algorithms represent the ligand pose as a chromosome, encoding the ligand conformation, translation, and position as individual genes[231]. Genes are subject to modifications by genetic operations (mutation, crossover, migration). A population of possible solutions is evolved using the genetic operations, and the resulting population is evaluated using the scoring function. Only a fraction of candidate solutions is kept for the next generation[215,218].

In general, the conformational sampling is handled sufficiently well and poses in close resemblance of the bioactive conformation are frequently among the proposed candidate solutions[210,232]. How to reliably identify good docking solutions from the candidates is still an open question as current scoring functions often fail to assign highest ranks to the biologically relevant poses[210,232,233].

Scoring functions

A scoring function should be able to reliably distinguish between binders and non-binders, identify the bioactive conformation among all generated conformations, and produce correct rankings in terms of the ligands' potential free energies of binding[210]. Several scoring functions have been proposed for this purpose. In 2008, Moitessier *et al.* referenced more than 30 variations, which can be grouped into three categories[7,234]:

1. Physically motivated force fields,
2. Empirical scoring functions,
3. Knowledge-based scoring functions.

Force field scoring functions

Force field scoring functions are based on modeling physical atomic interactions of the ligand and the receptor molecule[235]. Non-bonded interaction terms typically include *Van der Waals* (VDW) and electrostatic interaction as well as the internal strain energy. Force field parameters are usually derived from experimental data or *ab initio* quantum mechanical calculations[236]. In Eq. 8 the function used in the initial implementation of the software DOCK is given as an example of a force field scoring function[236]. It consists of the Lennard-Jones VDW potential and an electrostatic term.

$$E = \sum_i \sum_j \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + 332 \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \right], \tag{8}$$

where r_{ij} is the distance between protein atom i and ligand atom j, A_{ij} and B_{ij} are VDW repulsion and attraction parameters, $q_i q_j$ are atomic point charges, ε is the dielectric function, and the factor 332 converts the electrostatic energy into kcal/mol. In DOCK, the force field parameters are taken from the AMBER force field[237], excluding the explicit hydrogen-bonding terms. To account for solvent effects the distance dependent dielectric function $\varepsilon(r_{ij})$ was added to the Coulombic term. To consider the solvent effect more rigorously, water molecules can be explicitly included. Examples implementing this approach are free energy perturbation[238], and thermodynamic integration[239]. Both methods are computationally demanding and currently not suitable for fast virtual screening[209,240]. A computationally more efficient, but more coarse-grained approach is to treat the solvent as a continuum

dielectric medium[232]. Examples of such implicit solvent models are Poisson–Boltzmann or generalized-Born surface area models[241]. In 2013, the Nobel Prize in Chemistry was awarded for the development of multiscale models for complex chemical systems[242]. The work honored includes the method to calculate protein-ligand interactions more accurately by combining *Quantum Mechanics* and *Molecular Mechanics* force fields (QM/MM)[243]. In this approach, QM is used for the local ligand environment, while MM is used for the remaining system. Rigorous QM/MM methods are currently still too computationally demanding to be used as scoring functions in high-throughput virtual screening[244].

Empirical scoring functions

In empirical scoring functions, also termed "regression-based" scoring functions, the estimation of the free energy of binding, ΔG_{bind}, is decomposed into a set of weighted energy terms. An example including terms for hydrogen-bonding, metal ligation, hydrophobic effects, and ligand flexibility is given in Eq. 9, generalized from the commonly used ChemScore scoring function[245].

$$\begin{aligned}\Delta G_{\mathrm{bind}} = \Delta G_0 + \Delta G_{\mathrm{hb}} \sum_{\mathrm{hb}} f(\Delta \mathrm{r}) f(\Delta \alpha) + \Delta G_{\mathrm{met}} \sum_{\mathrm{met}} f(\Delta \mathrm{r}) \\ + \Delta G_{\mathrm{lipo}} \sum_{\mathrm{lipo}} f(\Delta \mathrm{r}) + \Delta G_{\mathrm{rott}} f(N_{rot})\end{aligned} \tag{9}$$

The individual terms are obtained by multiple linear regression in order to reproduce experimentally determined binding affinities. f is a penalty function accounting for radius (Δr) and angle (Δα) deviations from the ideal geometries, and the number of frozen rotational bonds N_{rot}. The development of empirical scoring functions was pioneered by Böhm with his work on SCORE1, which was later incorporated into LUDI for docking[246,247]. Since then, a variety of empirical scoring functions has been developed with different empirical energy terms and varying sets of training data[210]. Empirical scoring functions are trained on interactions observed in crystal structures, and thus only attractive interactions are considered and additional repulsive terms have to be added to correctly account for repulsive contacts, *e.g.* clashes with the receptor[248].

Knowledge-based scoring functions

In contrast to empirical scoring functions, explicit binding affinities are not used in knowledge-based scoring functions. Instead, they utilize the statistical discrepancies between observed and expected distributions of atom-pair distances. The idea is that ligand-protein atom pairs frequently found at a certain distance are more likely to represent a favorable interaction. Binned and normalized occurrence frequencies of atom-type pairs are converted to pair-wise potentials using the inverse Boltzmann relation (Eq. 10)[1].

$$E(i,j) = -k_B T ln \frac{\rho_{ij}^{observed}(r)}{\rho_{ij}^{expected}(r)}, \quad (10)$$

where k_B is the Boltzmann constant, T is the absolute temperature and ρ_{ij} are observed and expected frequencies for atom types *i* and *j* at distance *r*. The final score is calculated as the sum of all protein-ligand atom pairs within a given cutoff distance. Interpretation of the scores is difficult as they often consist of hundreds of small contributions[234]. Popular examples of knowledge-based scoring functions used in docking are PMF[249] and DrugScore[240]. More recently, knowledge-based approaches have been combined with machine-learning methods[208]. Their particular advantage is the implicit inclusion of ligand-protein interactions, avoiding the explicit modeling of function terms and the arguable assumption of additivity[250].

The comparison of different scoring functions has been the subject of numerous studies[233,251,252]. Most of these studies conclude that current scoring functions are able to correctly predict the binding pose, but fail at correctly predicting the binding affinity and "virtually no correlations could be observed between the docking score and *in vitro* binding affinities"[233]. Schneider reasoned that "[…] flexible fit phenomena, the role of water molecules, protonation states in proteinaceous environments, and the entropic and enthalpic contributions and compensations upon complex formation are not satisfactorily addressed by the existing virtual screening methods"[11]. Keeping these caveats in mind, several successful prospective studies emphasize the practical benefit of molecular docking for virtual screening, especially in comparison to biochemical HTS[7,253]. While high-throughput docking was initially computationally too demanding for large-scale screens, this has changed with the constantly increasing computational power. In a recent study on dopamine D_3

receptor ligands, Shoichet and coworkers screened over three million compounds using high-throughput molecular docking[254]. About two trillion ligand-protein complexes had to be evaluated with DOCK3.6[255]. A receptor homology model and the recently available D_3 crystal structure was used to perform two separate screens. Interestingly, the observed overlap of the top 1,000 docking hits from the two screens was only 9%. The authors selected a total of 51 top-ranked compounds from both screens and experimentally confirmed 11 compounds as D_3 antagonists, six from the homology model, and five from the crystal structure, with affinities in the range of 0.2 and 3.1 μM.

The methods and respective applications presented in this chapter demonstrate the versatility of virtual screening approaches. However, virtual screening efforts are limited by the structural diversity of the screened compound libraries. The *de novo* design methods presented in the next chapter circumvent this by allowing the construction of novel compounds.

1.2 Computer-assisted molecular *de novo* design

While virtual screening and the corresponding biochemical HTS may be used to find active molecules in a library of synthesized molecules, *de novo* design methods are required to generate NCEs "from scratch"[15]. Whereas HTS is limited to a focused part of known chemical space, *de novo* design is intrinsically innovative and facilitates exploration of chemical space[1]. This may justify the increased costs per *de novo* designed molecule, compared to its HTS counterpart[256]. Most of the *in silico* counterparts to bench synthesis mimic the iterative drug discovery process. In comparison to static compound library screening, iterative *de novo* methods have the advantage that several dimensions (*e.g.* potency, toxicitiy, ADME) can be measured and optimized with the prospect of continuous improvement in each iteration[6]. The first programs for computer-assisted *de novo* design were introduced more than two decades ago, prominent tools being ALADDIN[257], GROW[258] and LUDI[196], and with their seminal article on peptides and peptide-like ligands Moon and Howe pioneered the field of automated computer-assisted *de novo* design[258]. Ever since, a plethora of different *de novo* design approaches have been published. In a recent publication, Schneider and Baringhaus referenced more than 50 *de novo* design tools[15]. D*e novo* algorithms have to address three pivotal questions in order to successfully design NCEs[259]:

1. How are new molecules assembled?
2. How is the quality of a molecule evaluated?
3. How to efficiently navigate in vast chemical space?

De novo design aims at exploring the chemical space, which Lipinski and Hopkins compared to the cosmic universe in its vastness, with chemical compounds populating space instead of stars[260]. Initially, molecules were constructed iteratively on an atom-by-atom basis. This offered the advantage of greatest possible flexibility by allowing the generation of any possible structure. A caveat of this concept is that the huge number of possibilities impedes systematic searching of the total search space, ultimately leading to a large fraction of chemically unstable or synthetically inaccessible designs without desired drug-like properties[1]. While chemical space is indisputably huge, its estimated size for synthetically accessible small molecules is up to 10^{200},[28] from which only a small fraction is relevant for medicinal chemistry[261].

Examples for focused chemical space are drug-like, biological active, and receptor-relevant chemical (sub)space[28]. This fact can be exploited by restricting the search for NCEs in these relevant regions of chemical space, essentially by taking smart shortcuts[15].

1.2.1 Fragment-based *de novo* design

With the introduction of fragment-based assembly strategies in computer-based design approaches, the search space has been reduced to a feasible size. By using larger molecular fragments instead of individual atoms, the number of decisions required to construct a molecule is considerably reduced, and most of the bonds present in the designed product are predetermined by the used fragments. Additionally, fragments often bind macromolecular targets with high ligand efficiency[262], which designate them as attractive starting points for drug discovery[15]. Most of the early *de novo* design programs were exclusively receptor-based, and products were directly assembled in the binding cavity of the protein. One approach is to superimpose different ligand-bound protein structures by backbone alignment and identify strategic ligand bonds brought to close proximity by the alignment. Ligands are dissected at these positions and recombined to yield new candidate molecules. An example of this approach is BREED[263]. A drawback is that the chemical variety in the designs is restricted by the fragments present in crystalized ligand-protein complexes. Related approaches rely on molecular docking techniques instead of experimentally determined crystal structures. In the software IADE, molecules are also initially fragmented, but instead of using fragments from other known ligands they are replaced with assumed bioisosteric analogs[264]. The reconstructed molecules are then evaluated using docking or field- and shape-based similarity. A related strategy employing bioisosteric replacement for *de novo* design is implemented in SHOP using the alignment-independent GRID similarity[265].

Another possibility is to automatically place fragments in the binding cavity guided by molecular docking. After preferred positions for seed fragments are determined, they can be developed to candidate molecules using fragment growing or linking strategies. Growing iteratively adds fragments to one starting seed fragment, while linking connects fragments in the binding site with suitable linkers[1]. A schematic

overview of the described fragment strategies is shown in Figure 7. If fragments commonly observed in drug-like compounds are used for assembly, the designed compounds are expected to more likely also resemble drug-like properties[1].

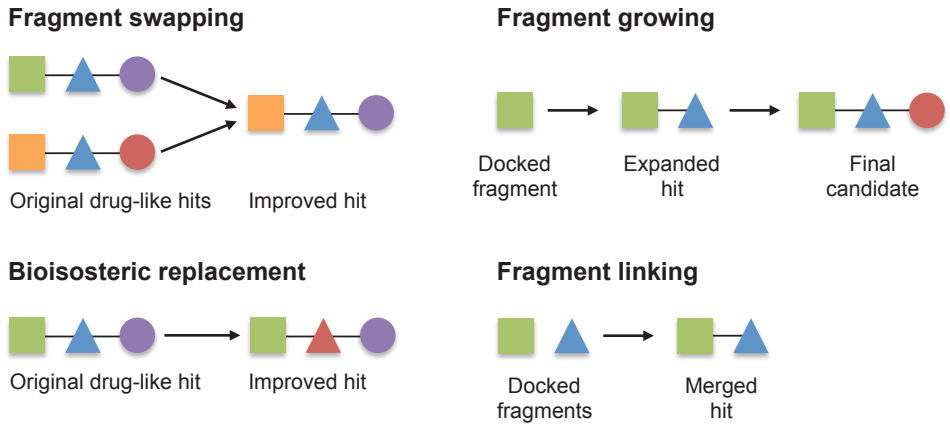


Figure 7. Fragment strategies for receptor-based *de novo* design. Same shape but different color indicates related receptor interaction properties. (Adapted from Ref.[3])

The fragment poses are evaluated using the methods described for molecular docking including force-fields as well as empiric and knowledge-based scoring functions (*cf.* Chapter 1.1.2, molecular docking). While most fragment approaches are appealing due to a fast design strategy, they suffer from relying on the assumption of additivity of the contributions of individual fragments to the docking score, and as a consequence to the binding affinity[262,266,267]. This assumption is only approximately true in cases where the binding mode and orientation of the individual fragments is also meaningful for the whole designed product[15]. Additionally, it has been realized that superadditivity (overproportional increase of binding affinity compared to individual fragment contributions) can occur due to the loss of rotational and translational entropy upon complex formation[268]. A possible solution might be to evaluate the whole product instead of individual fragments, which in turn eliminates the convenience of a small search space due to combinatorial explosion. Therefore, stochastic and deterministic local optimization strategies are required to identify high-scoring solutions[15].

An elegant method to obtain motivated fragments is by applying virtual *retro*-synthesis, most prominently the *REtrosynthetic Combinatorial Analysis Procedure* (RECAP)[269], to a library of drug-like molecules. RECAP defines eleven cleavable substructures based on reactions commonly observed in medicinal chemistry, which

are further used to dissect the molecules. By using the same rules for reassembly, designed products are more likely to be synthetically accessible, which is an important requirement for practical *de novo* design. The first software to implement a RECAP-based design scheme was TOPAS[270]. Context-dependent building block reactivity and physical availability of the required building blocks are challenges for the successful realization of designed compounds and in general it is not guaranteed that the anticipated reaction is possible[1,8]. One way to handle synthetic feasibility is to apply a synthetic accessibility filter in post-processing, *e.g.* implemented in SYLVIA[271], or to use synthesis planning software to recommend possible synthesis routes, *e.g.* employed in CAESA[272].

Reaction-based approaches are the latest incarnation of fragment-based methods that explicitly account for the synthetic tractability of the designed compounds[273]. Formalized reaction schemes, mirroring "real" synthesis protocols, are used as connection rules and readily available building blocks (*e.g.* purchasable from commercial vendors) as fragments. This does not only increase the likelihood to design synthetically feasible molecules but also suggest a synthesis route for immediate realization of the proposed compound. The software SYNOPSIS was among the first reaction-based *de novo* tools[274], in which virtual synthesis is guided by a collection of 70 organic reactions. Furthermore, an advanced protocol incorporating context-dependent reactivity ensured the eligibility of the building blocks for the chosen reaction. An apparent drawback of using only a limited set of reaction is the drastic restriction of the accessible chemical space, which could be problematic if, for example, novel targets are addressed. Several studies estimated the size of chemical space accessible using combinatorial reactions[275]. Schneider and coworkers calculated a size of 10^7 one-step products for a set of 58 robust reactions[276]. Another example is the *Pfizer Global Virtual Library,* which consists of up to 10^{18} virtual products, incorporating 1244 virtual reactions with more than 3000 experimentally validated synthesis protocols[277]. Compared to the estimated size of the total synthetically accessible chemical space (between 10^{20} and 10^{24})[278] there is still a distinctive difference, but this should only have a minor impact on practical drug discovery endeavors in the authors' opinion.

While fragment-based approaches drastically reduce the search space size, it still exceeds the size for exhaustive enumeration. Initially, deterministic design strategies

(*e.g.* breadth-first and depth-first searching) implementing the additivity principle were used with growing or linking approaches to quickly navigate the search space[259]. Stochastic search algorithms offer an alternative that can be used for optimization without relying on fragment additivity. While they cannot guarantee finding the globally optimal solution, they usually generate collections of acceptable candidate solutions in reasonable time[259]. Due to constituting a related problem, there is a profound overlap of methods used in molecular docking pose optimization. In the following, a brief overview of search methods frequently used in molecular docking programs is given. A detailed overview of search strategies used in popular *de novo* programs is also given in the literature[259,279]. A widely used stochastic search strategy in early design programs was *Monte Carlo* (random) sampling combined with the *Metropolis Criterion* (MCMC) to guide the search process[259]. The first program implementing this strategy was CONCEPTS with an atom-based strategy[280]. Among the first to use MCMC for fragment-based design was SMoG, using a knowledge-based scoring function to evaluate the randomly designed candidates for their protein-binding potential[281]. Evolutionary algorithms are the second group of search methods that were frequently adopted by *de novo* design programs. Algorithms belonging to this group are genetic algorithms, genetic programming[282], and evolution strategies[23]. Glen and Payne reported one of the earliest applications of a genetic algorithm to guide the compound construction in a fragment-based *de novo* process in their Chemical Genesis software[283]. Other prominent examples using evolutionary algorithms include TOPAS[270] and SYNOPSIS[274]. With the focus of recent *de novo* design programs on just a few reliable sampling methods, the evidence of several retrospective and prospective studies regarding their ability to identify suitable solutions suggest that "[…] one might consider the task of chemical space navigation solved."[8].

Ligand-based *de novo* design is a complementary strategy for receptor-based design, which is applicable without a known protein structure. While the same concepts for compound construction and sampling are applicable, the scoring during the design process has to be done independently of the macromolecular target structure. Therefore, candidate designs are compared to one (or multiple) reference compounds in a predefined descriptor space, yielding a measure of similarity or even a prediction of relevance. The chemical similarity concept and the methods used are

equivalent to the ones previously described for virtual screening (*cf*. Chapter 1.1.1). Among the first to describe a solely ligand-based design tool incorporating fragments was the work of Globus *et al.*, based on a genetic graph algorithm and all-pairs-shortest-path similarity to a single reference compound[284]. In a contemporaneous study, the software TOPAS was introduced, which employed an evolutionary algorithm[23] to navigate the search space, but used pharmacophore and substructure fingerprint molecular similarity for scoring the designed products[270]. TOPAS was used in several seminal prospective drug discovery studies at Roche, including the design of inhibitors for the human Kv1.5 potassium channel[270], and reverse agonists for the cannabinoid-1 receptor (CB-1), which were also validated in *in vivo* mouse models[285]. It is noteworthy that in the CB-1 study the whole process from *de novo* design to hit series identification was accomplished in four months.

Even though the majority of ligand-based *de novo* studies utilize structural or pharmacophoric similarity for evaluating the designs, it is also possible to make use of problem-specific QSAR models. *De novo* design with a QSAR scoring function is closely related to *inverse QSAR*, which aims at constructing molecules closely matching a given predicted optimal position in the descriptor or property space[286]. The first software tool to integrate a QSAR scoring function was PRO_LIGAND[287], using a scoring approach similar to *Comparative Molecular Fields Analysis* (CoMFA)[288]. At the same time, a prospective study combining LUDI (*de novo* design) with GRID[195] (QSAR scoring) was published[289]. In this prospective study, LUDI and GRID were employed to identify possible substituents for a known inhibitor of human synovial fluid phospholipase A_2 receptor in order to increase affinity. Since then, several additional *de novo* tools incorporating QSAR methods have been published[259].

There is also increasing interest in applying advanced machine-learning techniques to *de novo* design, which provide a "top-down" knowledge-based scoring concept, complementing the "bottom-up" concept of established methods. Not surprisingly, this mirrors the ongoing development in the virtual screening community and is fueled by the increasing amount of publicly available data mandatory for model building. Even though the first studies describing the design of peptides using ANN machine-learning methods have been introduced more than 15 years ago[290], there are only few small-molecule *de novo* design studies incorporating machine-learning

techniques. A distinguished example recently published by van der Horst *et al*. features the combination of a genetic algorithm for compound design with adenosine subtype specific SVM regression models for scoring[291]. It is of interest to note that these authors did not include the synthetic feasibility as a design objective, which could be a reason why they did not select designed compounds for synthesis. Instead, they identified promising scaffold patterns in the designed library and experimentally validated two out of six selected scaffolds, decorated with methyl substituents.

Numerous structure- and ligand-based *de novo* design success stories have been published in the last 25 years and automated *de novo* design has proven to deliver innovative hit- and lead-compounds[1,3,7,8,292,293]. Consequently, *de novo* design can be considered an established tool in the toolbox of commercial drug discovery programs[8,205,294].

1.2.2 Multi-objective optimization

Developing a drug is a balancing act between a broad set of different objectives, most prominently the affinity to the main target (or a multi-target panel), selectivity, and pharmacokinetic / toxicological properties (ADMET; *Absorption, Distribution, Metabolism, Excretion, Toxicity*). Deficiencies in ADMET properties are the leading cause of failure in late stages of drug development or withdrawal of already marketed drugs[295]. Historically, the development was done sequentially, first optimizing affinity and afterwards taking care of additional objectives[296,297]. With the increasing pressure on drug development projects due to escalating costs of drug discovery[6] and the realization that considering the crucial ADMET properties as early as possible in the discovery process might be beneficial for the overall success[295], this paradigm is starting to shift towards multi-objective optimization approaches[297]. By simultaneously considering multiple primary objectives in candidate solutions, the amount of trial and error design rounds might be reduced (Figure 8a), ultimately leading to a more efficient discovery process, while avoiding potential liabilities that could lead to later-stage failures[298].

In a multi-objective problem, a suitable solution must fulfill two or more primary objectives. In the case of conflicting objectives, a compromise has to be made and the identification of candidate solutions is in general a non-trivial task[298]. If multiple

objectives have to be considered simultaneously, costs increase sharply for conventional biochemical HTS approaches. Complementary computational methods including multi-objective library and *de novo* design might be cost-effective alternatives[299]. A straightforward way to consider multiple objectives is to combine them in one aggregated weighted-sum score, which can then be used for conventional single-objective optimization[300]. In the case of contradicting objectives mean-sum methods have the drawback that solutions are ranked best that are mediocre for all objectives, which is possibly not ideal[1]. An alternative is to predefine desired values for the individual objectives and use a geometric mean of desirability functions for optimization (Figure 8b)[301]. The software MOOP-DESIRE is an example for a desirability-based multi-objective optimization applied to library design[302]. The same group also used MOOP-DESIRE for designing non-steroidal anti-inflammatory drugs with optimized analgesic, anti-inflammatory, and ulcerogenic pharmaceutical profiles[303].

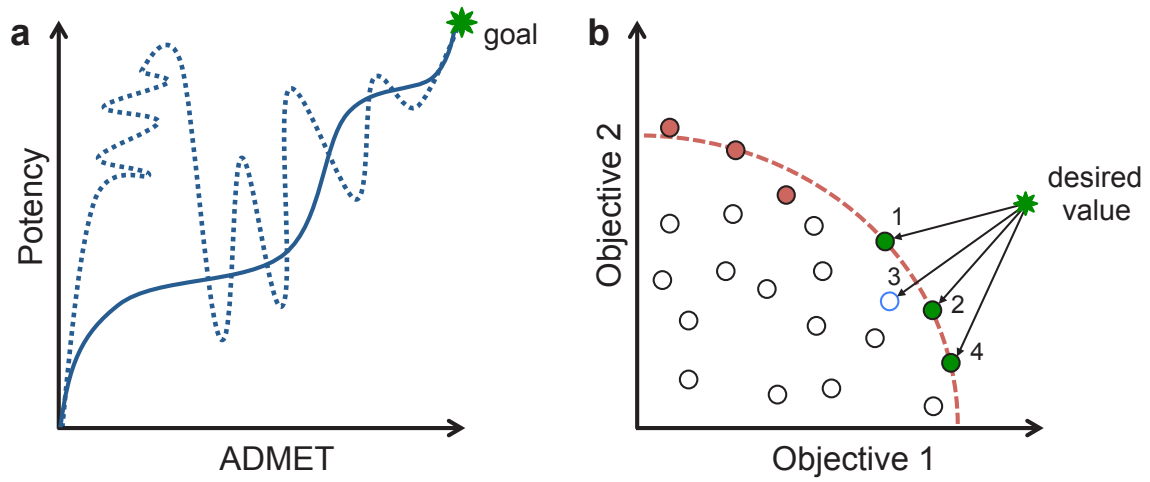


Figure 8. (**a**) Drug discovery strategies. Sequential single-objective optimization for two objectives (dashed line). Initially potency is optimized and consecutively ADMET properties. Multi-objective optimization methods simultaneously consider both objectives enabling more efficient convergence (solid line). Green star is the anticipated optimum. (Adapted from Ref.[297]) (**b**) Multi-objective prioritization. The dashed red line represents the Pareto frontier and filled circles are Pareto-optimal solutions. Green circles are solutions closest to the desired multi-objective value with ranks according to the geometric mean of the desirability functions. (Adapted from Ref.[100])

Pareto optimization does not require an explicit definition of individual weights or optimum values. Instead, it identifies a set of solutions (the Pareto frontier) that is optimal in the sense that no other solution dominates it in all objectives. The set of pareto-optimal solutions represents various compromises among the objectives and allows the user to choose the most suitable solutions[300]. With two objectives, the

Pareto frontier is a curve and can easily be visualized (Figure 8b). With more objectives it becomes a surface or hypersurface. It was already introduced in 1989 by Goldberg with a genetic algorithm to identify the ensemble of Pareto-optimal solutions[230]. Among the first to use Pareto optimization in the field of chemoinformatics were Gillet *et al.* with their study on library design using the MoSELECT software[304]. Soon after, Brown *et al.* introduced Pareto-optimization for *de novo* design[305]. Another recent example is the design software MEGA[306]. One apparent drawback of Pareto-based compound prioritization is that with an increasing amount of objectives, the number of solutions on the Pareto-optimal hypersurface tends to increase exponentially[307].

If there is a primary and several secondary objectives, an unsophisticated yet effective approach is to design candidates for the primary objective and filter (or flag) the designed solutions afterwards with respect to the remaining objectives[308].

1.3 Polypharmacology

One of the dominant objectives of drug discovery projects in the last decades has been to find NCEs that selectively modulate one single macromolecular target with high potency[5,309], so-called "magic bullets"[5]. This concept was already identified by Ehrlich in 1904[310]. The "one drug-one target" strategy is closely connected to the linear causality model of "one protein-one gene-one disease", which have mainly influenced the definition of standard drug design protocols we encounter today[299,309]. While there has been notable success with this single-target centric approach for diseases with clearly defined cause and mechanism, it is now recognized that the incomplete knowledge of target networks and drug interaction profiles have strongly biased the perception of drug selectivity[309]. Recent studies indicate that most therapeutically effective drugs interact with multiple proteins and selective drugs are less frequently encountered[309]. Studies investigating drug promiscuity estimate that, on average, drugs interact with approximately six targets[311,312] and only 15% of analyzed drugs interacted solely selectively[309]. Given the limited assays' availability compared to the total number of potential biological targets, these numbers are likely to increase with additional targets being investigated[309].

Multiple on-target binding is perceived to be essential for efficacy, while binding to off-targets induce undesired side effects encountered in many marketed drugs[313]. Especially for complex, polygenic diseases, including the majority of neurological diseases and cancer, a polypharmacology design is essential, and several marketed drugs already exhibit a distinctive multi-target profile[4,313,314]. An illustrative example is the antipsychotic drug clozapine, already discovered in the 1960s[315], which exhibits a highly complex pharmacological profile with high affinities for a variety of GPCR receptors including serotonin, dopamine, muscarinic and adrenergic receptors[316]. While clozapine is extremely effective in treatment of schizophrenia, and in reducing suicidal tendencies, it also has severe side effects, including agranulocytosis, seizures, weight gain, and diabetes, caused by unwillingly targeted GPCR receptors[316]. Identifying the appropriate subset of receptors responsible for efficacy, while avoiding side effect-inducing ones, is clearly a future strategy for further drug development[4].

Another fact facilitating polypharmacology drug design is the realization that many diseases are connected to multiple compensatory signaling pathways, which are often

found to be remarkably resilient to perturbation[317]. This realization is supported by large-scale genomics studies showing that single-gene knockouts targeting a single pathway frequently have no effect on the phenotype and multiple simultaneous knockouts are required in order to induce a phenotypic effect[317]. An example is the complementarity of the mitogenic Ras/mitogen-activated protein kinase (MAPK) and survival phosphoinositide 3-kinase (PI3K)/Akt pathways[318]. Targeting just a single pathway in tumor or leukemia treatment has proven ineffective due to extensive cross talk and compensatory regulations between the pathways[318]. Drugs suffering from this are kinase inhibitors selectively acting on individual proteins in either of the signaling cascade. Blocking both pathways at the same time by combining PI3K and MAPK selective inhibitors has shown to be beneficial[319].

These findings implicate that systematically searching for multi-target drugs globally affecting disease-associated networks should be considered more often instead of solely relying on designing selective drugs[4]. Additionally, it has been realized that the rational design of multi-target profiles could be beneficial for drug efficacy as well as drug safety[309].

1.3.1 Multi-target drug discovery

Combination therapy, *i.e.* the simultaneous administration of a mixture of selective drugs with different mechanisms of action, is a well-established therapeutic approach to treat polygenic diseases[5,299]. Prominent examples are anticancer chemotherapy, the treatment of infectious diseases, and the treatment of central nervous system diseases[320-322]. While these therapies are effective, they suffer from the combined side effects of the individual drug entities, and the risk of possible drug-drug cross interactions[299,323]. It is also challenging to balance the differences in bioavailability, pharmacokinetics, and metabolism[323]. Combining therapeutic effects of the drug cocktail in a single compound containing multiple biological properties could be of advantage compared to combination therapy[5,324]. The design of multi-target ligands for predefined polypharmacology profiles can be challenging[100,299]. Essentially a set of potentially contradicting objectives is added to the already complex multi-objective drug discovery problem. Two distinct strategies have emerged for the rational multi-

target ligand design: The combination of pharmacophores derived from selective ligands, and data mining in chemogenomics databases[5,299].

The combination of pharmacophores observed in selective ligands is currently the most commonly used method for designing multi-target ligands[5]. To achieve the desired *in vivo* efficacy profile and therapeutic effect, the affinities towards the individually addressed targets have to be carefully adjusted[5]. This can be a challenging task. Dual-target compounds are most frequently designed instead of poly-target compounds to limit the costs associated with multi-target strategies[299]. After promising targets located on complementary pathological pathways with known selective ligands are identified, the observed pharmacophoric patterns contributing to selective binding can be joined in a single molecule. Comparison of binding pockets could be used to assess *a priori* if the selected targets can potentially be addressed with a single molecule[325]. A straightforward way to achieve dual-target ligands is to synthesize drug conjugates by combining two selective ligands with a cleavable linker. Frequently, an ester linker is used that is cleaved by plasma esterases to release the individual drugs[5]. Conjugates can also be linked permanently if a linking position can be identified that does not affect the binding affinity of the individual drugs. Drug conjugates are often larger and more complex compared to the selective drugs, which is associated with problematic pharmacokinetics, lower *in vivo* efficacy, and limited oral availability[326]. Optimization of pharmacokinetics while retaining dual-target affinities of drug conjugates is particularly challenging[5]. To gain candidate compounds with drug-like properties, pharmacophores can also be fused or completely integrated in a single pharmacophore profile to limit the compound size and complexity[5]. The identification of pharmacophore hybrids can be supported by computational methods. In a recent study, Achenbach *et al*. employed their multiSOM approach to identify hybrid 5-lipoxygenase (5-LO) and soluble epoxide hydrolase (sEH) dual-target ligands[327]. Characteristic substructures were identified for each target, and target-specific SOMs were built to represent the target profiles. The authors screened fragment libraries with combined SOMs, and discovered fragments binding both targets. One of the fragments was further expanded to a ligand efficient dual-target inhibitor (5-LO 0.05 μM; sEH 0.17 μM). In general, fragments seem to offer a valid

starting point for multi-target drug development due to their inherent binding promiscuity[299].

Data mining of large chemogenomics databases is an alternative approach to design compounds matching a multi-target profile. Without the limitations associated with the combination of pharmacophores, it might be possible to simultaneously address several target-objectives[299,328]. Usually, such polypharmacology profiles are a combination of desired on-target modulation, and avoidance of side effect related off-targets[328]. In the SOSA approach, a library of drugs is screened for compounds partially matching an aspired target profile[329]. Using traditional optimization, the initial observed profile is then modified to yield the desired multi-target activity profile, possibly exchanging the main and side activity in this process. Starting with an existing drug should increase the likelihood of obtaining analogs with favorable pharmacokinetics and safety profiles[329]. In a recent study, Besnard *et al*. combined target-specific naïve Bayes models with a genetic algorithm to adaptively evolve known drugs to a desired multi-target profile[100]. The authors implemented a ligand transformation scheme based on a collection of medicinal chemistry transformations frequently used in the literature. The applicability to polypharmacology design was demonstrated by evolving an approved acetylcholinesterase inhibitor into a brain-penetrable ligand with a distinctive GPCR polypharmacology profile. In this study, More than 800 predictions were validated experimentally, and 75% of the predictions were confirmed.

1.3.2 Target prediction

With the revival of compound screenings based on cell or organismal phenotypically readouts, there is an increasing demand of identifying biological targets for orphan chemical entities, *i.e.* molecules without a known macromolecular target[330,331]. Another trend that could profit from target identification is the repurposing of approved drugs for novel therapeutic indications. With the increasing complexity and requirements of clinical trials and additional hurdles in the approval process for novel drugs, repurposing offers an alternative that effectively reduces risks in clinical activities[6,332]. A third area of application for target identification methods is the

prediction of side effect-related drug-target interactions to identify potential clinical risks early in the drug discovery process[330,333].

Conventional biochemical approaches for target identification include affinity purification using immobilized compounds and "pulling-down" target proteins from cell extracts, yeast three-hybrid technique with hybrids of small molecules and cDNA libraries, and the screening of protein micro arrays with small-molecule probes[330]. Experimental constraints, *e.g.* different levels of proteins in cell extracts, can prevent the straightforward application. Therefore, they can be complemented by computational target identification methods, including chemical similarity searching, data mining in annotated chemogenomics databases, and panel docking[331].

Similarity searching with any molecular descriptor and similarity metric can be used for target identification[331]. A query compound is compared to a database of known ligands with annotated targets. The annotated targets of the most similar compounds point to potential targets for the orphan compound. There are several issues with this simplistic approach[331]: When are two compounds similar enough to be considered for target prediction? If there are several targets, how are they ranked? How is the prediction influenced by target class bias in the database?

By clustering the database with an unsupervised learning method, *e.g. Self-Organizing-Maps* (SOMs)[108,109], some of the issues are addressed. The SOM algorithm detects cluster boundaries in the high-dimensional descriptor space, which can be interpreted as applicability domain for target prediction[334-336]. Only compounds inside a SOMs neuron perception field are used for target identification. Schneider *et al.* used a SOM trained on their expert curated collection of bioactive reference compounds[37] to identify protein-tyrosine phosphatase 1B as a new target for aspirin[37].

A potential target bias can be eliminated by comparing whole sets of target-specific ligand instead of simple pair-wise similarities[331]. A recent example for such an approach is the *Similarity Ensemble Approach* (SEA)[337]. SEA calculates the similarities between molecules of two sets and retains all similarities above a predefined threshold. The sum of similarity scores is then normalized based on an empirical extreme value distribution model for random distributed sets of equivalent size in order to avoid a bias due to different set sizes. Finally, an *Expectation value* (E-value) is calculated that expresses the likelihood of seeing a score at least as high just

by chance, motivated by the BLAST pairwise sequence alignment theory[338]. The E-value can also be used to rank target predictions. Keiser *et al*. successfully applied SEA for drug repurposing and side effect-related off-target prediction[339]. The authors predicted targets for 878 US *Food and Drug Administration* (FDA) approved small-molecule drugs, based on 246 target sets drawn from the *MDL Drug Data Report* (MDDR) database[340]. 30 predictions were experimentally tested and 23 new drug-target associations could be confirmed. As an example, the dopamine D_4 receptor was identified as off-target for the marketed drug doralese with even greater potency compared to the known therapeutic targets $\alpha_{1A/B}$ adrenergic receptor. Several of the 23 confirmed off-targets belonged to the aminergic GPCR family, a class with well-known cross-activity[339]. Nevertheless, also four unexpected cross-boundary off-targets were discovered in the study. In a related effort, Mestres and coworkers developed a ligand-set approach based on the group fusion similarity searching technique. The authors predicted potential off-targets for 767 drugs, using a database of 109,766 compounds annotated to 684 therapeutically relevant targets[341].

Chemogenomics databases can be analyzed with data mining methods to infer activity spectra for target panels[331]. Inductive machine-learning methods are commonly used for this task[331,342]. Already in the 1990s, the software PASS was introduced to predict activity spectra for molecules[343,344]. It is based on a multilevel neighborhood atoms descriptor and linear regression to yield a binary classification for each target with probabilities of being active or inactive. Initially, predictions for 300 pharmaceutical effects were included, which was later extended to more than 2000 using a library of over 60,000 annotated compounds[345].

Naïve Bayesian classifiers are frequently used to estimate class labels from large collections of data points (*cf*. Chapter 1.1.1, machine-learning). Multiple-category Laplacian-modified naïve Bayes models have been used by Nidhi *et al*. for the prediction of 964 targets[346]. The models were trained on the WOMBAT chemogenomics database[347] using binary circular substructure fingerprints (ECFP)[36]. A retrospective evaluation, using 85% of the WOMBAT database as training data and predicting the remaining 15%, indicated a high accuracy of the method. Considering the top three predictions for each query compound, 92% of the

compounds were correctly assigned to their annotated targets. In a second study, targets for all MDDR database compounds were predicted to systematically deconvolute the generic therapeutic annotation in the MDDR database to specific macromolecular targets associated with the effect[346]. In a similar approach, Paolini *et al*. used Naïve Bayes models with a large-scale database containing 4.8 million molecules, of which over 275,000 molecules were classified as being biologically active[348].

While molecular docking has a long history in receptor-based drug design, it has only recently been investigated for its target identification potential[349]. Instead of docking a collection of molecules to one binding site, the opposite strategy is used in inverse docking approaches[349]. Molecules are docked in parallel to a set of prepared target protein structures, and the individual docking scores are used for ranking the targets. Compared to ligand-based approaches, this strategy is computationally demanding and requires the appropriate preparation of a heterogeneous collection of protein binding pockets suitable for small-molecule docking on a large scale. Inverse docking suffers from the inaccuracy of current fast scoring functions and a high false negative rate[349]. Despite these drawbacks, it has been successfully applied in prospective target fishing projects, including the identification of potential targets for natural products[350] and to "deorphanize" a small focused library containing a novel chemical scaffold[351]. Other receptor-based approaches include the comparison of protein-ligand binding sites[352] and multi-target pharmacophore models[353].
Additionally to the described chemoinformatic tools, it is also possible to identify new targets for small molecules based on the phenotypic response they elicit. For example, targets can be related using drug side effects obtained from literature mining[354], or experimentally by means of gene-expression profiles, which have to be thoroughly computationally analyzed in order to make robust predictions[355].

1.4 Chemical space visualization

Understanding the distribution of molecules in chemical space is essential for answering the fundamental question in a drug discovery project: Which molecule should be synthesized next? In a sense, it is a "voyage to the unknown", in which an excellent navigation is required to avoid rough terrain, and find an easy and efficient route to the desired goal[356]. For large collections of compounds combined with available knowledge about their respective biological activity, different approaches have been formulated on how to build informative maps, which can provide an intuitive access to the data and allow analysis, visualization, and navigation of the underlying chemical space[15,28,260,357-359]. Initially, visualization was solely used to investigate the diversity of libraries and find uncovered areas of chemical space[360]. With the continuing use of biochemical HTS also for novel biological targets, the composition of the supporting compound libraries has to be constantly revised. Instead of solely relying on maximizing the number and diversity of compounds, focused libraries are increasingly used for screening[20,361]. For designing appropriate focused libraries, visualization techniques can be applied to identify and explore regions of chemical space with potential biological interest and to understand the content and relative distribution of compounds in a library[362]. An early example of a software incorporating different visualization methods with the purpose of providing data mining capabilities to drug discovery researchers is ChemSpaceShuttle[363].

In drug discovery, multivariate datasets are frequently encountered, with data originating from *in silico* calculations, phenotypic screenings, biological assay panels, ADMET profiling, or functional genomics[28]. The most basic way to visualize such data is to create histograms for individual properties and compare those by visual inspection or statistical methods[357]. This type of analysis is usually restricted to the comparison of a few properties. To overcome this restriction, visualization methods project data from a high-dimensional space into an easily accessible lower-dimensional space (two or three dimensions for reasons of visualization), while preserving most of the relevant information[359,364].

1.4.1 Dimensionality reduction

For dimensionality reduction, unsupervised learning methods are frequently applied, which aim at finding hidden structures in high-dimensional data[28,365,366]. Two methods commonly applied are *Principal Component Analysis* (PCA)[367] as well as the closely related *Multi Dimensional Scaling* (MDS)[368]. PCA finds the direction in the data in which the data varies most, and the potentially correlated variables are transformed in a set of uncorrelated variables. PCA preserves the covariance structure of a set of variables by computing the eigenvectors and eigenvalues of the data covariance matrix. The data is linearly projecting, according to the matrix of eigenvectors. For visualization purposes, only two or three eigenvectors with the greatest eigenvalues are used for projection, while the others are discarded. In MDS, a low-dimensional mapping is computed that preserves the pair-wise distance matrix by minimizing a distance based cost function. In the case of classical MDS, a Euclidean distance matrix is used, and the low-dimensional mapping found is identical to the PCA solution[369].

In contrast to linear projections, Kohonens' SOMs[109,370], initially introduced to chemistry by Zupan and Gasteiger[371], is an unsupervised learning algorithm. It is capable of revealing nonlinear relations whilst preserving the topology of the data by approximation of the data density. It is closely related to artificial neural networks as it is based on a regular lattice (usually two-dimensional) of artificial neurons whose weights are adapted to match the training vectors in high-dimensional data space. Neurons are connected to their adjacent neighbors and influence each other during training according to the neighborhood function. Influence is gradually reduced during the training process, enabling adaption to the global shape in the beginning as well as adaption to finer local structures at the end of the optimization. In the SOM projection the local neighborhood is preserved. Data points located in close proximity in the planar SOM projection were also close in the high-dimensional data space.

SOMs have been widely applied in drug discovery[356,372], *e.g.* for identifying compound clusters belonging to specific target families[373,374], comparing or shaping compound libraries[375,376], or for designing target-specific or multi-target modulating ligands[327,377].

A probabilistic extension of the SOM is the *Generative Topographic Mapping* (GTM) algorithm[378]. GTM generates a probability distribution in the high-dimensional data

space by means of low-dimensional latent variables. The nonlinear transformation from the latent to data space is achieved using a RBF network. By analogy with the SOM algorithm, a regular grid of nonlinear basis functions covers the latent space with associated radial symmetric Gaussians in the data space. Due to the probabilistic definition, Bayes' theorem can be used to calculate the inverted projection from data to latent space. A GTM is trained using the *Expectation-Maximization* algorithm to maximize the likelihood of the data with respect to the model. For a HTS campaign covering five GPCR targets, Maniyar *et al.* applied GTM to gain informative and discriminative visualizations and compared it to PCA and SOM[359]. The authors also used an extension to GTM, *Hierarchical GTM* (HGTM)[379], which combines multiple GTM models in a tree hierarchy, each focusing on different sub-spaces of the data space. For the dataset used in the study, visualizations generated with GTM and HGTM showed a clearer separation of target clusters compared to PCA or SOM.

In the distance preserving *Nonlinear Mapping* (NLM)[380], each data point is represented in the low-dimensional mapping and all relative distances are taken into account. Thus, distances between points on the map are directly related to the similarity in high-dimensional space. Mappings are found by minimizing Sammon's stress function[380] by means of standard optimization methods. Due to the preservation of all relative distances, without special focus on *e.g.* local distances, the global shape of the compound distribution is preserved. In 2003, Agrafiotis introduced an approximation for nonlinear mapping, suitable for large datasets commonly found in drug discovery applications[381].

Methods focusing on preserving the local neighborhood have recently been proposed[382-385]. The assumption is that the data is truly located on a low-dimensional nonlinear manifold. To truthfully represent data on or near a low-dimensional manifold, it is crucial to keep the low-dimensional representations of similar data points close together, while in traditional linear mappings it is important to keep dissimilar data points far apart[386].

A detailed overview of dimensionality reduction with a focus on drug discovery application is given in Chapter 8.1.

1.4.2 Structure-activity relationship visualization

If a set of reference compounds with additional properties is available, it is possible to model a property landscape of this data (with the activity landscape being a special case). Visualization methods complement mathematical QSAR methods in a graphical and intuitive way. They are not only suitable for analyzing congeneric compound series during compound optimization, which is the primary field of application for classical QSAR approaches, but are also applicable for hit identification in combination with structurally diverse sets of compounds from different sources[387]. Visualization methods are descriptive in nature and do not directly aim at the prediction of new compounds[358]. They can aid medicinal chemists in rationalizing SAR features based on increasingly large amounts of diverse data. When used in conjunction with complementary predictive methods, visualization can act as a guide during the modeling process[15]. While SAR analysis can also be accomplished by experienced medicinal chemists, solely relying on their chemical knowledge and intuition, the size of the datasets to be analyzed usually prevent manual in-depth analysis, and favors computational approaches for decision support[388]. In general, activity landscapes integrate molecular similarity and potency information, and are suitable for large-scale SAR analysis[388]. This rather broad definition covers a diverse set of different methodologies[388,389]. In this chapter we focus on four types of activity landscape visualization:

1. Analog-centric visualization
2. Comparisons of structural similarity and activity similarity
3. 2D / 3D surface representation
4. Similarity networks

A simple way to assess the SAR for a series of related molecules (chemical analogs) is the R-group analysis[390]. Within this analysis, molecules are decomposed into core scaffolds and attached R-groups, which are then related to the observed activities. A recent example is SAR Map, which arranges the decomposed R-groups in a rectangular grid. Each cell represents a single compound, color-coded with the corresponding activity[390]. The inherent drawback is that only SAR datasets arranged around a specific scaffold are suitable for this analysis and multiple examples for each R-group are required to gain additional SAR insights. By considering molecules not

only as a set of R-groups but as a whole, and relating them by means of global molecular similarity, it becomes feasible to investigate structural diverse datasets.

Plotting the pair-wise structural similarity versus the difference in activity is a straightforward way of analyzing a structure activity dataset[391,392]. In *Structure-Activity-Similarity* (SAS) maps, all compound pairs are systematically compared, and each pair is represented by a point in the scatterplot[391-393]. Structural similarity is calculated based on any structural representation, and using an appropriate similarity metric. Activity similarity for compounds *i* and *j* is calculated as

$$S_{ij} = 1 - \frac{|A_i - A_j|}{A_{max} - A_{min}}, \tag{11}$$

where A_i gives the activity of compound i and $A_{max} - A_{min}$ the range of observed activities in the whole dataset. Points can be color-coded according to the activity of the compound pair (*e.g.* sum or maximum of paired activities), adding an additional layer of information[388]. As all possible compound pairs are depicted in SAS maps, the number of points in the scatter plot increases quadratically with the number of compounds, thus limiting this approach to the analysis of small sets of important compounds[393].

Different areas in the SAS plot can be assigned to distinct SAR characteristics[388]. Of special interest are "scaffold-hopping" (low structural similarity – high activity similarity), and "activity cliff" areas (high structural similarity – low activity similarity) (Figure 9a). The term "scaffold-hopping" was introduced by Schneider *et al.* with their use of topological pharmacophore search for virtual screening[14], while "activity cliff" was first described by Lajiness[394]. Scaffold-hopping potential is a desired behavior for a given chemical reference space, as it enables medicinal chemists to extensively modify a structure while still maintaining biological activity[62]. On the other hand, activity cliffs are problematic as small changes in the structure lead to unexpected changes in activity. One has to keep in mind that activity cliffs are an artificial concept and reflect the fact that we are still lacking an appropriate chemical reference space representation capturing the important factors determining biological activity[26]. It might not be surprising that relatively small structural changes have the potential to drastically influence activity, depending on the protein environment the molecule interacts with upon binding. The influence of

"magic methyl" on protein-ligand interaction is a prominent example, where the addition of a single methyl can result in an up to 100-fold activity boost[395]. The addition of an unsatisfied hydrogen bond donor can also have drastic energy consequences, *e.g.* burying a hydroxyl group in hydrophobic pockets lead to a desolvation penalty of up to 21 kJ/mol, which translates to a activity loss of 3.7 log units[396]. Activity cliffs can reveal substitution sites essential for activity, and thus have high SAR information content. In general they are nonetheless not necessarily beneficial for understanding global trends that could help improve activity[388].

An example of SAS map visualization for a focused QSAR dataset is shown in Figure 9b. The benchmark dataset by Rücker *et al.*, previously used to evaluate QSAR performance[169,397], consists of 144 PPARγ ligands with pK_i affinities obtained by scintillation proximity assays. This result in 10,244 compound pairs included in the SAS map. For the molecular representation Morgan fingerprints[35] (radius 2, 1024 bits) were used. The number of depicted compound pairs leads to a crowded plot, which limits the expressiveness of this visualization. In Figure 9c, representative pairs from each SAR quadrant are shown for one origin compound. Several detailed perspectives on activity cliffs and related concepts have recently been published[393,398-401].

For the global analysis of compound sets combined with additional properties, 2D/3D landscapes are an especially intuitive visualization. More than two decades ago, Kauffmann *et al.* introduced rugged fitness landscapes for the analysis of species coevolution[402]. For SAR analysis, Maggiora and coworkers were among the first to describe the concept of theoretical schematic 3D landscape views, which are reminiscent of topographic maps or actual landscapes[403]. For the categorization of SAR phenotypes, they compared areas of continuous SAR (small distances in chemical space that lead only to moderate changes in activity) to "gentle rolling hills" of a smooth landscape, while discontinuous SAR (small distances in chemical space that have large impact on activity) were compared to rough landscapes[403]. The different characteristics have a direct impact on the amount of compounds required to sufficiently represent a specific area. While only a few compounds are required for continuous SAR regions, extensive sampling is necessary for areas of discontinuous SAR[403].

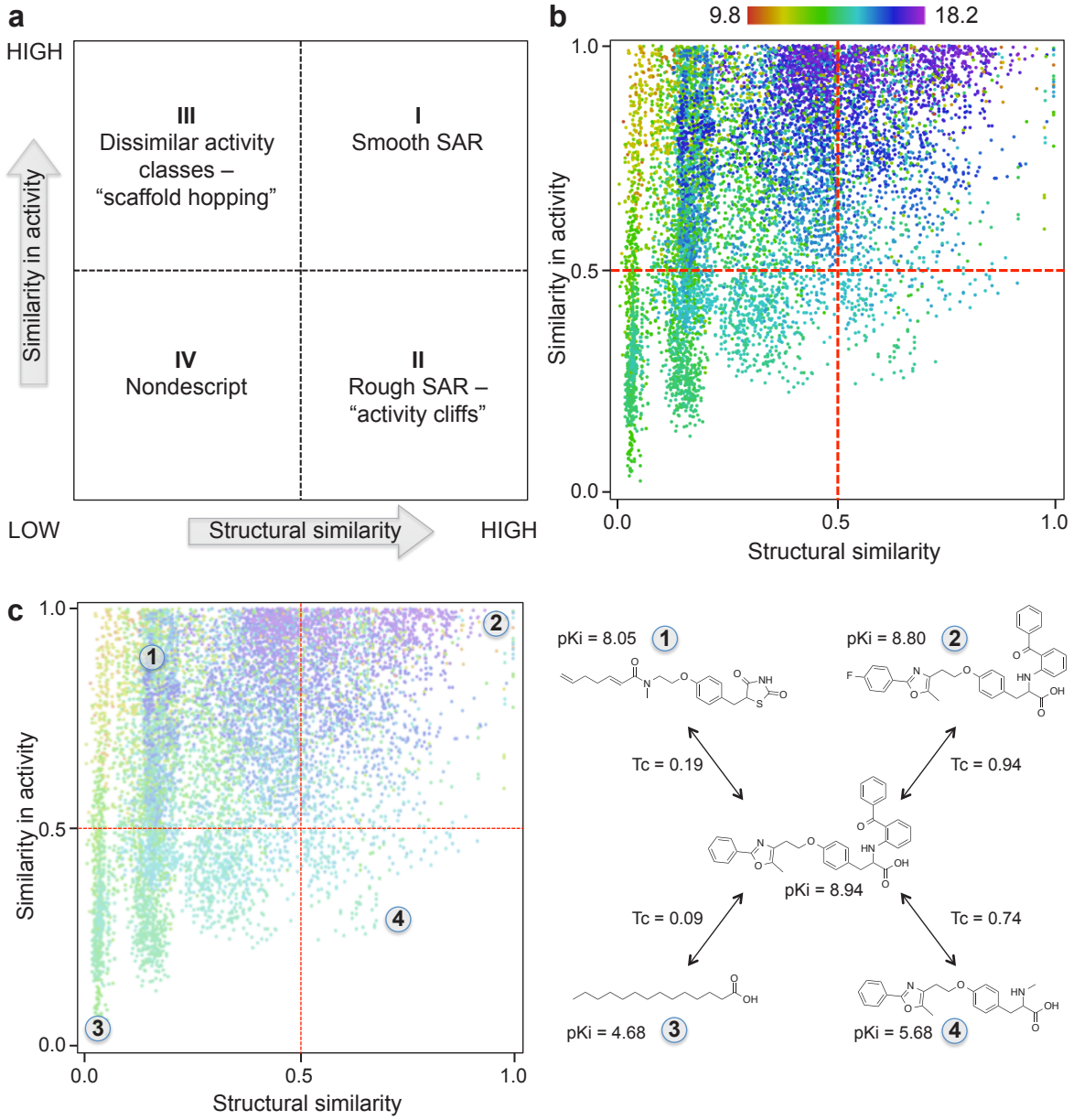


Figure 9. SAS maps (**a**) Schematic representation of a SAS map with characteristic SAR areas **I-IV** highlighted. (**b**) SAS map visualization of a focused dataset with 144 PPARγ pK_i measurements[397]. Activity similarity was calculated using Eq. 11, and molecular similarity was calculated using the Tanimoto coefficient with RDKit Morgan Fingerprints (radius = 2) as molecular descriptor. Coloring according to the sum of activities. (**c**) Example compound pairs taken from each of the SAR quadrants. The central compound is the same in all depicted pairs. Numbers indicate the position of the respective pair in the SAS map.

For the visualization of SAR datasets, molecules are described by molecular descriptors, which span a high-dimensional chemical space. Dimensionality reduction methods are applied to reduce the dimensionality and define a sparsely distributed 2D chemical (reference) space. Additionally, compound positions are "decorated" by an auxiliary property as a third dimension, most prominently biological activity. To approximate a continuous and smooth surface from the sparsely distributed compound property values (data points), interpolation or data fitting functions are

applied. The obtained 3D surface is then colored with a gradient according to the elevation, and underlying data density[404]. Clusters of molecules with similar properties lead to distinct plateaus, while diverging properties in close proximity are lacking a clear trend in the corresponding landscape patch. This gives rise to landscapes with varying topologies, mainly influenced by the underlying target SAR, the chosen molecular descriptor, and dimensionality reduction methodology. Such response surface landscapes are suitable for large-scale analysis of global or local SARs in the context of a specific molecular representation. Regions with an accumulation of attractive compounds are rapidly identified as well as regions that should be avoided for subsequent investigations. The concept of "activity islands", initially described for abstract chemical space representations by Schneider *et al*.[259], can easily be transferred to 3D landscape visualization. An island consists of molecules belonging to a distinct structural class that share a function in regard to their primary target. Note that activity islands are not necessarily caused by biological distinctions, *e.g.* different mechanisms of action or alternative binding modes, but could be artificially evoked by an inadequate molecular representation[26,405]. A comparison of different molecular descriptors and modeling techniques might be required to find an adequate representation.

A 2D projection of the PPARγ dataset, previously used for SAS analysis in Figure 9, is shown in Figure 10a. MDS was applied to calculate a 2D mapping from the Tanimoto distance matrix. An exemplary 3D landscape derived from the projection is shown in Figure 10b. pK_i values were used as an additional property. Trends in chemical space are immediately visible. Weakly active molecules are distributed in one half of the focused chemical space, while most of the potent compounds are located in two distinct activity islands, which are separated by a sparsely populated area. This region could be a promising target for further chemical exploration efforts.

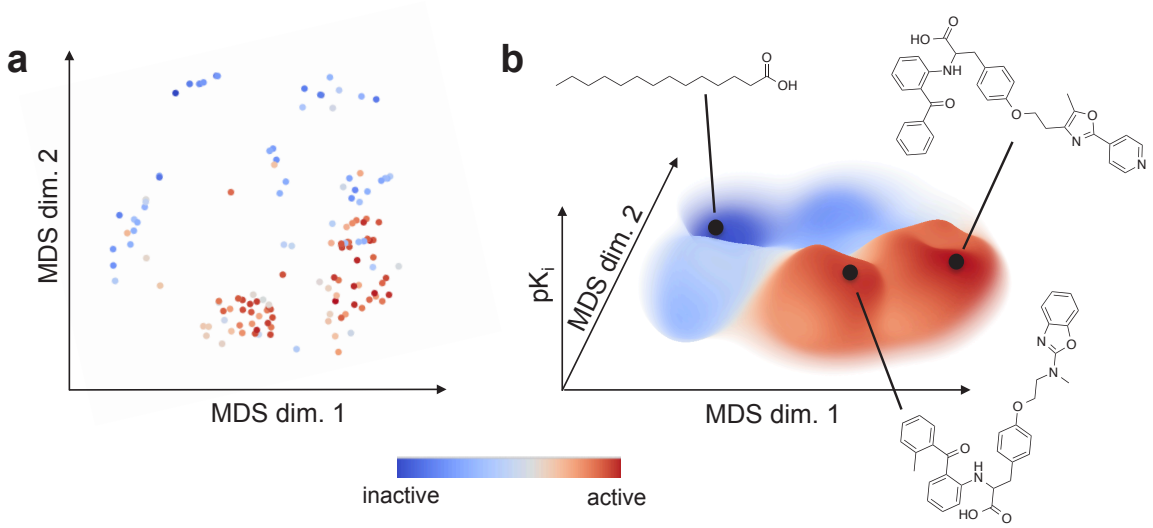


Figure 10. (**a**) 2D chemical space projection found by multidimensional scaling (MDS) for 144 PPARγ inhibitors (*cf*. Figure 9 for details) colored by p*K*i values. (**b**) 3D landscape visualization with highlighted compounds from selected areas of the landscape created with LiSARD. Surface color according to the interpolated activity value.

In principle, the interpolated surface could be used as a nonlinear QSAR model for quantitative prediction, but usually the accuracy is not sufficient[387]. Predictive modeling should instead be performed based on the original high-dimensional space. Hence, 3D landscapes are best used in a qualitative manner to visualize trends of global SAR and provide global views of compound distributions[387].

Graph or network visualization of molecular relationships has a long history in biosciences, for example for the construction and analysis of metabolic, gene regulation, or proteomics networks[406-408]. Network representations of compound sets have a long tradition in chemoinformatics. Minimum spanning trees calculated for two-dimensional representation (MDS or PCA) of physicochemical properties were among the first approaches to visualize molecule relations using graphs[409,410]. Other approaches derived networks directly from the chemical graph *e.g.* by calculating edge-deletion matrixes[411,412], or by relating molecules by their *Maximum Common Substructure* (MCS)[413].

Recently, networks derived from structural similarity metrics have gained increasing attention. In these network visualizations each graph node represents a compound and edges similarity relationships of linked compound pairs. In threshold networks only nodes having a similarity exceeding a specific value are connected, with the consequence of a sparsely connected graph[414].

Network-like Similarity Graphs (NSGs) are an example for such threshold networks[415]. Structural similarity is calculated as the Tanimoto similarity, based on binary fingerprint representations. Edges are added between two nodes if the Tanimoto similarity exceeds a specific threshold. Nodes are colored according to the activity or other properties, *e.g.* a selectivity ratio. An additional information layer is added by sizing the nodes regarding a local discontinuity score. The score is high for closely related compounds with a heterogeneous distribution of activity values. In NSGs, the node layout in the two-dimensional plane is determined by a force-directed layout algorithm solely on the basis of node connectivity. A direct consequence is that actual similarity values are not reflected by edge length[415]. Similarity networks can also be analyzed by means of graph theory. An example is the estimation of the importance of individual compounds in large compound libraries and the identification of key compounds, bridging different structure classes[414].

For the focused QSAR dataset of PPARγ ligands, a NSG is shown in Figure 11a. The activity islands discovered in Figure 10b are also clearly visible. The heterogeneity of the clusters can be visually evaluated, which is assisted by scaling the nodes according to the local discontinuity score. For a diverse dataset taken from ChEMBL database, the NSG consists of a collection of smaller clusters, evenly distributed on the plane by the layout algorithm. The global structure is not easily perceived, and it seems as if NSGs are especially suited for analyzing smaller sets of interrelated compounds, while global chemical space analysis is the domain of 3D landscapes.

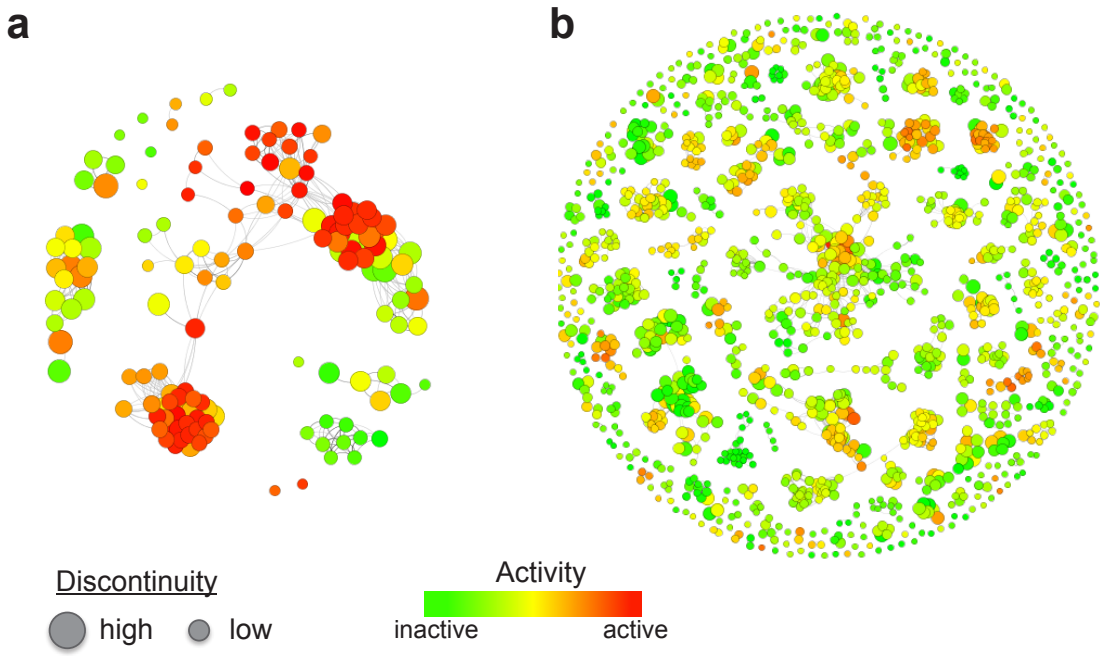


Figure 11. *Network-like Similarity Graph* (NSG) visualization. An edge between two nodes is drawn if the Tanimoto similarity exceeds 0.65. Node size reflects the compound local discontinuity score. Clusters with small circles indicate smooth, and big circles rough SAR with respect to the associated neighbors. Visualization created using SARANEA[416]. (**a**) A focused QSAR set of 144 PPARγ inhibitors (*cf*. Figure 9 for details). (**b**) Diverse set of 1524 compounds collected from ChEMBL with annotated human dopamine D_4 receptor affinity > 10 µM.

By extending the similarity network idea to similarity analysis of whole ligand sets, it is possible to relate drug targets by their known ligands. In ligand set threshold networks nodes represent ligand sets and are linked by edges according to the minimum level of set similarity[417]. Similarity is not restricted to simple set-wise compound similarity comparison, but can also be calculated based on representative ligand set properties, which can be evaluated *e.g.* by correlation analysis of SOM-derived pharmacophoric features[334].

2 Aims of this thesis

This study aims at identifying hit- and lead-compounds that exhibit a multi-target affinity profile, while simultaneously fulfilling secondary criteria including high ligand efficiency, solubility, and acceptable ADMET properties. By first identifying compounds matching the main objective – most commonly high affinity to the main target – and then optimizing the remaining objectives, multiple goals can be consecutively addressed. With contradicting objectives, the sequential approach easily leads to suboptimal compound prioritization, and consequently requires lengthy optimization. Adaptive molecule optimization according to multiple objectives simultaneously could spur a more efficient hit and lead discovery process, with potentially lower compound attrition. Although several retrospective studies have confirmed the benefits of multi-objective optimization, but prospective studies investigating the potential of this emerging computational technology are still rare. Within the scope of this work, several aspects of computer-aided multi-target design are examined in a prospective setting, with a focus on potential applicability to actual drug discovery.

Working hypothesis: Chemical space visualization and predictive machine-learning models can aid the drug design process by suggesting macromolecular targets for a given molecule, and generating innovative, multi-target modulating compounds with the desired properties.

To test this hypothesis, the aims of this doctoral thesis are the following:

1. Development of locally adaptive, multi-objective landscape visualization software as a visual guideline for molecular design.
2. Extension of the topological pharmacophore descriptor (CATS) by aromatic features (CATS2), and evaluation of the parameter influence on virtual screening retrieval performance and scaffold-hopping potential.
3. Development of a multi-target machine-learning model using Gaussian process regression based on publicly available chemogenomics data.
4. Adaption of a nature-inspired optimization concept to combinatorial molecular design, and assessment of the method's applicability to focused library design, large-scale compound sampling, and *de novo* compound generation.
5. Synthesis and testing of *de novo* designed compounds.

3 Results

3.1 Neighborhood-preserving visualization of adaptive structure-activity landscapes and application to drug discovery

In this initial study, a three-dimensional landscape visualization method is introduced as an intuitively accessible roadmap of chemical space. It contains a neighborhood preserving dimensionality reduction technique in order to preserve meaningful relations between molecules in their low-dimensional representation. The extension of the method to incorporate multiple objectives in combined multi-objective landscapes is demonstrated. To assess the benefits of visualization support, the proposed method LiSARD was applied to a representative "real-world" drug discovery project dataset (provided by F. Hoffmann-La Roche Ltd., Basel), focusing on human somatostatin subtype 5 receptor (hSST5R) as therapeutic target.

3.1.1 Abstract

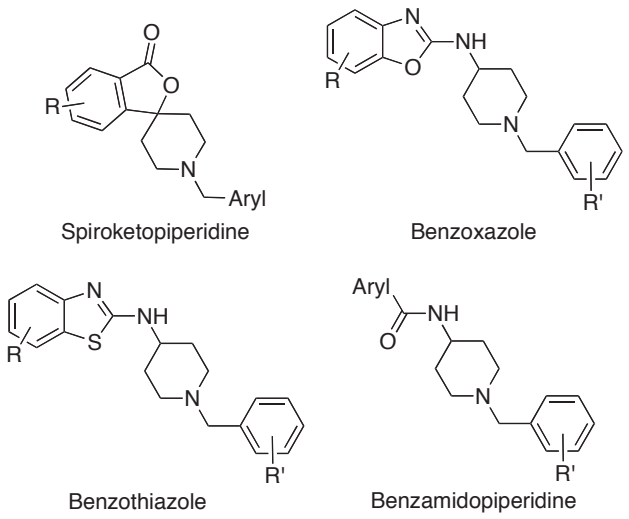
Compound optimization from primary hits to pharmaceutical lead structures by organic synthesis is largely guided by the chemical feasibility and tractability of the candidate compounds, and the specific knowledge and intuition of the medicinal chemists involved. Here, we present a modeling approach that assists synthetic chemists in decision-making and molecular design by visualizing and rationalizing structure-activity and -property relationships as "SAR landscapes".

3.1.2 Introduction

Visualization and analysis of *Structure-Activity Relationship* (SAR) or "fitness" landscapes have been research topics in computational medicinal chemistry for approximately two decades[259,356,374,399,404,410,418-422]. *Principal Component Analysis* (PCA)[423] and *Projection to Latent Structures* (PLS)[424] yield linear, statistically interpretable SAR models and data projections from typically high-dimensional property spaces. Due to the underlying mathematical models, the solutions provided by *nonlinear* projection are often more accurate, but also evade immediate

interpretation. Despite this apparent drawback, nonlinear projection techniques like the *Self-Organizing Map* (Kohonen network, SOM)[114,370,376], Sammon mapping[380,425,426], *Multidimensional Scaling* (MDS)[427], and *Stochastic Proximity Embedding* (SPE)[428,429] – to name just the most prominent approaches – have demonstrated their particular usefulness for SAR modeling. Their appeal lies in the ability to appropriately mirror the typically nonlinear dependencies between a structural (constitution-, topology-, conformation-based) molecular representation and some measured bioactivity or property.

We present an advanced modeling approach to SAR landscape visualization that results in easily interpretable biological response surfaces in chemical space (*Ligand-induced Structure-Activity Relationship Display,* LiSARD). The LiSARD algorithm generates interactive graphics that can be used as intuitive roadmaps for molecular design and optimization. As a first practical application, we analyzed human somatostatin receptor subtype 5 receptor (hSST5R) antagonists. This class-A G-protein coupled receptor is involved in several physiological processes, *e.g.*, NMDA receptor activation and control of hormonal secretion[430,431]. In a chemogenomics study aimed at finding non-peptidic hSST5R antagonists, approx. 3000 compounds of which the majority belonged to four structural classes, were synthesized and tested at Roche (Scheme 1)[432-435].



Scheme 1. Scaffold classes of compounds synthesized and tested for hSST5R activity (or antagonization).

Although this application of LiSARD to a real-life dataset from Roche is of a retrospective nature, it serves as a proof-of-concept study to test the applicability of innovative approaches for interactive SAR visualization in medicinal chemistry.

3.1.3 Material and Methods

Structural data

Compound structures were standardized using the "wash" function in MOE v2010.10 (The Chemical Computing Group Inc., Montreal, Canada). Properties were computed with MOE. CATS descriptors were computed using the *speedcats* software (0-9 bonds, type-sensitive scaling), as described[14,436].

Dimensionality reduction

Stochastic Neighbor Embedding (SNE) defines two conditional probabilities: i) p_{ij}, the probability that a data point $\boldsymbol{\xi}_i$ has $\boldsymbol{\xi}_j$ as its neighbor (Eq. 12), and ii) $q_{j|i}$, the induced probability that point i picks point j as its neighbor as a function of the low-dimensional images $\mathbf{x}_i$ of all data points $\boldsymbol{\xi}_i$ (Eq. 13). The cost function minimized in the embedding is a sum of the *Kullback-Leibler* (KL) divergences between the original (p_{ij}) and induced (q_{ij}) distributions (Eq. 14).

$$p_{ij} = \left. \exp\left(-\frac{\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2}{2\sigma_i^2}\right) \middle/ \sum_{\mathrm{k}\neq\mathrm{i}} \exp\left(-\frac{\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_k\|^2}{2\sigma_i^2}\right), \text{ where} \right. \tag{12}$$

σ_i was chosen by a binary search, such that the entropy of the distribution over neighbors was equal to log$k$, where k is the number of local neighbors or "perplexity".

$$q_{ij} = \left. \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \middle/ \sum_{\mathrm{k}\neq\mathrm{i}} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2) \cdot \right. \tag{13}$$

$$C = \sum_{\mathrm{i}} \sum_{\mathrm{j}} p_{ij} log \frac{p_{ij}}{q_{ij}} = \sum_{\mathrm{i}} KL(P_i||Q_i). \tag{14}$$

We used the implementation of SNE from the Matlab Toolbox for Dimensionality Reduction v0.7.2[437] and Matlab 7.10.0 (The MathWorks Inc., Natick, USA).

Surface calculation

The Nadaraya-Watson estimator was applied to fitting a surface to the projected data points[438,439]. The value for an unobserved location is estimated as a locally weighted average of the given data, using a kernel as weighting function. For a set of n observations $(\mathbf{x}_i, y_i)$ with $\mathbf{x} \in \mathbb{R}^2$ and $y \in \mathbb{R}$ the Nadaraya-Watson estimator is defined as given in Eq. 15.

$$\hat{\mathrm{m}}(\mathbf{x}, \mathbf{h}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \varkappa_{\mathbf{h}}(\mathbf{x} - \mathbf{x}_i) y_i \Big/ \sum_{\mathrm{i}=1}^{\mathrm{n}} \varkappa_{\mathbf{h}}(\mathbf{x} - \mathbf{x}_i) \text{, where} \qquad (15)$$
$$\varkappa_{\mathbf{h}}(\mathbf{x}) = \frac{1}{h_1 h_2} \varkappa\left(\frac{x_1}{h_1}, \frac{x_2}{h_2}\right).$$

Here, $\mathbf{h} = (h_1, h_2)$ is a vector of bandwidths, and $\varkappa(\boldsymbol{x})$ a multivariate kernel function, for which we used the multivariate Gaussian kernel (Eq. 16).

$$\varkappa_{\mathbf{H}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{H}|^{-\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{x}^T \mathbf{H}^{-1} \mathbf{x}}. \qquad (16)$$

The bandwidth matrix $\mathbf{H}$ is defined as $\mathbf{H} = diag(h_1^2, h_2^2)$[440]. The optimal bandwidth h_1, h_2 was estimated from the data according to the *Normal Reference Rule*[441]. As data points are in general unevenly distributed, a fixed bandwidth only represents a compromise for both densely and sparsely populated areas. We thus combined the *Local Density Adaptive Bandwidth Estimator* with the Nadaraya-Watson approach to obtain the local bandwidth $\mathrm{h}(\mathbf{x})$ (Eq. 17)[442,443].

$$\mathrm{h}(\mathbf{x}) = k\left(\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \varkappa_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)}\right)^{-1}. \qquad (17)$$

The factor k defines the degree of smoothing. We used k = 1. The bandwidth varies with the estimation position, inversely proportional to local data density. All surface interpolations were calculated in our visualization software LiSARD, which was implemented in Java SE 6 (Oracle Corporation, Redwood Shores, USA). For license requests contact the authors.

Chemical dissimilarity assessment

The dissimilarity between compounds was calculated as the distance between descriptor values of the compounds. In this study, we used the Euclidean distance between the compounds numerical descriptor representation. The Euclidean distance d_{ij} between descriptor vectors p_i, p_j is calculated with Eq. 18, where n is the dimensionality of the descriptor.

$$d_{ij} = d(p_i, p_j) = \sqrt{\sum_{k=1}^{n} (p_{ik} - p_{jk})^2}. \quad (18)$$

Comparison of dimensionality reduction methods

Three chemical compound libraries containing drug-like bioactive molecules (LOPAC Library of Pharmacologically-Active Compounds (Sigma-Aldrich, St. Louis, USA), 1280 compounds; COBRA Collection of Bioactive Reference Compounds v10.3[37], 11230 compounds; hSST5R project data, 2965 compounds) served as reference data for the comparison of four different dimensionality reduction techniques (MDS[427], SPE[428,429], SNE[382], PCA[423]). We analyzed the preservation of the K = 10 nearest neighbors of each compound in the original high-dimensional space (150-dimensional topological CATS pharmacophore descriptor, scaled by background feature occurrence)[14,436] and in the three-dimensional (3D) projection. Five quality indices were calculated: *Trustworthiness* W_T, *Continuity* W_C, *Mean Relative Rank Error* (MRRE) W_n, MRRE W_v, and the *Local Continuity Meta-Criterion* (LCMC), as detailed below.

For high-dimensional vectors $\boldsymbol{\xi}$ with distance δ_{ij} and their lower-dimensional projection $\mathbf{x}$ with distance d_{ij}, ρ_{ij} is the rank of $\boldsymbol{\xi}_j$ with respect to $\boldsymbol{\xi}_i$ and r_{ij} for $\mathbf{x}_j$ and $\mathbf{x}_i$. The co-ranking matrix is defined as $\mathbf{Q} = [q_{kl}]_{1\le k,l\le N-1}$ with $q_{kl} = \left|\{(i,j): \rho_{ij} = k \text{ and } r_{ij} = l\}\right|$. On the co-ranking matrix the following three subdivisions can be defined (Eq. 19):

$$\begin{aligned} \mathbb{UL}_K &= \{(i,j): 1 \le i \le K \text{ and } 1 \le j \le K\}, \\ \mathbb{LL}_K &= \{(i,j): K < i \le N-1 \text{ and } 1 < j \le K\}, \\ \mathbb{UR}_K &= \{(i,j): 1 \le i \le K \text{ and } K < j \le N-1\}. \end{aligned} \quad (19)$$

Accordingly, the T&C indices[444] (Eq. 20) were computed as:

$$\begin{aligned} \mathrm{W}_T(K) &= 1 - \frac{2}{G_K}\sum_{(\mathrm{k,l})\in \mathbb{LL}_K}(k-K)q_{\mathrm{kl}}, \text{ and} \\ \mathrm{W}_C(K) &= 1 - \frac{2}{G_K}\sum_{(\mathrm{k,l})\in \mathbb{UR}_K}(l-K)q_{\mathrm{kl}}. \end{aligned} \tag{20}$$

The normalization factor $G_K = N\,min\{K(2N-3K-1),(N-K)(N-K-1)\}$ considers the worst and scales the values to [0,1]. N is the data count.

Additionally, we computed MRRE indices[444], which incorporate the rank difference in the first K ranks (Eq. 21):

$$\begin{aligned} \mathrm{W}_n(K) &= \frac{1}{H_K}\sum_{(\mathrm{k,l})\in \mathbb{UL}_\mathrm{K}\cup\mathbb{LL}_\mathrm{K}}\frac{|k-l|}{l}q_{\mathrm{kl}} \\ \mathrm{W}_v(K) &= \frac{1}{H_K}\sum_{(\mathrm{k,l})\in \mathbb{UL}_K\cup\mathbb{UR}_K}\frac{|k-l|}{k}q_{\mathrm{kl}}. \end{aligned} \tag{21}$$

The normalizing factor $H_K = N\sum_{k=1}^{K}\frac{|N-2k+1|}{k}$ considers the worst-case scenario.

LCMC[445] is a measure for the preservation of the first K ranks (Eq. 22):

$$\mathrm{U}_{LC}(K) = \frac{K}{1-N} + \frac{1}{NK}\sum_{(\mathrm{k,l})\in \mathbb{UL}_K} q_{\mathrm{kl}}. \tag{22}$$

It is desirable to achieve high $\mathrm{W_T}$, $\mathrm{W_C}$ and LCMC values, and low $\mathrm{W_n}$ and $\mathrm{W_v}$ values.

3.1.4 Results and Discussion

Projection and visualization of chemical data

For SAR landscape analysis, compounds need to be represented by meaningful structural attributes ("descriptors") that correlate with the measured activities. In this study, we employed a topological pharmacophore representation (CATS descriptor)[14,436], which leads to a 150-dimensional feature space containing information of both molecular structure and potential ligand-receptor interaction points. Molecules that are neighbors in such a chemical space are more likely to have similar properties and activity than compounds with a large pairwise distance. This means that for dimensionality reduction and data visualization the preservation of the local neighborhood (context) might be more important than preservation of distances. We evaluated different dimensionality reduction methods (PCA[423], MDS[427], SPE[428,429], SNE[382]) for their ability to preserve the high-dimensional

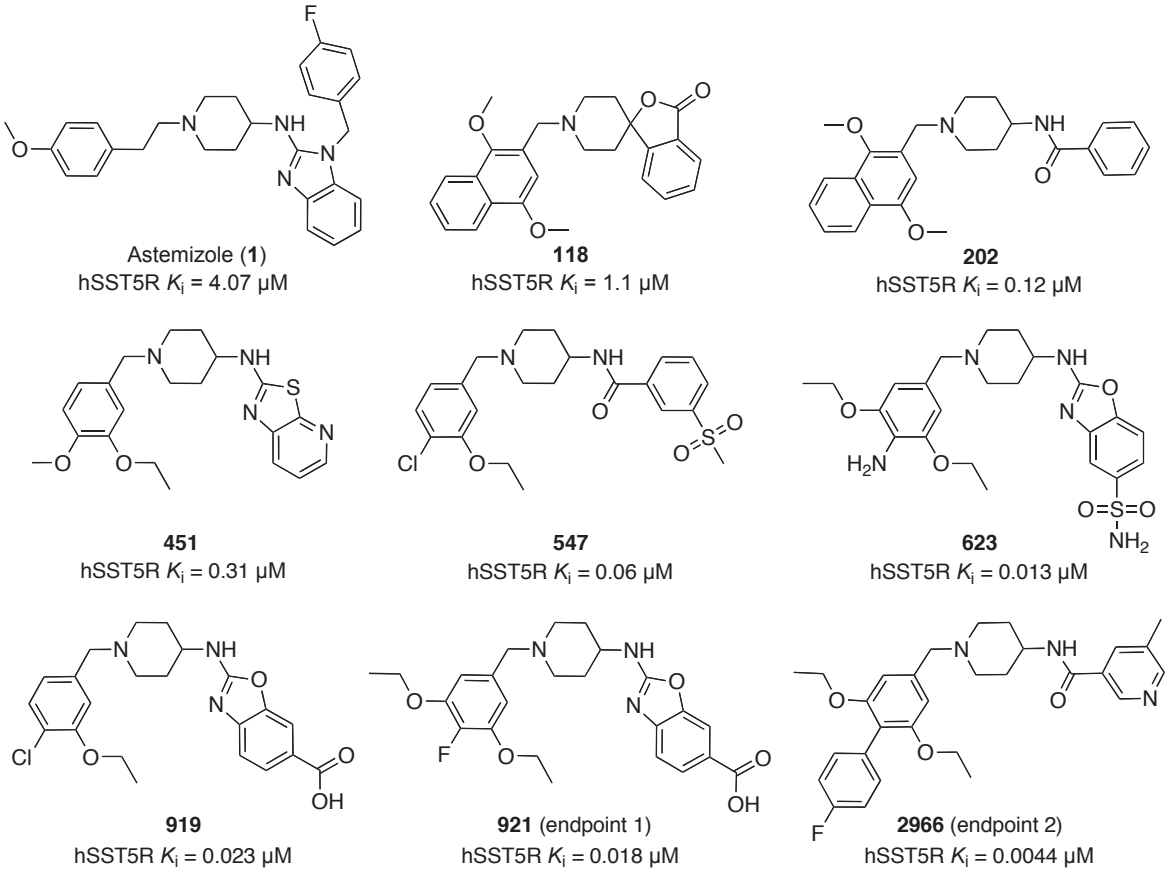
neighborhood in the low-dimensional projection for different chemical libraries (Table 3). Within the given framework, SNE scored best for all calculated measures on all datasets. SNE aims at finding a projection from the original data space to the lower dimensional embedding space such that the pairwise neighborhood distribution of points in the data and embedding space are approximately the same. For visualization of the hSST5R compound distribution, the 150-dimensional descriptor space was projected to a three-dimensional (3D) space by SNE. We observed a projection trustworthiness of 96% and continuity of 99% for the hSST5R data.

Table 3. Comparison of four different dimensionality reduction methods (MDS, SPE, SNE, PCA) for three chemical libraries (LOPAC, COBRA, hSST5R). All indices were calculated for K = 10 nearest neighbors.

	LOPAC				COBRA				hSST5R			
	MDS	SPE	SNE	PCA	MDS	SPE	SNE	PCA	MDS	SPE	SNE	PCA
Trustworthiness W_T	0.89	0.85	0.92	0.85	0.89	0.84	0.93	0.89	0.80	0.81	0.96	0.90
Continuity W_C	0.97	0.92	0.98	0.95	0.97	0.92	0.99	0.97	0.95	0.89	0.99	0.99
MRRE W_n	0.11	0.14	0.07	0.14	0.11	0.16	0.06	0.10	0.20	0.19	0.03	0.10
MRRE W_v	0.03	0.07	0.02	0.04	0.02	0.07	0.01	0.02	0.05	0.10	0.01	0.01
LCMC	0.16	0.20	0.44	0.21	0.11	0.04	0.33	0.13	0.06	0.06	0.60	0.19

For ease of interpretation, 2D projections are preferably used for analyzing the structure of the SAR landscape, with the third dimension typically being an experimental measure of activity (*e.g.*, pIC_{50}). To allow for multiple views on such a landscape and reduce the risk of artifacts and potential misinterpretation, LiSARD enables manual rotation of the 3D compound cloud to select a suitable view on a 2D plane (Figure 12a). Local activity values are computed on the fly from the available data points by Gaussian kernel regression with adaptive bandwidths and represented as a colored surface (Figure 12b). As a result, a continuous surface is spanned over the data points. Its height corresponds to some observed variable, *e.g.* local compound potency, and the transparency of the landscape indicates the local confidence of the model. This continuous SAR landscape representation resulted from tight interaction between the computer scientists and medicinal chemists. It offers the advantage of identifying local SARs by rotating the compound cloud and reducing the risk of over-interpretation of a static graphical model.

In this study, experimentally determined biological activity values were added to the projected data points as the third dimension, which required interpolation between the data points. For continuous SAR estimation Bajorath and coworkers recently employed a geostatistic method called *kriging*[404]. As this technique turns out to be computationally demanding and a limiting factor for dynamic data visualization, we decided to implement a Kernel regression technique, which is suitable for large datasets and interactive calculation of fitness landscapes.



Scheme 2. Selected compounds and their potency on hSST5R. Higher compound index numbers indicate later stages of the project.

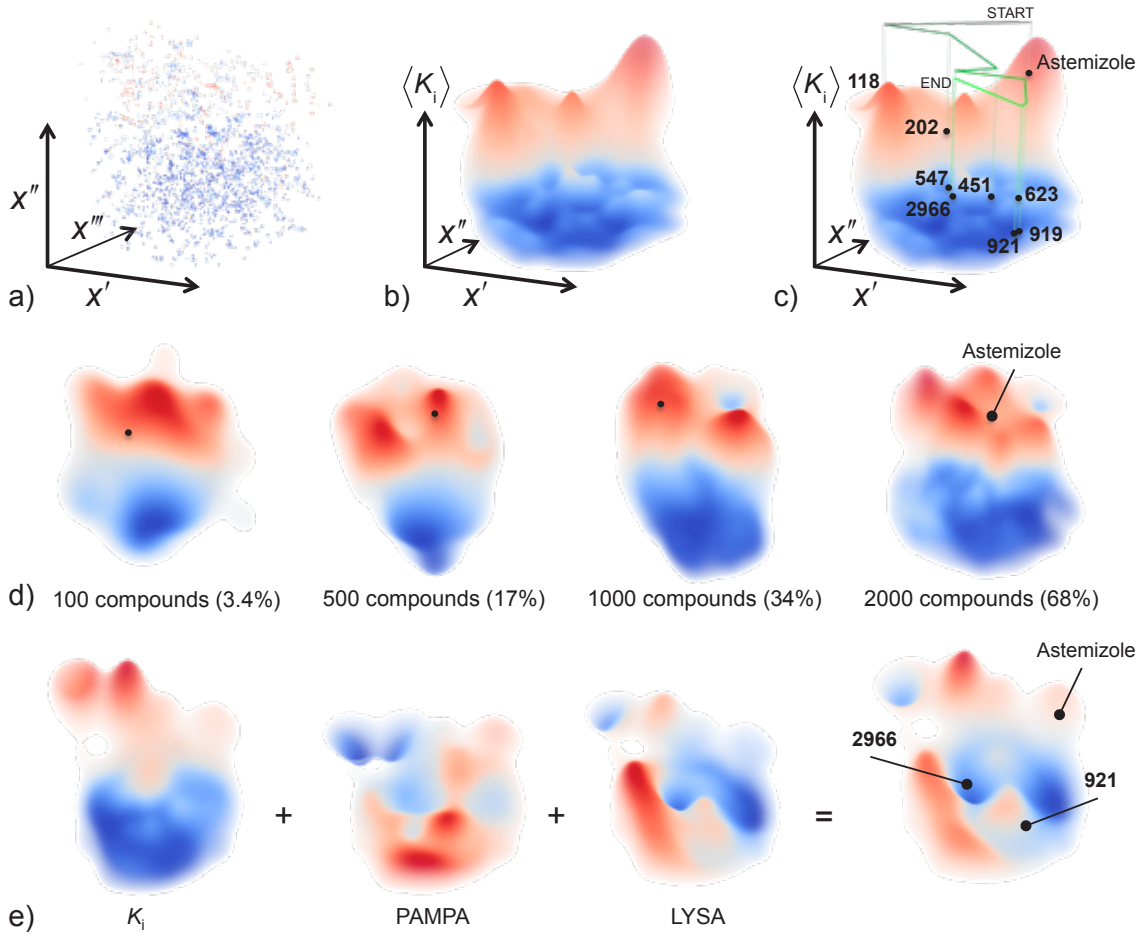


Figure 12. Landscape generation and visualization by LiSARD. a) Three-dimensional compound distribution obtained by compressing a 150-dimensional compound representation (CATS pharmacophore feature vector) using *Stochastic Neighbor Embedding* (SNE)[382]. Dots represent compounds; *blue*: low k_i; *red*: high k_i. b) View on the hSST5R SAR landscape. Note that only two (x', x'') of the three dimensions shown in a) are used. Surface regions with low transparency correspond to areas of high model confidence. Coloring of the landscape is according to average local potency $\langle k_i \rangle$. c) Progress of the hSST5R project over time. The sample trajectory starts with the reference Astemizole and ends at compound **2966**. Compound numbers correspond to strategically preferred hits selected by the project team. d) Adaptive evolution of the structure-activity landscape for hSST5R agonists over project time. The snapshots contain increasing levels of detail that can be captured depending on the available number of compounds synthesized and tested. Note that active and inactive compounds contribute equally to the model (*blue* = low k_i, *red* = high k_i). e) Superimposition of different landscapes for various relevant drug properties results in a multidimensional "fitness landscape" for drug design (here: hSST5R). The locations of the reference Astemizole and the two endpoints of medicinal chemical optimization (**921, 2966**) are shown. (*blue* = preferred regions, *red* = tabu regions).

Trajectories in chemical space, adaptive SAR landscapes, and multidimensional optimization

Visualization of the chemical space visited during a drug design project can help avoid *tabu*-areas containing unwanted chemotypes and properties. Figure 12c presents the trajectory of hSST5R project progress over time. Strategic decision points are marked, and the corresponding lead structures are presented in Scheme 2. Using a chemogenomics strategy, Astemizole **1** and the spiropiperidine class were identified as chemical entry points. Until then no small-molecule hSST5R antagonists were known. Both starting points evolved into the benzothiazole and the bioisosteric benzamidopiperidine series, which were optimized in parallel.

In the early project phase, compound potency and selectivity towards the related H1 receptor (Astemizole was marketed as a H1R antagonist and withdrawn later) were the driving criteria to obtain potent and selective tool compounds. Compound **921** (endpoint 1) represents such an intermediate candidate. This phase was followed by a multidimensional optimization strategy aiming at the best compromise between potency and physicochemical as well as pharmacokinetic parameters. Optimization finally resulted in compound **2966** (endpoint 2), which was tested *in vivo*[446].

Visualization of SAR landscapes at all project stages provides an additional criterion that is based not only on the actives found so far, but equally on the inactive compounds. Figure 12d demonstrates the adaptive nature of the SAR landscape models. Depending on the number of compounds synthesized and tested, and on the project status, increasingly fine-grained models are computed. Using two thirds of the data, the final shape of the SAR landscape (*cf*. Figure 12b) is clearly visible. It is of note that even the first approximate landscape model computed from only 100 compounds correctly structures chemical space into desired (blue) and "tabu" (red) regions (Figure 12d). Having access to such knowledge at an early project stage provides valuable information for hit prioritization, and helps focus on relevant areas in chemical space earlier so that optimized lead structures may be identified faster.

Monitoring the SAR landscape over project duration certainly is a desirable feature for medicinal chemists to explore innovative structural variations of a chemotype and avoid walking in circles and areas with potential off-target liabilities. Multiple activities and properties can be displayed simultaneously in LiSARD, thereby enabling multidimensional optimization with the aim to avoid compounds that have

an undesired pharmacological activity and property profile. Figure 12e presents the superimposition of the landscapes for experimentally determined potency (hSST5R antagonism), membrane permeability (PAMPA) and aqueous solubility (LYSA)[447-449]. Such a multi-dimensional "fitness landscape" can be readily obtained by adding up the individual landscape functions and subsequent re-scaling of the *z*-axis to obtain *pseudo*-probabilities.

3.1.5 Conclusion

In this proof-of-concept study, we have demonstrated that a dynamic view on adaptive SAR landscapes can support molecular design by providing project-specific visual aids for compound prioritization. Potential compound liabilities can be avoided, multiple properties can be considered at a time, and the information contained in both active and inactive compounds is optimally exploited for early hit prioritization and progress monitoring.

3.1.6 Publication details and contributions

Authors

Michael Reutlinger,[a] Wolfgang Guba,[b] Rainer E. Martin,[b] Alexander I. Alanine,[b] Torsten Hoffmann,[b] Alexander Klenner,[a] Jan A. Hiss,[a] Petra Schneider,[a] Gisbert Schneider [a]

[a] ETH Zurich, Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

[b] F. Hoffmann-La Roche Ltd., Discovery Chemistry, 4070 Basel, Switzerland.

Author contributions

M.R. designed and developed the software LiSARD, conducted the research, analyzed the results, prepared the figures, and contributed to the manuscript.

Reference

Reutlinger M, Guba W, Martin RE, Alanine AI, Hoffmann T, Klenner A, Hiss JA, Schneider P, Schneider G (2011) Neighborhood-preserving visualization of adaptive structure-activity landscapes: Application to drug discovery. *Angew. Chem. Int. Ed.* **50**, 11633-11636.

Licence

Source of funding

The research was supported in parts by the Deutsche Forschungsgemeinschaft (DFG, FOR1406, TP4), the Swiss National Science Foundation (grant no. 205321-134783), and the OPO-Foundation Zurich.

3.2 Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for "orphan" molecules

The chemical space visualization LiSARD introduced in Chapter 3.1 utilizes a molecular representation. The topological CATS pharmacophore descriptor, which was used in the study, does not explicitly consider aromatic pharmacophores in its original implementation[42]. In this second study, the CATS descriptor was extended by the aromatic pharmacophore pattern and evaluated for its ability to enrich bioactive compounds with an emphasis on the scaffold hopping potential. Additionally, the viability of CATS for identifying new biological targets for a given chemical structure was assessed.

3.2.1 Introduction

Drug discovery is driven by the identification of *New Chemical Entities* (NCEs)[450,451]. Virtual screening and *de novo* design techniques have been proven to serve this purpose, thereby complementing experimental biochemical and biological approaches[11]. Still, it remains a matter of debate, which particular molecular representation and similarity index are preferable for a given drug target in order to identify appropriate NCEs with minimal synthetic and testing effort involved[61,452-455]. Ligand-based chemical similarity approaches have also been effectively applied to large-scale activity and target prediction for known drugs, some of the prominent methods being PASS developed by Poroikov *et al*.[345], the techniques conceived by Mestres and coworkers[341,456,457], and the *Similarity Ensemble Approach* (SEA) implemented by the Shoichet group[333]. Here, we compared several popular two-dimensional molecular representations for their ability to retrieve actives (enrichment potential) and chemotypes (scaffold-hopping potential) from a collection of druglike bioactive compounds. Subsequently the applied *Chemical Advanced Template Search* (CATS)[14,42] was applied to predicting potential drug targets for a virtually assembled combinatorial compound library, from which we synthesized and successfully tested candidate compounds. The results demonstrate that CATS is not only suited for its intended purpose of NCE retrieval by scaffold-hopping[62], but also

for reliable target profiling of "orphan" virtual molecules[342,458]. It thereby complements the suite of available validated tools for target prediction.

3.2.2 Material and Methods

Synthesis and analytics

Chemical synthesis was performed with a Biotage® Initiator microwave synthesizer (Upsala, Sweden). Aminopyridine (1.0 mol. eq.), aldehyde (1.0 mol. eq.), isocyanyde (1.0 mol. eq.) and perchloric acid (11 mol%) were dissolved in EtOH (1.1 ml × $mmol^{-1}$). The solution was heated at 170ºC for 5 minutes under microwave irradiation . The resulting crude product was purified *via* preparative HPLC using CH_3CN:H_2O (+0.1% trifluoroacetic acid in each phase) as eluent, in a gradient of 5-50% CH_3CN run over 16 minutes, to afford compounds **2** and **3** as yellow oils.

Compound **2** (methyl 2-((2-(2,4-dimethoxyphenyl)imidazo[1,2-*a*]pyridin-3-yl)amino)acetate), 81%: ^{1}H-NMR (CD_3OD, 400.13 MHz): δ 3.44 (3H, s, OCH_3), 3.67 (2H, s, CH_2), 3.75 (3H, s, OCH_3), 3.80 (3H, s, OCH_3), 6.57-6.61 (2H, m, Ar-*H*), 7.32-7.36 (1H, m, Ar-*H*), 7.53 (1H, d, *J* = 8.0 Hz, Ar-*H*), 6.77-7.75 (2H, m, Ar-*H*), 8.68 (1H, d, *J* = 2.4 Hz, Ar-*H*). ^{13}C NMR (CD3OD, 100.61 MHz): δ 48.54, 52.52, 56.18, 56.48, 99.77, 107.08, 108.44, 112.51, 117.43, 123.13, 126.42, 129.04, 132.62, 133.56, 137.46, 159.91, 164.60, 173.19. HRMS-ESI calc. ($C_{18}H_{19}N_3O_4$+H^+): 342.1448, found: 342.1448.

Compound **3** (methyl 2-(1-methyl-1*H*-pyrrol-2-yl)-3-((2-morpholinoethyl)amino)imidazo[1,2-*a*] pyridine-7-carboxylate), 74%: ^{1}H-NMR (CD_3OD, 400.13 MHz): δ 3.08 (2H, m, C*H*$_2$), 3.21 (2H, t, *J* = 6.4 Hz, C*H*$_2$), 3.37 (2H, m, C*H*$_2$), 3.45 (2H, t, *J* = 6.4 Hz, C*H*$_2$), 3.72 (3H, s, C*H*$_3$), 3.81 (2H, m, C*H*$_2$), 3.97 (2H, m, C*H*$_2$), 4.05 (3H, s, C*H*$_3$), 6.32 (1H, dd, *J* = 3.8 Hz, Ar-*H*), 6.63 (1H, dd, *J* = 3.8 Hz Ar-*H*), 7.06 (1H, m, Ar-*H*), 7.93 (1H, dd, *J* = 1.6 and 7.2 Hz, Ar-*H*), 8.40 (1H, m, Ar-*H*), 8.80 (1H, dd, *J* = 0.8 and 7.2 Hz, Ar-*H*). ^{13}C NMR (CD_3OD, 100.61 MHz): 35.01, 41.04, 53.33, 53.84, 57.27, 64.81, 110.04, 114.52, 115.63, 116.03, 116.31, 118.78, 126.40, 127.62, 131.28, 134.24, 137.07, 165.19. HRMS-ESI calc. ($C_{20}H_{25}N_5O_3$+H^+): 384.2030, found: 384.2031.

We used dynamic light scattering (Brookhaven 90Plus) to determine potential aggregation of compound **3** in aqueous solution with 1% DMSO. Aggregate particles were observable at concentrations ranging from 15.5-250 µM.

Self-organizing map

We use our software tool *molmap* for generating a toroidal SOM containing 160 clusters arranged in a 16×10 rectangular grid, as described previously[334-336], with number of training cycles = 10^6 and Gaussian neighborhood radius = 8.

CATS molecular descriptor

Descriptor calculation was performed with a proprietary Java-based software tool (for licensing options, contact G.S.). Free online access to demonstration software is provided at URL: http://modlab-cadd.ethz.ch/

Biochemical activity determination

Activity against PI3Kα was measured by Reaction Biology Corp. (Malvern, PA, USA) in a 10-dose IC_{50} determination (n = 3), in the presence of 10 μM ATP. Preliminary DNA topoisomerase and gyrase inhibition tests were performed with a compound concentration of 5 mM by Inspiralis Ltd (Norwich, UK).

3.2.3 Results and Discussion

A framework for retrospective evaluation of similarity searching runs with different molecular representations ("descriptors") was established on basis of the COBRA collection of druglike bioactive compounds[37], employing Euclidean distances for metric descriptors and the Tanimoto coefficient for fingerprint descriptors[10]. COBRA contains 12,642 manually curated entries with 980 target protein subtype annotations. For 170 macromolecular drug targets with a minimum of 20 annotated active ligands per target, each compound annotated as "active" was selected as a query in turn, and compared to all remaining compounds in the screening pool in terms of molecular descriptor similarity, finally yielding sorted results lists with the most similar or least distant pool compounds sorted to the top. Although there are large collections of bioactive compounds available in the public domain[459], we used the carefully compiled COBRA collection to (i) reduce the risk of erroneous activity

data and faulty compound structures[460], and (ii) avoid redundancy with existing tools that are based on such public structure-activity data. In addition, we intend to probe the value of a comparably small but well curated reference compound pool for target prediction.

We used a representative set of descriptors and fingerprints for benchmarking. "Morgan" fingerprints, closely related to *Extended-Connectivity Fingerprints* (ECFP), are based on radial assessment of non-predefined potentially infinite molecular fragments[36]. The "AtomPair" descriptor can be seen as a CATS predecessor merely denoting the occurrence of all pairs of atoms at a given topological distance[59]. The "MACCS" keys represent substructure-based fingerprints,[461] and the "RDkit" fingerprint implements a Daylight-like fingerprint based on hashed molecular subgraphs[462]. Latter fingerprints and descriptors were calculated using the open-source software package RDkit[463]. Finally, the "MOE2D" descriptor consists of a standardized vector of physicochemical properties provided by the Molecular Operating Environment (v2011, Chemical Computing Group, Montreal).

At this point, we analyzed two versions of CATS vectors, namely the originally described CATS1[14,42] and CATS2, which distinguishes lipophilic from aromatic atoms during typing, thereby resulting in more pharmacophore type pairs and consequently a higher dimensionality of the descriptor than CATS1, which lacks the aromatic atom type. For both descriptors we employed "types scaling", which mitigates the potential dominance of prevalent pharmacophore feature types, and a maximal correlation distance of 10 bonds[436]. An example of CATS descriptor calculation is presented in Figure 13.

We employed the *Receiver Operating Characteristic* (ROC) related BEDROC score for actives-retrieval benchmarking[53]. For our study, the alpha level of the BEDROC method was set to 160.9, which corresponds to the top 1% of the screening list contributing 80% of the score. Murcko scaffold[464] diversity among the set of actives within the top 1% of respective screening lists served as measure for scaffold-hopping potential.

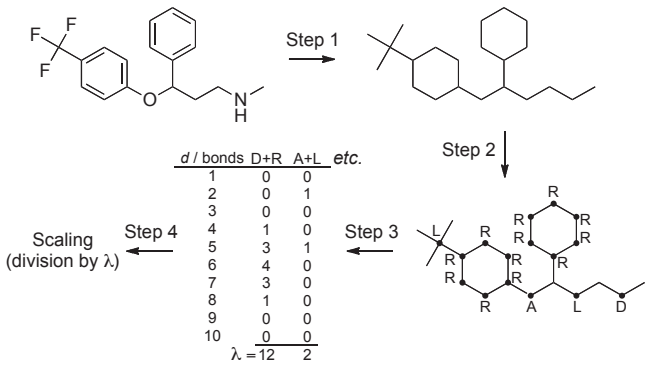


Figure 13. Principle of CATS descriptor calculation. The molecular structure (Step 1) is reduced to the molecular graph, and feature types are assigned (Step 2; L, lipophilic; R, aromatic; A, hydrogen-bond acceptor; D, hydrogen-bond donor). Then, atom pairs for all feature pairs are counted (Step 3), and the final descriptor values are scaled (Step 4). Here, the raw values were divided by the respective λ value (sum of atom type pair occurrences). Note that not all vertices in the molecular graph are considered "pharmacophoric". These possess no feature types.

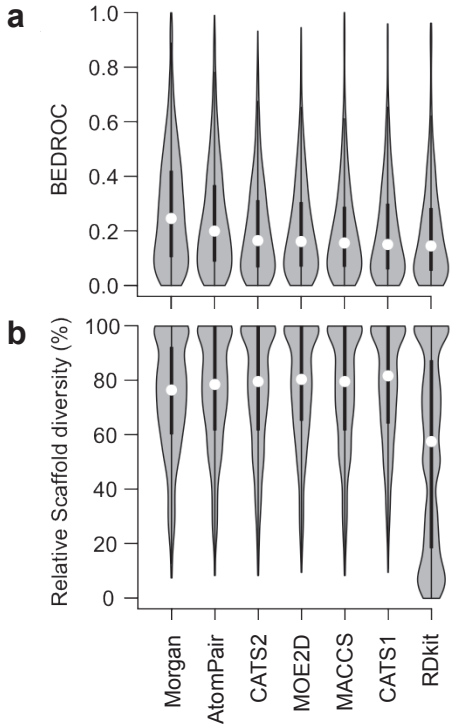


Figure 14. Comparison of molecular representations for their abilities to retrieve known actives (**a**) and scaffolds (**b**) from a collection of druglike bioactive compounds (COBRA). Violin plots show the shapes (gray), medians (white circle) and quartiles (thick lines) of the distributions.

Albeit state-of-the-art radial fingerprints and atom-pair fingerprints outperformed CATS descriptors in terms of the number of actives retrieved (Figure 14a), they ratify their intent of design by delivering the overall highest ratio of diverse scaffolds among retrieved actives. Scaffold-hopping potential was determined by examining the distribution of relative scaffold diversities *r*, which is the ratio of differing scaffolds *s* to the number of retrieved actives *n* among the top 1% of respective

screening runs. While *s* correlates to the BEDROC scores when comparing different descriptors, *r* unveils the CATS1 descriptor as the most suitable descriptor for scaffold-hopping among the compared molecular representations (Figure 14b). In terms of BEDROC scores estimating the enrichment potential, radial fingerprints (Morgan) and Carhart-type atom pairs (AtomPair) performed similar, as did the CATS2 and MOE descriptors, while MACCS, CATS1, and RDkit fingerprints formed a third group (Figure 15a). With respect to scaffold-hopping potential, the groups vary, with CATS1 and MOE2D pairing up, as well as CATS2 and MACCS (Figure 15b). It might thus be advisable to select one method from each group for similarity searching and compare ranked results lists, *e.g.* by data fusion[465]. We wish to point out that the grouping of methods depicted in Figure 15 should be treated with caution, as the dendrograms are likely to vary for other reference data sets and chemotype/target coverage.

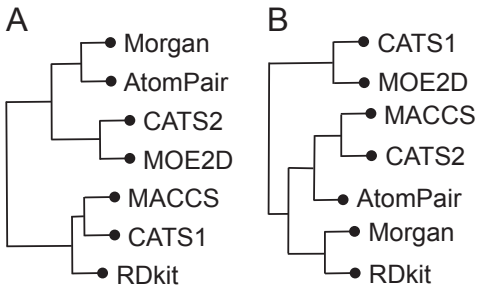


Figure 15. Similarity of molecular representations in terms of their enrichment (A) and scaffold-hopping potential (B). Pair-wise, one-sided Wilcoxon rank sum tests[466] were performed for the BEDROC score distributions of the descriptors. Clustering the obtained *p*-values with Ward's method[467] resulted in the depicted dendrograms.

The outcome of this limited benchmark study is in agreement with a large-scale systematic analysis of 2D fingerprint methods by Sherman and coworkers, who conclude "(..) *if the objective of a screen is to identify novel, diverse hits, then a less specific atom-typing scheme may be more appropriate"*[50]. The CATS representation of molecular graphs and pharmacophoric features serves this purpose of finding new chemotypes. When using the descriptor, one should not expect highest possible enrichment of actives among the top-scoring virtual hits, but can anticipate surprising new ideas for synthesis and activity testing.

This intended permissiveness ("fuzziness")[468,469] of the CATS molecular representation, which is achieved by coarse-grained atom-typing and feature pair

correlation, not only enables scaffold-hopping but may also be used for predicting mutual targets of structurally diverse bioactive ligands. Here, we started from an Ugi-type three-component combinatorial synthesis (Scheme 3)[470] and tested whether we could use CATS for "de-orphanizing" some of the compounds by target identification. All prospective experiments were carried out with the CATS2 implementation.

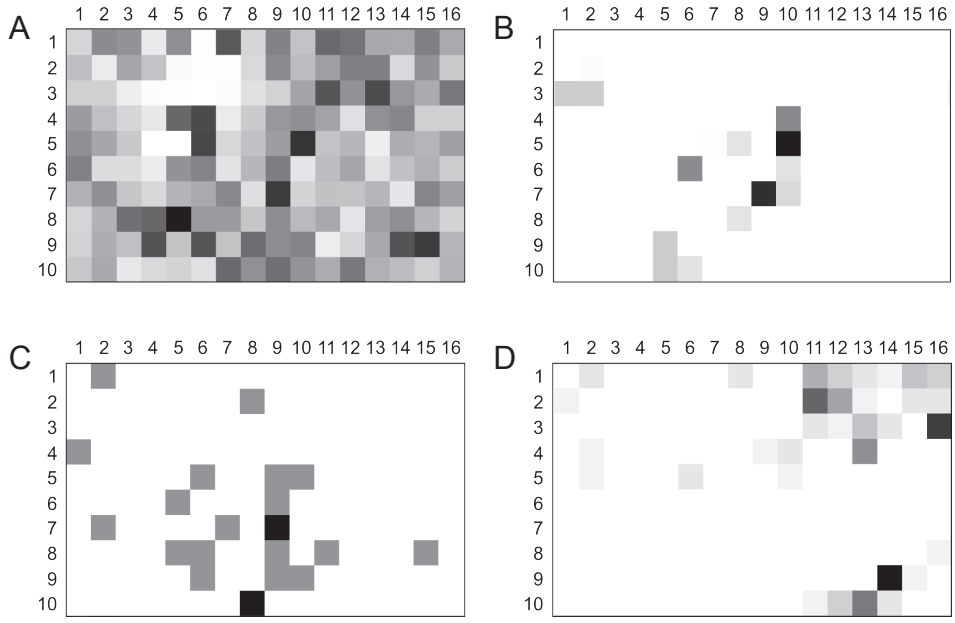
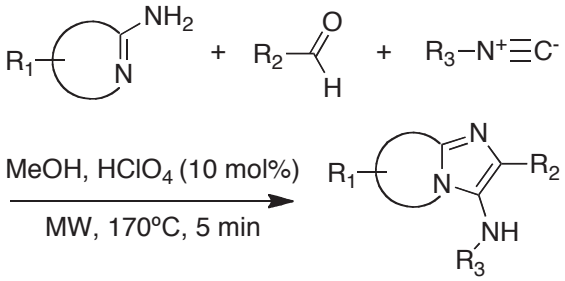
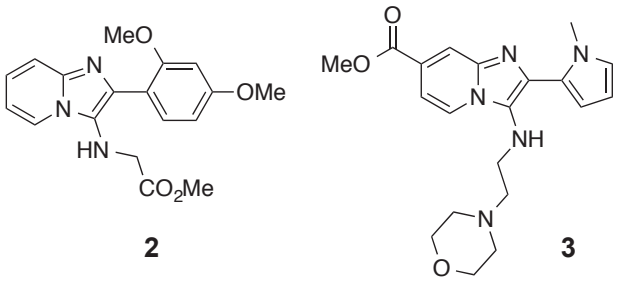


Figure 16. Toroidal self-organizing map (SOM) visualizing overall data density (**A**), distribution of the virtual combinatorial library (**B**), known PI3K inhibitors (**C**), and known muscarinic receptor ligands (**D**). 16×10 data clusters ("neurons", Voronoi fields) are shown as squares. Gray shading represents local compound density (note that the shading in each plot is scaled between minimal and maximal values). Compound **2** is located in cluster (9,7), compound **3** in cluster (10,5). For compound **2** an overlap with PI3K inhibitors is predicted. Compound **3** is found in a cluster that contains muscarinic receptor ligands.



Scheme 3. Ugi-type three-component reaction used for constructing a virtual combinatorial library and synthesizing selected compounds.



Scheme 4. Structures of compounds **2** and **3**.

We constructed a virtual combinatorial library from 12 aminopyridines, 40 aldehydes and 8 isocyanide building blocks, resulting in 3840 virtual products (Scheme 3 and Supplementary Information Chapter 8.2.2). To predict potential bioactivities for these compounds we computed their CATS similarity values to known drugs and lead structures (COBRA v11.10). Briefly, we trained a self-organizing neural network (SOM, Kohonen network) on the pool of COBRA reference compounds and the virtual combinatorial products, followed by visualization of compound distributions as a two-dimensional toroidal map (Figure 16)[325,376,377]. For the purpose of prediction, we only considered annotated targets of the reference compounds that were co-clustered with the combinatorial products. In this way, target predictions are limited to a conservative "application domain" of a reference compound cluster, and the risk of false-positive prediction is reduced[334-336]. For further target prioritization, we computed *p*-values from the similarity score distribution of the complete training data[471]. The *p*-values are an estimate of the probability of making a false-positive prediction (type-I error).

For the whole library, this method suggested six targets with average *p*-values < 0.01: phosphoinositide 3-kinase (PI3K), biphenyl-2,3-diol 1,2-dioxygenase, diacylglyceride O-acyltransferase, smoothened receptor, interleukin receptors, and cytochrome P450 reductase. We decided to investigate the PI3K prediction in more detail because this enzyme is a relevant drug target in antitumor research. Of note, the underlying scaffold was previously shown to afford PI3Kα inhibitors[472].

First, we synthesized and tested the nine top-predicted compounds for PI3Kα inhibition. In total, four of them exhibited the desired activity. Compound **2** turned out to be the most active (IC_{50} = 131 μM). Although the measured activities might be considered as weak, this result verifies the CATS+SOM-based approach for target prediction.

We then synthesized and tested an additional set of 57 compounds from the virtual combinatorial library, for which the highest joint prediction scores for PI3K and DNA topoisomerases were computed. Previous studies suggested that simultaneous inhibition of these two enzymes might allow for more efficient chemotherapy with reduced chemoresistance of tumor cells[473]. Molecules with a target profile that includes both these targets will constitute an important step in anti-cancer research. Moreover, the scaffold of our library has already been proven to produce bioactive compounds against both those targets[472,474]. In fact, in the present study six of our compounds, at a concentration of 75 μM, turned out to be moderately active against PI3Kα, where compound **3** was the most potent (IC_{50} = 230±30 μM). We wish to point out that we cannot completely rule out measurement artifacts caused by compound aggregation[475]. None of the 57 synthesized compounds inhibited human DNA topoisomerse II (EC 5.99.1.3), but in a preliminary test four of them inhibited bacterial DNA gyrase, a bacterial type II topoisomerase (EC 5.99.1.3) (data not shown). Apparently, the scaffold of the combinatorial library positions R-group vectors appropriately, but proper side-chain functionalities are required for potency and target selectivity. There is ample opportunity for optimizing compound **2** in this regard by including additional building blocks in the combinatorial synthesis. The SOM projection shown in Figure 16d may serve as a guide for structure optimization[185,376,377,476,477], as compound **2** is located in a sparsely populated region of the activity island formed by known muscarinic receptor ligands. Side-chain alteration could steer the design towards the center of the distribution thus potentially improving potency[356,478].

For comparison, we also predicted targets for the obtained PI3Kα inhibitors using SEA[337]. In SEA, compound **3** yielded no target predictions at all when using ChEMBL[72] as reference data. For the remaining compounds SEA reported maximal Tanimoto similarity below 0.35 and *E*-value > 1.2, rendering them low confidence predictions. Compound **2** was suggested as ligand of quinone reductase 2 (NQO2) and melatonin receptor 1B (MTNR1B). PI3K was not reported by SEA.

Finally, it is of particular note that CATS suggested human muscarinic receptor 1 (M1) ranking first on the target list computed just for compound **3**. In a first cell-based functional assay[479] compound **3,** in a concentration of 10 μM, actually exhibited substantial M1 agonistic activity yielding 34±5% of the effect caused by 100

nM acetylcholine. Follow-up concentration-dependent activity determination yielded an approximate EC_{50} of 5 µM for compound **3** (Figure 17). This result confirms the CATS+SOM-based target prediction as viable and de-orphanizes compound **3** as a novel (no entry in CAS[480]) functional M1 receptor agonist.

3.2.4 Conclusions

The results of this study corroborate CATS+SOM as a useful similarity approach for identifying pairs of molecules with similar bioactivity but different molecular scaffolds. Inclusion of the aromatic feature type in the CATS2 implementation increased enrichment in a retrospective analysis. Results of a preliminary prospective target-profiling study demonstrate that (i) the CATS2 descriptor may be employed to predict targets of virtually generated compounds with potential applications in *de novo* design and drug re-purposing, (ii) relying only on a single prediction algorithm bears the danger of missing relevant drug targets or focusing on false-positive predictions, and (iii) different molecular descriptors (here: CATS2; SEA with ECFP4 fingerprints) in combination with its associated knowledge base (here: COBRA or ChEMBL) complement each other in their domains of applicability. It will therefore be worthwhile to construct a prediction tool that is based on multiple reference databases, descriptors and models, *e.g.* as a jury decision approach. Whether activities in the micromolar range give rise to desired poly-pharmacology effects or turn out to be actually sufficient for drug re-purposing certainly depends on the particular pharmacological activity, therapeutic area, and intended application[309,332]. Many more practical examples will be required to allow for a statistically motivated assessment. Irrespective of the shortcomings of each method, our study validates ligand-based target prediction as viable for rapid compound profiling in medicinal chemistry and chemical biology.

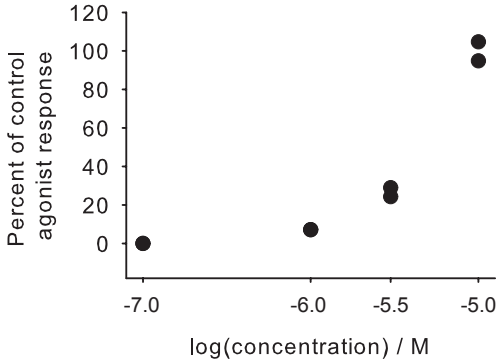


Figure 17. Concentration-dependent agonistic activity of compound **2** on the human M1 receptor. Acetylcholine served as positive control agonist (EC_{50} = 1.9 nM). At ligand concentrations > 10 µM compound **3** aggregated and interfered with the measurement (data not shown).

3.2.5 Publication details and contributions

Authors

Michael Reutlinger,‡[a] Christian P. Koch,‡[a] Daniel Reker,‡[a] Nickolay Todoroff,[a] Petra Schneider,[a] Tiago Rodrigues,[a] Gisbert Schneider [a]

‡ M.R., C.P.K. and D.R. contributed equally to this work.

[a] ETH Zurich, Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

Author contributions

M.R., C.P.K., and D.R. contributed equally.
M.R. implemented the CATS2 descriptor, conducted the retrospective evaluation, designed the research, analyzed the results, and contributed to the manuscript.

Acknowledgements

The Chemical Computing Group Inc. (Montreal, Canada) provided a research license of MOE.

Reference

Reutlinger M, Koch CP, Reker D, Todoroff N, Schneider P, Rodrigues T, Schneider G (2013) Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for "orphan" molecules. *Mol. Inf.* **32**, 133-138.

Licence

Source of funding

This research was supported by the Swiss National Science Foundation (grant 205321-134783), Deutsche Forschungsgemeinschaft (FOR1406TP4), and OPO-Foundation Zurich.

3.3 Combining on-chip synthesis of a focused combinatorial library with *in silico* target prediction reveals imidazopyridine GPCR ligands

Motivated by the results reported in Chapter 3.2, the decision to incorporate two complementary molecular representations in a machine-learning based method applicable to biological target prediction was made. *Gaussian Process* (GP) regression was used to build predictive affinity models based on data obtained from the ChEMBL chemogenomics database. The GP regression not only provides an expected affinity value but also an estimate of uncertainty, which can be combined in a single prediction score to reflect both aspects. The approach was theoretically evaluated for its predictive capabilities using a diverse range of protein targets. The computational approach was further complemented with an on-chip chemical synthesis system to prospectively evaluate its ability to identify new targets for a focused combinatorial library.

3.3.1 Abstract

For the example of the Ugi three-component reaction we report a fast and efficient microfluidic-assisted entry into the imidazopyridine scaffold, where building block prioritization was coupled to a new computational method for predicting ligand-target associations. We identified an innovative GPCR-modulating combinatorial chemotype featuring ligand-efficient adenosine $A_{1/2B}$ and adrenergic $\alpha_{1A/B}$ receptor antagonists. Our results suggest tight integration of microfluidics-assisted synthesis with computer-based target prediction as a viable approach to rapidly generate bioactivity-focused combinatorial compound libraries with high success rates.

3.3.2 Introduction

The fast pace of drug discovery programs is supported by high-throughput screening campaigns to identify new chemical entities, where the underlying screening compound collections benefit from combinatorial libraries with lead- and drug-like properties[481,482]. While numerous synthesis protocols are available, a reliable assessment of potential macromolecular targets of these compounds is desirable for

the compilation of bioactivity-focused combinatorial libraries. We demonstrate this concept taking the Ugi multicomponent reaction[483], which have shown robustness in producing both tool compounds and drug candidates[484,485], as an example. Imidazopyridines may be considered a privileged scaffold given their diverse range of macromolecular drug targets[470,474,486-491]. While entry into this chemotype through an Ugi-3 component reaction has been reported[470,486,487], these methods do not allow for the quick assembly of combinatorial libraries and scaling up. Therefore, our initial efforts focused on developing a robust and scalable process in flow using a continuous synthesis system equipped with low-pressure, pulsation-free syringe pumps. The setup included a 3-2-way solenoid valve to allow for automated building block filling-dispensing cycles. The amine and benzaldehyde components were dissolved in ethanol, together with perchloric acid, while the isocyanide component was pumped independently. Stock solution concentrations were adjusted to afford the desired final building block concentration in the microreactor. A borosilicate DeanFlow chip with a total volume of 5 µl and a zig-zag mixing zone was used as the primary reactor (Figure 18a). Alternatively, we used a KombiMix chip with a reaction volume of 13 µl (Figure 18b). The protocol was then scripted with Cetoni Q_{mix} Elements software to automate all steps, including washing of the microfluidic channels.

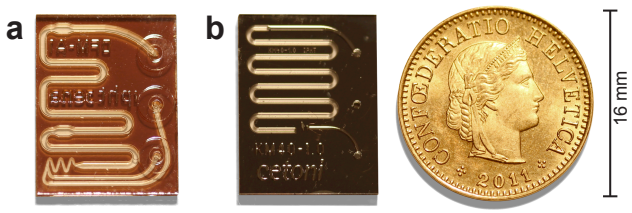


Figure 18. DeanFlow (**a**) and KombiMix (**b**) microreactor chips.

3.3.3 Material and Methods

Computational

For training the Gaussian process models[169,171] we used the ChEMBL database (version 14) containing 1,213,242 distinct compounds with 10,129,256 bioactivities for 9,003 targets[72]. Protein targets with fewer than 200 annotated human bioactivities were excluded. All activity end-points were standardized to p*Affinity* = -

$\log_{10}$(*activity*). The final affinity data set consisted of 209,293 compounds with 431,313 bioactivities for 469 human targets. Post-processing was conducted using Python (www.python.org) and Knime v.2.6.0.[492] Molecular structures were standardized using the "*wash*" function in MOE 2012.10 (The Chemical Computing Group Inc., Montreal, Canada); log*P*(o/w) was calculated with MOE. Two different molecular descriptors were calculated for each compound: topological pharmacophores (CATS2, 0-9 bonds, type-sensitive scaling)[486], and an ECFP-like topological circular fingerprint (Morgan fingerprint, *radius* = 4, 2048 bit; RDKit: www.rdkit.org)[36]. Predictive models were implemented using Matlab R2012b (The MathWorks Inc., Natick, USA) and the GPML toolbox v3.1 (www.gaussianprocess.org). We assessed prediction quality by 10-fold stratified cross-validation (cross-validated squared correlation coefficient, Q^2; *Mean Absolute Error*, MAE). The *Boltzmann-Enhanced Discrimination of ROC* (BEDROC; $\alpha = 56$, top 3% contribute 80% to the score) was used to quantify the early enrichment performance[53]. We used the lower confidence-bound p*Affinity* estimate throughout this study: $prediction = \mu_* - \sigma_*^2$, where μ_* is the model's predictive mean and σ_*^2 the predictive variance. To distinguish from random predictions we calculated the *Mahalanobis Distance* (MhD) of an activity *prediction*: MhD(*prediction*) = (*prediction* − μ_r) / σ_r, where μ_r and σ_r are the mean and standard deviation of a randomized predictive distribution. The background consisted of 50000 randomly selected molecules from ChemDB[493].

Synthesis.

Stock solutions of building blocks were prepared in ethanol. The amine and aldehyde components were premixed, and perchloric acid was added. Two independent syringe pumps delivered the amine/benzaldehyde/perchloric acid solution and the isocyanide solution at suitable flow rates. The reaction chamber containing the microchip was heated at different temperatures and the crude product was collected in a vial. The crude mixtures were purified by preparative HPLC (acetonitrile:H_2O + 0.1% formic acid in each solvent) using a gradient of 30-95% or 5-50% acetonitrile over 16 minutes. Microfluidics hardware and the Q_{mix} Elements software were from Cetoni (Korbußen, Germany). Microwave synthesis was performed in a Biotage Initiator (Uppsala, Sweden) in 1-2 ml vials, as described[486].

Testing.

Tests for determination of ligand binding and K_i values were performed at Cerep on a fee-for-service basis (Le Bois l'Evêque, B.P. 30001, 86600 Celle l'Evescault, France, www.cerep.fr). Assays were performed according to literature-described protocols (functional assays: hA_1 ref.[494]; hA_{2B} ref.[495]; ha_{1A} ref.[496]; ha_{2B} ref.[497]). Functional EC_{50} values obtained for the test compounds were converted to K_i values using the Cheng-Prusoff equation (Eq. 23):

$$\boldsymbol{K}_i = {\boldsymbol{EC}_{50}}/{\mathbf{1} + (\boldsymbol{A}/\mathrm{EC}_{50A})}, \tag{23}$$

where A = concentration of reference agonist in the assay, EC_{50A} = EC_{50} value of the reference agonist (Figure S9).

Direct binding to the human adenosine A_1 receptor was measured in a radioligand assay as described[498]. Percentage of binding was expressed as the mean of two independent measurements

3.3.4 Results and Discussion

As an initial screening of reaction conditions we performed sequential and automated synthesis of compound **1**. Conversion rates were derived from ^{1}H NMR spectra (Figures 19a, 19b). In first instance we investigated optimal flow rates and reaction temperatures, using 10 mol% of catalyst and a final concentration of each building block equal to 0.3 M, as described previously[486]. Generally, reactions at lower temperatures (30 and 70ºC) performed better than at 170 and 200ºC. Additionally, we conducted control reactions in glassware at room temperature and 30ºC for two hours, showing conversion rates of 73% and 80%, respectively. The results pinpoint the usefulness of a microreactor, both for improving conversion rates and drastically shortening reaction times. Reactions carried out under higher flow rates (30 and 60 µl × s^{-1}) performed worse than their 3.75 µl × s^{-1} and 7.5 µl × s^{-1} counterparts, possibly due to shorter residence times in the reactor chip. We observed the highest conversion at intermediate temperatures (70 and 100ºC). Interestingly, at 70ºC the reaction appears to be tolerant to a wide range of flow rates, while at 100ºC a rate of 15 µl × s^{-1} is preferable.

Having studied the best binary combination of temperature and flow rate, we screened for the ideal catalyst loading and final concentration of building blocks – 10 mol% and 0.3 M, respectively (Figure 19b). Comparable conversion rates were obtained in a microwave procedure and the setup described herein (94% *vs.* 93%, respectively)[486]. Of note, these results were obtained using a lower reaction temperature in the flow system (100ºC in flow *vs*. 170ºC for microwaves) and shorter reaction times (0.3 seconds in flow *vs.* 15 minutes in microwaves). Finally, the optimized reaction conditions were compared in the DeanFlow and KombiMix microreactors. While **1** was converted to 93% in the DeanFlow chip, an 88% conversion rate was observed in the KombiMix.

With these results in hand we synthesized a small focused library of imidazopyridines **4-15** (Figure 19c) using the DeanFlow reactor chip, and predicted potential biological targets with Gaussian Process regression models, built for 469 drug targets that are annotated in the ChEMBL database (version 14)[72]. Given a query compound, the computer model predicts p*Affinity* values for each target, which goes beyond related computational tools[100,333]. Furthermore, to ensure meaningful, non-trivial and high value predictions we calculated the *Mahalanobis Distance* (MhD) of the predicted values to the predictions made for a large collection of randomly selected molecules. Here, we considered only drug targets for which we obtained p*Affinity* > 5.5 and MhD > 0.5 standard deviations. With these mildly restrictive criteria we predicted an average of 18 targets per compound. Basically due to the low p*Affinity* bound this number exceeds other theoretical considerations and experimental findings reporting between 2 and 10 targets per drug, depending on the target class[499,500]. We obtained an average of four targets per imidazopyridine compound with the more conservative boundaries p*Affinity* > 6 and MhD > 1.

Keeping the permissive estimate we selected a total of 41 targets with high p*Affinity* predictions for further study. For these targets the model yielded favorable cross-validated accuracies of Q^2 = 0.68±0.10, MAE = 0.65±0.11 and BEDROC = 0.67±0.15 (all values *mean*±*stddev*)[51,52,501]. We finally selected five targets based on majority predictions for the whole library, potential pharmaceutical interest and assay availability. p*Affinity* values were in the micromolar range (Table 4), notwithstanding the models' high predictive variance. This observation emphasizes the potential scaffold novelty compared to known ligands in the ChEMBL database. In fact, to the

best of our knowledge, imidazopyridines with this framework have not been reported as adenosine or adrenergic receptor ligands (Chemical Abstracts Service, SciFinder, https://scifinder.cas.org/).

Having predicted potential macromolecular targets for all synthesized compounds, we selected those compounds for testing for which we obtained robust p*Affinity* predictions. For one of the prominent targets, phosphoinositide 3-kinase, activity had previously been reported for the underlying imidazopyridine scaffold[486], which corroborated the prediction. For proof-of-concept, we then explored a range of predicted G-protein coupled receptor (GPCR) targets aiming at the discovery of a new activity island in chemical space. In radioligand displacement assays probing direct ligand-receptor binding and cell-based functional activity assays, 71% of the compounds were found to be active as predicted (Table 4). More specifically, compounds **6** and **10** presented antagonistic K_i values of 2-3 μM, respectively, against the adrenergic a_{1B} receptor, while compound **5** showed similar low micromolar antagonistic potency against the adrenergic a_{1A} and adenosine A_{2B} receptors. Compounds **11** and **15** turned out to be potent direct A_1 receptor ligands (84% and 89% binding at 100 μM, respectively), but inactive in the functional cell-based assay. Additional tests will be required to determine selectivity profiles in a full GPCR panel screen.

Several quality indices have been suggested to guide hit prioritization in drug discovery[19,502]. Accordingly, our compounds fully qualify as lead structure candidates (Table 4). For example, compound **10** is a scarcely decorated, yet highly ligand-efficient chemical entity (*LE* = 0.40; *SILE* = 3.23) that might justify development as an adrenergic a_{1A} receptor antagonist. On the other hand, albeit less ligand efficient than **10**, compound **6** presents a better balance between affinity and computed log*P*(o/w) (*LLE* = 3.46 *vs.* 1.74). Most importantly, the leads presented herein are dissimilar to their nearest neighbors from the training data (structural similarity *Tanimoto* = 0.16–0.30, Table S1) and would likely not have been selected using straightforward substructure-based similarity searching.

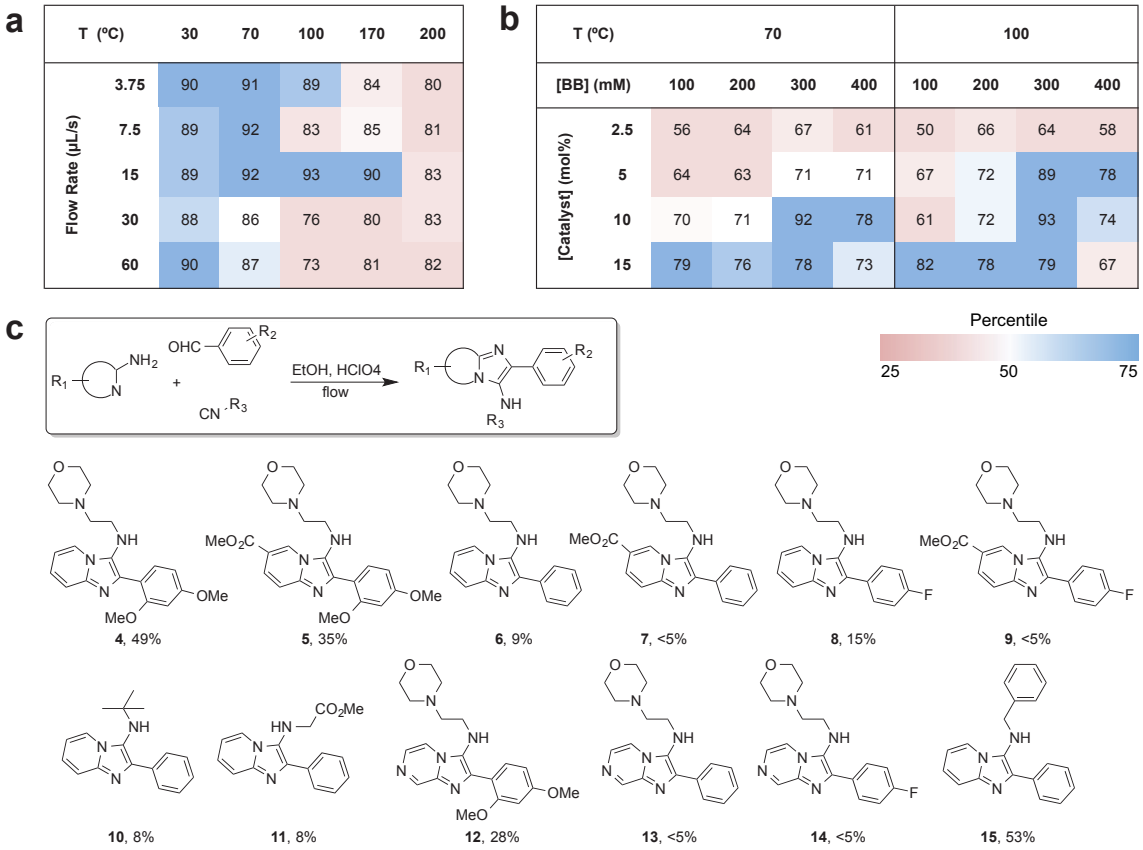


Figure 19. Synthesis of imidazopyridines in flow: (**a**) Screening of optimal flow rate and temperature (T), at constant catalyst loading (10 mol%) and *Building Block* (BB) concentration (0.3 M); (**b**) Screening of optimal catalyst loading and building block concentration, at fixed flow rate (15 µl/s) and temperature (70 and 100ºC); (**c**) Focused library synthesized in the present study and isolated yields.

Table 4. Summary of results for selected compounds **3**, **6**, **8**, **10**, **11**, **12** and **15**.

Cmpd.	Target	Predicted p*Affinity*	Mahalanobis distance	Experimental pK_i or % binding	*LE*[a]	*LLE*[b]	*SILE*[c]
3	a_{1A}[d] / PDE10A[e]	5.7 / 5.7	0.7 / 0.8	<4 / <4	-	-	-
6	a_{1B}[f]	6.2	2.4	5.6	0.33	3.46	3.04
8	a_{1A} / A_{2B}[g]	5.8 / 6.5	0.7 / 2.4	5.4 / 5.2	0.30 / 0.29	3.07 / 2.86	2.87 / 2.76
10	a_{1B}	6.1	2.0	5.7	0.40	1.74	3.23
11	A_1[h]	5.7	3.2	> 80%[i]	-	-	-
12	A_{2B} / PDE10A	6.4 / 5.8	2.6 / 1.7	<4 / <4	-	-	-
15	A_1	6.0	3.3	> 80%[i]	-	-	-

[a] ligand efficiency; [b] lipophilic ligand efficiency; [c] size-independent ligand efficiency; [d] adrenergic a_{1A} receptor; [e] phosphodiesterase 10A; [f] adrenergic a_{1B} receptor; [g] adenosine A_{2B} receptor; [h] adenosine A_1 receptor; [i] radioligand assay; activity values are averaged from duplicate measurements.

3.3.5 Conclusion

Altogether, our chemistry-driven approach to target-focused combinatorial library design, in an expeditious and efficient manner, led to the identification of a molecular framework targeting four GPCRs. The results highlight the imidazopyridine scaffold

as a privileged motif and demonstrate how the integration of emerging technologies in drug discovery, such as on-chip synthesis and computational target prediction, may advance hit and lead identification in chemical biology and molecular medicine. In light of recent advances in lab-on-a-chip technologies[503-505], one could even envisage a fully automated hit finding automaton that integrates computational target prediction and building block selection for the microfluidic-assisted synthesis and testing of candidate compounds.

3.3.6 Publication details and contributions

Authors

Michael Reutlinger,[a] Tiago Rodrigues,[a] Petra Schneider,[a] Gisbert Schneider [a]

[a] ETH Zurich, Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

Author contributions

M.R. designed and conducted the research, developed the target prediction, performed the synthesis, analyzed the results, and contributed to the manuscript.

Reference

Reutlinger *et al*. (2014) Combining on-chip synthesis of a focused combinatorial library with *in silico* target prediction reveals imidazopyridine GPCR ligands. *Angew. Chem. Int. Ed.* **53**, *582-585*.

Licence

Source of funding

This research was financially supported by a research grant from the OPO Foundation, Zurich.

3.4 Combinatorial chemistry by ant colony optimization

Combinatorial reactions theoretically give access to large sets of diverse molecules. In this study, *Ant Colony Optimization* (ACO) was introduced as a method to efficiently identify focused compound libraries from the complete virtually accessible combinatorial space. To assess its capabilities, ACO was initially applied to peptide synthesis to find sets of MHC-1 binding peptides and further extended to arbitrary combinatorial reactions. The enhanced small-molecule ACO design process was then evaluated for its potential to adaptively prioritize small-molecule building blocks and guide a design process.

3.4.1 Abstract

We present the implementation and practical application of ACO[506,507] to adaptive peptide design, which implicitly generates a task-specific molecular similarity metric (*Molecular Ant Algorithm*, MAntA). We chose the design of novel major histocompatibility complex I (MHC-I, mouse H-2K^b allele) molecule-stabilizing octapeptides as an example, and demonstrate that this modelling method may be used to find local clusters of peptides with desired properties ("activity islands")[37], without the necessity for full peptide library enumeration. In a second practical application, we extend the combinatorial approach to combinatorial chemistry, taking the *Ugi three-Component Reaction* (Ugi-3CR) as an illustrative example. We show that the computer-based design method is able to identify ideally suited molecular building blocks, and present a successfully synthesized and tested combinatorial product exhibiting effective inhibitory activity against human blood clotting factor Xa. This result qualifies ACO-based approaches for efficient combinatorial synthesis planning aiming at new hit and lead compounds for the pharmaceutical sciences.

3.4.2 Introduction

Peptides are currently experiencing a renaissance as tool compounds and lead structures in pharmaceutical research and chemical biology[508], specifically through

combinations of computational, chemical, and biological approaches[509,510]. While solid-phase synthesis and *in vitro* activity testing allow for analysing several thousand peptides at a time, exhaustive peptide libraries become prohibitively impracticable with growing peptide length. As an elegant alternative, phage display technologies offer parallel access up to approximately 10^{15} sequences[511,512]. While this biochemical approach will deliver peptides with desired properties and activities for a large variety of applications, it also suffers from several limitations. For example, very hydrophobic sequences and peptides that kill or otherwise affect the host bacteria used for phage production will elude identification by phage display. When time and resources are limited and a model or design hypothesis is available, *de novo* computer-based peptide generation constitutes an alternative for providing solutions of the combinatorial peptide optimization problem. This methodology is based on a predictive model of peptide activity (the objective or "fitness" function), and a robust optimization method for navigation in sequence space towards regions of high-predicted fitness[290].

According to the *Principle of Strong Causality*[23] – a term derived from technical optimization, that was rephrased in the context of quantitative structure-activity relationships and molecular design as the *Chemical Similarity Principle*[24] – systematic combinatorial optimization requires a type of function-related order among the molecular building blocks that are used for compound construction. Otherwise, one would always perform a "random" search for optimal products. In other words, a *context-sensitive* similarity metric is needed so that small structural variations between molecular building blocks result in only small changes of the predicted and measured activity. Molecular similarity could thus be considered as a context-dependent property[356,425]. Nature-inspired optimization methods have been shown to lead to practical solutions of this combinatorial design task and related problems in molecular modelling[513,514].

3.4.3 Materials and Methods

Ant colony optimization of peptide sequences (MAntA).

The MAntA algorithm is based on the MMAS approach introduced by Stützle and Hoos[515], which is one of the most successful *Ant Colony Optimization* (ACO)

algorithm variants in practice[507]. Here , we provide a brief description of MMAS, based on the original paper by these authors, to highlight the modifications that were implemented in MAntA. Originally, MMAS was developed to improve exploitation while avoiding early stagnation. It differs from other ACO algorithms (*e.g.,* Ant System[516] or Ant Colony System[517]) in three key aspects:

1) Only a single ant adds pheromones to the path.
2) The pheromone intensity is bound to an adaptive min/max interval.
3) Pheromone levels are initialized at the interval's upper bound to foster initial exploration.

MMAS is best described using the well-studied traveling salesman problem (TSP): "*Given a complete graph with n vertices (cities) and distances d as edge labels, find the shortest closed tour visiting each of the cities exactly once*." Let *m* be the number of ants (*colony size*) and $\tau_{ij}(t)$ the amount of pheromone deposited on the edge (*i,j*) at time step *t*. In each iteration each ant constructs a tour based on a probabilistic decision rule, which is biased by the pheromone trail $\tau_{ij}(t)$ and by locally available heuristic information η_{ij}, which for TSP is usually defined as $1/d_{ij}$. Ants prefer cities, which are close to their actual location and have a high pheromone concentration on the connecting edge. An ant *k*, currently located at position *i*, chooses to go to position *j* with probability (Eq. 24):

$$p_{ij}^{k}(t) = \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}{\Sigma_{j\in\mathcal{N}_i^k}[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}} \text{ if } j \in \mathcal{N}_i^k \,, \tag{24}$$

where α and β determine the relative importance of pheromone concentration and heuristic information and $\mathcal{N}_i^k$ represent the set of vertices (cities) that have not yet been visited by ant *k*. After all ants have completed the construction a fraction of the pheromone evaporates, and only edges involved in the best solution receive additional pheromone according to the following update rule (Eq. 25):

$$\tau_{ij}(t+1) = \rho\, \tau_{ij}(t) + \Delta\tau_{ij}^{best} \text{ , where } \Delta\tau_{ij}^{best} = 1/\mathrm{f}(s^{best}), \tag{25}$$

and ρ is the pheromone persistence. $\mathrm{f}(s^{best})$ gives the cost of either the global-best or iteration-best solution, where latter has been used in MAntA. To avoid stagnation of the search process, MMAS ensures that the pheromone concentration does not

exceed the interval $[\tau_{min}, \tau_{max}]$. After a new best solution has been found, these boundaries are updated according to Eq. 26 and Eq. 27:

$$\tau_{max} = \frac{1}{1-\rho}\frac{1}{f(s^{gb})}, \tag{26}$$

$$\tau_{min} = \frac{\tau_{max}(1-\sqrt[n]{p_{best}})}{(avg-1)\sqrt[n]{p_{best}}}, \tag{27}$$

where $f(s^{gb})$ is the cost of the best solution so far, $avg = n/2$, and p_{best} is the probability that the best solution found is constructed in a fully converged MMAS process. The process of tour construction, solution evaluation and pheromone deposition is iterated until a convergence criterion is met. In MAntA we limited the iterations by a fixed amount of evaluated unique solutions (fixed budget constraint).

In MAntA, the MMAS algorithm has been modified to allow for solving the subset selection problem for peptide design. Because of the lack of a path, ants deposit pheromones directly on the nodes or in the case of peptides on the amino acids at each sequence position. Accordingly, $\tau_{ij}(t)$ is the pheromone concentration for amino acid *j* at position *i*, and η_{ij} is the heuristic information. When constructing a peptide, an ant probabilistically selects exactly one amino acid per sequence position according to Eq. 28:

$$p_{ij}(t) = \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{j\in\mathcal{N}_i}[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}} \text{ with } j \in \mathcal{N}_i \ , \tag{28}$$

where *i* is the ant's current position, values α and β determine the relative importance of pheromone concentration and heuristic information, and $\mathcal{N}_i$ is the set of possible amino acids at position *i*. Only the iteration-best solution is used for the pheromone update according to Eq. 25 to favour exploration of the search space. The solution cost $\Delta\tau_{ij}^{best}$ is calculated using the trained jury predictor with $\Delta\tau_{ij}^{best} = 1/(1-\mathcal{F}(s^{ib}))$, and $\mathcal{F}(s)$ is the score for solution *s*. If multiple ants share the best score, the solution components from all best solutions are updated with additional pheromone. In the presented applications the heuristic information was set uniformly. For the small-molecule example, the amino acids were replaced by combinatorial building blocks, and the solution cost was calculated as $\Delta\tau_{ij}^{best} = 1/\mathcal{F}(s^{ib})$, with $\mathcal{F}(s)$ the affinity information for solution *s* from the virtual assay. The

chosen values for the free parameters are given in Table 6. Values for p_{best}, ρ and *colony size* were selected based on systematic testing (Figure 20).

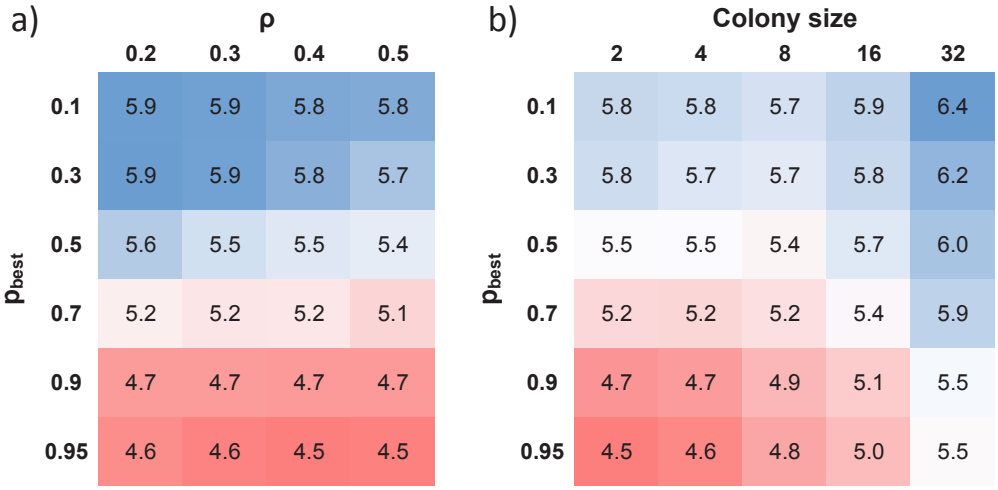


Figure 20. Systematic evaluation of the influence of parameter values of p_{best}, and ρ on the quality of the selected focused library. Values are given as the mean IC_{50} values [µM] of the 20 most active compounds. Cells are colored according to the IC_{50} values with a gradient from *blue* = lowest activity to *red* = highest activity.

Peptide scoring function.

We used a previously implemented cascaded classifier model[518] based on *Support Vector Machines* (SVM)[133] and multilayer *Artificial Neural Networks* (ANN)[107]. Model development and training is described in detail in the original publication[518]. Figure 21 shows a schematic of the model's architecture. It consists of two SVMs and two ANNs providing the input to a jury network, which computes a prediction score between zero and one. Koch *et al.* have recently performed an analysis of the residue positions and properties, which are most relevant for class separation[519].

Stochastic neighbor embedding (SNE).

For dimensionality reduction (PPCA descriptor) we used the implementation of SNE from the Matlab Toolbox for Dimensionality Reduction v0.7.2 and Matlab 7.13.0 (The MathWorks Inc., Natick, USA)[520]. The SNE perplexity parameter was set to 30 for all calculations. To focus on uncommon motifs, peptides with canonical residue motifs were eliminated beforehand.

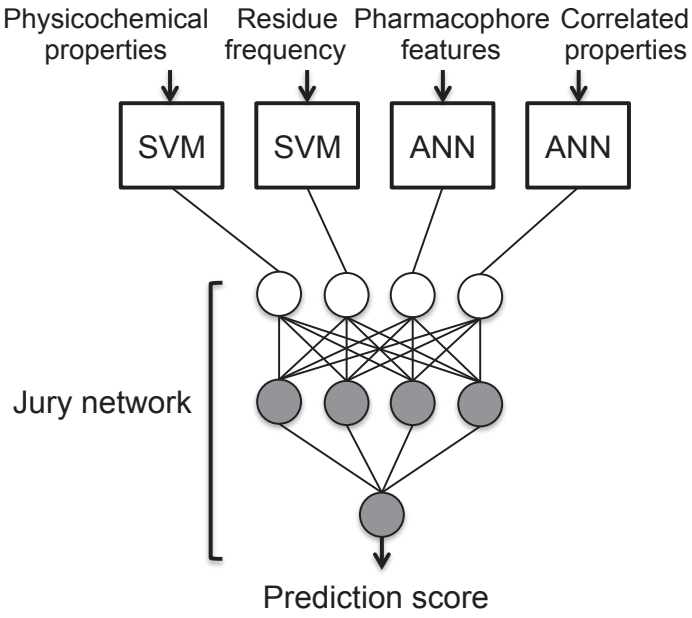


Figure 21. Schematic of the machine-learning model used for peptide activity prediction. It consists of a jury network that combines individual prediction scores from four classifiers (two *Support Vector Machines*, SVM, and two *Artificial Neural Networks*, ANN, as first-stage filters) to a single value between zero (non-stabilizing) and 1 (H-2K^b stabilizing). First-stage filters are based on different peptide representations.

Sequence logos.

For sequence logo depiction we used the web application WebLogo v2.8.2 (http://weblogo.berkeley.edu/logo.cgi)[521]. Only peptides with a prediction score > 0.8 were included in sequence logo calculation.

Peptide synthesis and analytics.

Peptides were synthesized on an Overture™ robotic solid phase peptide synthesizer (Protein Technologies, Tucson, USA) on a 10 µmol scale utilizing ten-fold excess of Fmoc-protected amino acids (200 mM) over Fmoc-Wang resin. Resins, amino acids and HCTU (O-(6-chloro-1-hydroxybenzotriazol-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate) were purchased from AAPPTEC (Louisville, USA). DMF (dimethylformamide), DCM (dichlormethane), diisopropylether, piperidine and TIPS (triisopropylsilane) were purchased from Sigma-Aldrich (Buchs, Switzerland); NMM (4-methylmorpholine) and TFA (2,2,2-Trifluoroacetic acid) from Fisher Scientific (Wohlen, Switzerland). Deprotection was performed using 20% piperidine in DMF for 2 x 5 min. Double coupling was executed conservatively in rotation 2 x 15 min under utilization of 1:1:4 200 mM amino acid / 200 mM HCTU / 800 mM NMM in DMF. After

deprotection and double coupling washing with DMF was performed for 4 x 30 sec. Automated cleavage was performed for 2 h with 95%/2.5%/2.5% TFA/H_2O/TIPS after multiple washing with DCM (4 x 30 sec). Product was precipitated out of the final TFA-peptide solution in ice-cold diisopropylether, rewashed and dried under nitrogen gas. Peptides were used in the experiments without further purification. Peptide products were analysed on a LC-20A rpHPLC instrument (Shimadzu, Reinach, Switzerland) using a C18, 110 Å, 1.8 µm, 100 x 3 mm column (Macherey-Nagel, Düren, Germany). A linear gradient of 30-95% ACN/H_2O(0.1% TFA) over 12 min with a flow rate of 0.5 ml x min^{-1} was applied. Shimadzu SPD-20A Prominence UV-VIS detector was used for detection at 210 nm. Masses were detected on positive mode between 500-1500 Da with a Shimadzu LCMS-2020 single-quad mass spectrometer (ESI+).

Peptide sequences with calculated molecular weight (*mw*, unit: Da), retention time (R_t, unit: minutes), and observed masses (*m+*):
AQFQYTNA (*mw* = 942.0, R_t = 1.01, *m/z* = 943.4, 944.7), KSIFFSNP (*mw* = 939.1, R_t = 1.16, *m/z* = 938.6), RGLSFTPG (*mw* = 833.9, R_t = 1.17, *m/z* = 834.5, 835.5, 836.5), VSPYFRAE (*mw* = 968.1, R_t = 1.03, *m/z* = 968.6), RFFDLYSL (*mw* = 1060.2, R_t = 3.46, *m/z* = 531.0, 531.9, 1060.6, 1061.6), IGWEIGTL (*mw* = 888.0, R_t = 3.29, *m/z* = 888.5, 889.5, 890.5), VNLRAYLL (*mw* = 961.2, R_t = 2.59, *m/z* = 961.6, 962.6, 963.6), KIFHLVSL (*mw* = 956.2, R_t = 2.29, *m/z* = 956.6), RTLNPPPL (*mw* = 907.1, R_t = 1.18, *m/z* = 907.6, 909.6), IAFPWSIK (*mw* = 961.2, R_t = 2.92, *m/z* = 961.6, 962.6), INLNPPPG (*mw* = 820.9, R_t = 0.97, *m/z* = 821.5), RSVQWINK (*mw* = 1030.2, R_t = 0.95, *m/z* = 516.1, 1030.7), ISPQGGPS (*mw* = 741.8, R_t = 0.92, *m/z* = 742.4), WWYAYHHI (*mw* = 1175.3, R_t = 1.94, *m/z* =588.5, 588.6), TWYLYNEI (*mw* = 1101.2, R_t = 3.16, *m/z* = 551.5, 551.6, 1101.6), EYYDYEAV (*mw* = 1051.1, R_t = 1.17, *m/z* = 526.4, 1051.5, 1052.3, 1053.7), SIINFEKL (*mw* = 963.1, R_t = 2.7, *m/z* = 963.7, 965.6), QDNGHDWI (*mw* = 984.0, R_t = 1.43, *m/z* = 984.5, 985.5).

Cell-based MHC stabilization assay.

The stabilization assay was conducted as previously described[522] with TAP-deficient RMA-S cells[523] and mouse mutagenized Rauscher-virus induced T-lymphoma cells (RMA cells) as control. Briefly, cells were cultivated in RPMI (Gibco-BRL, Karlsruhe, Germany) with 5% FCS (Biochrom, Berlin, Germany) at 37°C under 8% CO_2. Prior to the assay, the cells were kept at 26°C for 16h to promote stabilization of peptide-free MHC-I H-$2K^b$ on the plasma membrane. Subsequently, the cells were incubated for 30 min at 26°C with peptide concentrations ranging from 100 to 1 × 10−3 µg/ml (in 10 *x*-fold serial dilutions) followed by incubation for 45 min at 37°C inducing denaturation of unloaded H-$2K^b$. Remaining peptide-loaded H-$2K^b$ located at the cell surface was measured by flow cell cytometry by combining the H-$2K^b$ specific

monoclonal antibody B8.24.3 (G. Köhler, Basel Institute of Immunology) and FITC (fluorescein isothiocyanate) labelled rat anti-mouse IgG1 or goat anti-mouse IgG secondary antibodies (BD Pharmingen, Heidelberg, Germany). The mean fluorescence intensity (MFI) was taken as measure for the H-2K^b stabilizing effect by protein-bound peptide. SC_{50} values were calculated as the peptide concentration leading to half-maximal MFI (maximal FTI = response measured for the stabilizing peptide SIINFEKL). Measurements were performed with a FACSCalibur flow cell cytometer (BD Bioscience, Heidelberg, Germany), data analysis with WINMDI (The Scripps Research Institute, La Jolla, USA). SC_{50} calculation including data normalization and linearization was done with Excel 2011 (Microsoft Corporation, Redmond, USA).

Synthesis

Compound **16** (*tert*-butyl 4-(2-((3-carbamimidoylphenyl)amino)-2-(2-((3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2*H*-pyran-2-yl)oxy)phenyl)acetamido)piperidine-1-carboxylate).

Benzamidine (0.24 mmol) was dissolved with toluene (1 ml) and triethylamine (0.7 mmol). A suspension of aldehyde (0.24 mmol) in water (0.5 ml) was added, and the mixture left reacting at room temperature for 1 h. Phenylphosphinic acid (10 mol%) and isocyanide (0.24 mmol) were added subsequently to the solution. The reaction mixture was stirred for 24 h at room temperature.

Enzyme inhibition assays.

Assays were performed at Reaction Biology Corp. (Malvern, PA, USA), as follows: Factor VIIa (Biomol Cat. No SE-438), Cathepsin S, L and Fluorogenic Substrate (Peptide sequence: Z-Val-Val-Arg-AMC.HCl [Z=Cbz=Benzyloxycarbonyl; AMC=7-amino-4-methylcoumarin] (Biomol Cat. No P-199) were prepared in fresh reaction buffer (25 mM Tris pH 8.0, 100 mM NaCl, 0.01% Brij35; 1% DMSO in buffer was added before use). Enzyme solution was delivered into the reaction well, followed by the test compounds and substrate solution. The final concentration of the enzyme and substrate was 10 µg/ml and 10 µM, respectively. The enzyme activities were monitored (Ex/Em 355/460) as a time-course measurement of the increase in fluorescence signal from fluorescently labelled peptide substrate for 120 minutes at

room temperature. Same protocol using different buffers, and substrates was used for the other proteases. For the factor Xa (Biomol Cat. No SE-362) assay, a buffer consisting of 25 mM Tris pH 8.0, 100 mM NaCl, 0.01% Brij35, with 1% DMSO and 0.25mg/ml BSA added to the buffer before use was employed. The final enzyme concentration in the assay was 2 mU/ml. Pefafluor™ FXa, fluorogenic peptide substrate for factor Xa (sequence: CH_3SO_2-D-CHA-Gly-Arg-AMC-AcOH, final concentration of 10 µM) was used. For thrombin (Enzyme Research Lab Cat. No HT-1002a, final concentration of 20 mU/ml) assay the buffer consisted of 25 mM Tris pH 8.0, 100 mM NaCl, 0.01% Brij35, with 1% DMSO, 0.25mM $CaCl_2$ and 1.0 mg/ml BSA added to the buffer before use, and Pefafluor™ TH, fluorogenic peptide substrate for thrombin (sequence: H-D-CHA-Ala-Arg-AMC.2AcOH, final concentration of 10 µM) were used. Finally, for trypsin (Sigma Cat. No T-1426; final concentration of 0.8 µg/ml) the same buffer as for factor VIIa was used, as well as the same substrate as in the thrombin assay.

3.4.4 Results and discussion

Peptide design by ant colony optimization

The first goal of this study was to test the ability of the MAntA method to identify MHC-I H-2K^b stabilizing octapeptides without the requirement for full enumeration of all 20^8 ($\approx 10^{10}$) possible peptides. Computationally, sequence lengths up to eight residues can still be handled exhaustively[518,519]. Thus, the application of MAntA to H-2K^b stabilizing octapeptides should be regarded as a proof-of-concept study. We focused on sequences that display discrepancies between the presence of a known canonical MHC-I binding motif and the predicted H-2K^b stabilizing potential. The known canonical residue pattern defines residue positions 5 [Y, F] and 8 [aliphatic] as relevant "anchors" for H-2K^b-peptide interaction. The term "anchor" is an interpretation of observed residue conservation in known MHC-I binding peptides[524]. Accordingly we defined three classes of peptides for candidate selection:

Category A - Partial agreement with the canonical motif (either residue at position 5 or 8) and predicted as H-2K^b stabilizing;
Category B - Lack of canonical motif residues, but predicted as stabilizing;
Category C - Complete motif fulfilled, but predicted as non-stabilizing.

For prediction of MHC-I stabilizing potential, we used a machine-learning model as fitness function, which computes a score between 0 and 1 for a given octapeptide sequence. Score values greater than 0.5 suggest H-2K^b stabilizing potential with higher values corresponding to higher model confidence. We previously developed this cascaded ensemble model using 996 known H-2K^b stabilizing and non-stabilizing octapeptides with sustained predictive accuracy[518].

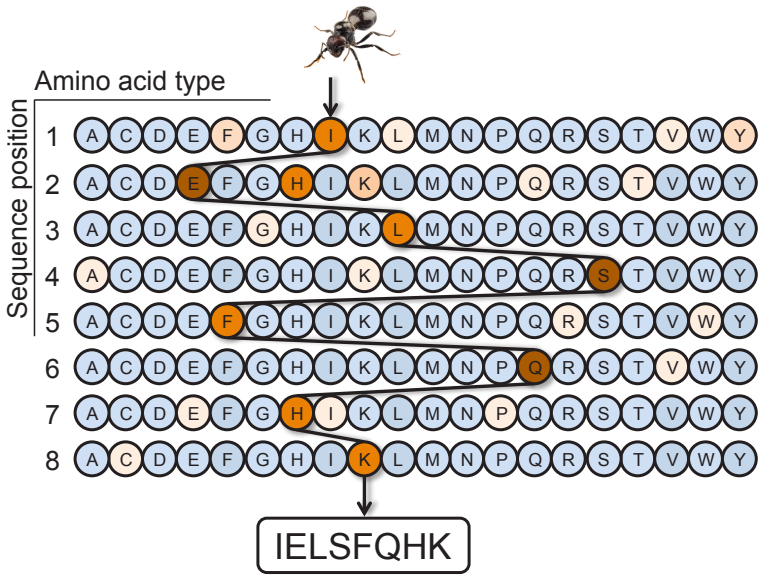


Figure 22. Schematic of an ant run showing the pheromone matrix of the search space. Eight horizontal bars represent the residue positions of the octapeptides. At each position, there are pheromone variables for 20 standard amino acids (circles). Colours represent the pheromone concentration at each amino acid (*blue* = low, *red* = high). The black line (top to bottom) represents an ant path corresponding to the currently preferred peptide sequence.

For target-driven sampling of peptides we employed our new method MAntA, which implements an ACO algorithm. It simulates the process how ants find the shortest path between a food source and their nest using pheromone trails. MAntA revises this concept so that a colony of artificial search agents (ants) builds solutions (peptides), evaluates the quality of the solutions (fitness, measured activity), and updates the ants' pheromone trails proportional to the quality of the solutions. The design of peptides is modelled as a subset selection problem with the constraint that exactly one amino acid has to be selected for each position (Figure 22). Pheromones are directly deposited on the nodes, which correspond to the amino acids at each position. In each peptide construction step, the ants select the next solution component (amino acid) based on a probabilistic decision rule. This probabilistic choice is biased by the pheromone trail and by available heuristic information. The

pheromone trail is updated after each design cycle according to the predictions by the fitness function (machine-learning model). Only the amino acid residues of the fittest solution (*i.e.*, the currently best peptide in terms of the computed score) receive additional pheromone, which is mathematically expressed by increasing probabilities. Consequently, preferred ant paths emerge in the search space (here: 20^8 peptides) that correspond to potential high-quality amino acid sequences. To reduce the risk of early convergence of the search process a small fraction of pheromones evaporates after each.

In the present study we generated artificial ant colonies with 10 ants and allowed a maximum of 100 peptides to be sampled. The colony size of 10 ants was chosen to boost initial explorative behaviour and increase the chance of discovering more diverse local optima. We ran the optimization 100 times. The resulting 10000 peptides were represented by the PPCA descriptor (19 principal component scores of residue properties)[106]. For visual inspection of the peptide distribution these high-dimensional property data containing 19 × 8 = 152 descriptor values were projected to three dimensions using a nonlinear method termed *stochastic neighbor embedding*, which minimises information loss during data projection and retains local compound neighborhoods (Figure 23)[382,525]. Apparently, peptides predicted to be active (grey-coloured dots in Figure 23a) possess limited residue diversity, as indicated by low Shannon entropy, while diverse residue patterns yielding great Shannon entropy values are characteristic for the inactive centre of the projected space. This observation suggests that the machine-learning model has extracted potential function-relevant features from MHC-I stabilizing peptides. Since individual MAntA runs converged in disjunct regions of sequence space resulting in different local clusters of peptides, we conclude that (i) there are multiple classes of peptide sequences that might stabilize the same MHC-I allele, and (ii) ant colony optimization converged to several local optima. Such a visualization of chemical space helps identify pharmacologically interesting regions and select promising candidate compounds (here: peptides) for synthesis and activity testing[525].

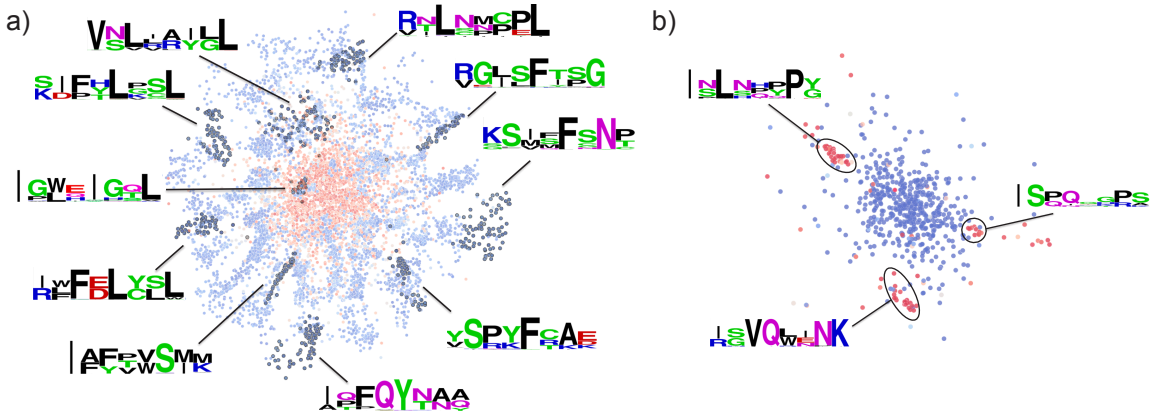


Figure 23. (**a**) Peptide distribution obtained by compressing the 152-dimensional PPCA representation to three dimensions using stochastic neighbor embedding (SNE). Dots represent individual peptides. Colouring is according to Shannon entropy (*blue* = low, *red* = high). Selected optimization runs are shown in darker colour and with the corresponding sequence logo for the peptides generated by MAntA. (**b**) Peptide distribution after elimination of sequences containing the canonical residue motif for MHC-I binding. Three distinct clusters with positive predictions remained. Peptide positions in the projection are coloured according to prediction score (*blue* = negative, *red* = positive).

The 100 independent MAntA runs resulted in 100 distinct clusters for the peptides with prediction scores exceeding a value of 0.8, from which we chose 10 well separated clusters (Figure 23a; Table 5 peptides 1-10). Peptides containing the canonical residue motif were eliminated after dimensionality reduction (Figure 23b; Table 5 peptides 11-13). From this reduced set three clusters were chosen for peptide picking. Residue patterns present in local sequence clusters were visualized as sequence logos[526]. From each selected cluster we picked one representative peptide compliant to the corresponding sequence logo for synthesis and testing (Table 5). Additionally, we included three category C peptides (Table 5 peptides 14-16). Evidently, prediction accuracy varies for the different peptide categories A, B and C. Based on bias in the training data, sequences being close to training samples are comparably easy to categorize, while the activities of sequences being profoundly different to the training data (B) are more difficult to predict.

Peptides were tested for their ability to stabilize murine H-2K^b on the surface of TAP-deficient mutagenized Rauscher virus-induced T lymphoma cells using an established assay system[527]. H-2K^b stabilization is a necessary step for complex formation with the cognate T cell receptor, and served as an indirect measurement of peptide binding to H-2K^b. The natural epitope sequence SIINFEKL served as a positive control[528], for which we determined SC_{50} values (peptide concentration yielding half-maximal H-2K^b stabilization on the plasma membrane) of 0.02 µM and 0.04 µM in two independent

measurements. The sequence QDNGHDWI was used as a negative control, which was confirmed by lack of measureable activity.

All category A peptides were predicted and experimentally confirmed to stabilize H-2K^b, the most potent peptide being AQFQYTNA (SC_{50} = 0.2 μM, 0.4 μM). This result confirms the sustained predictive potential of the machine-learning model used for peptide sampling, and reliable convergence of the search process. Of note, the stabilizing potencies of these peptides were an order of magnitude weaker than the positive control, which might be attributed to their only partial match with the full canonical residue motif.

For category B peptides completely lacking the canonical motif only the sequence IAFPWSIK exhibited some moderate activity (SC_{50} = 5.5 μM, 10.3 μM); the other three candidates were inactive in the cell-based assay. Apparently, the predictions of potential MHC-I binding peptides are dominated by partial or full compliance with the canonical motif. This likely is a consequence of model training, for which the majority of positive examples (59%) possessed appropriate residues in anchor positions 5 and 8[518,519].

Notably, presence of the motif alone seems to be insufficient for H-2K^b stabilization, as shown for the category C peptides: One peptide (EYYDYEAV) was predicted and experimentally confirmed as inactive, despite full compliance with the canonical residue pattern in the anchor positions. Two category C peptides (WWYAYHHI, TWYLYNEI) were moderately active but predicted as inactive.

The results of biochemical activity determination support the previously found relevance of appropriate residues in the anchor positions of MHC-I binding peptides. They also point to additionally required features, as shown by the sequence EYYDYEAV, which turned out to be inactive despite the appropriate residues in the anchor positions. Thereby, our machine-learning model not only correctly predicted this sequence as non-stabilizing, which clearly demonstrates that this model exceeds a simple pattern matching approach based on residue motifs, but also confirms earlier experimental findings[529]. Overall, we observed 11/16 = 69% correct predictions in this prospective screening study with MAntA. Based on the suggested peptide clusters resulting from this ant colony optimization algorithm we were able to select peptides with desired sequence features and corresponding biological activity.

Table 5. Results of cell-based activity determination.[a]

No.	Peptide category	Amino acid sequence	SC_{50} (µM)	Prediction
1	A	AQFQYTNA	0.2, 0.4	stabilizing
2	A	IGWEIGTL	0.5, 0.9	stabilizing
3	A	KSIFFSNP	0.5, 1.2	stabilizing
4	A	RFFDLYSL	0.6, 0.7	stabilizing
5	A	VNLRAYLL	0.6, 0.8	stabilizing
6	A	KIFHLVSL	1.3, 2.6	stabilizing
7	A	VSPYFRAE	2.8, 5.0	stabilizing
8	A	RGLSFTPG	3.1, 3.5	stabilizing
9	A	RTLNPPPL	3.9, 6.0	stabilizing
10	B	IAFPWSIK	5.5, 10.3	stabilizing
11	B	INLNPPPG	*inactive*	stabilizing
12	B	RSVQWINK	*inactive*	stabilizing
13	B	ISPQGGPS	*inactive*	stabilizing
14	C	WWYAYHHI	2.2, 6.5	non-stabilizing
15	C	TWYLYNEI	3.1, 9.1	non-stabilizing
16	C	EYYDYEAV	*inactive*	non-stabilizing
17	Positive control	SIINFEKL	0.02, 0.04	stabilizing
18	Negative control	QDNGHDWI	*inactive*	non-stabilizing

[a] Peptides were tested for their ability to stabilize H-2K[b] molecules on the surface of TAP-deficient cells (n = 2 independent experiments; SC_{50}: peptide concentration that resulted in half-maximal protein stabilization). Residues in agreement with the canonical binding motif are highlighted. Category A: partial agreement (either pos. 5 or pos. 8) and predicted as stabilizing; Category B: lack of canonical motif and predicted as stabilizing; Category C: complete motif and predicted as non-stabilizing.

Application of MAntA to combinatorial chemistry

As a pioneering application of ACO to combinatorial library design and molecular building block selection we chose the Ugi-3CR, which essentially represents a one-pot approach to the condensation of an aldehyde, amine and isocyanide[530,531]. We used MAntA to prioritize building blocks with the aim of finding an inhibitor of human factor Xa, as a representative of trypsin-like serine proteases involved in the blood-clotting cascade, which previously served as targets for stochastic nature-inspired ligand optimization[532-536]. The stock of structures contained 43 aldehydes, 15 amines, and 24 isocyanides, giving access to 15,480 potential products. Illgen *et al.* had synthesized and tested this whole combinatorial array and compiled a database containing the activity data of reaction mixtures (IC_{50} values)[537,538], which we employed as a virtual assay system for our computational approach. In a proof-of-concept study, we restricted the number of virtual products and tests to 300 (approx.

2% of the complete combinatorial library), because we wanted to mimic a realistic early-phase drug discovery scenario with a limited budget for synthesis and testing, *i.e.* without exhaustive compound generation. To this end we tested different ant colony sizes for their ability to compile focused low IC_{50} libraries (Figure 24a). From the results obtained, one can observe that with larger colony size the mean IC_{50} of the top 20 compounds increases. This may be due to the additional agents exploring a larger variety of local optima. Such behaviour is desirable if the fitness space contains diverse local optima. In the present study the task was to compile a focused compound library under simulated budget constraints, which resulted in the selection of a colony size of two ants (Table 6).

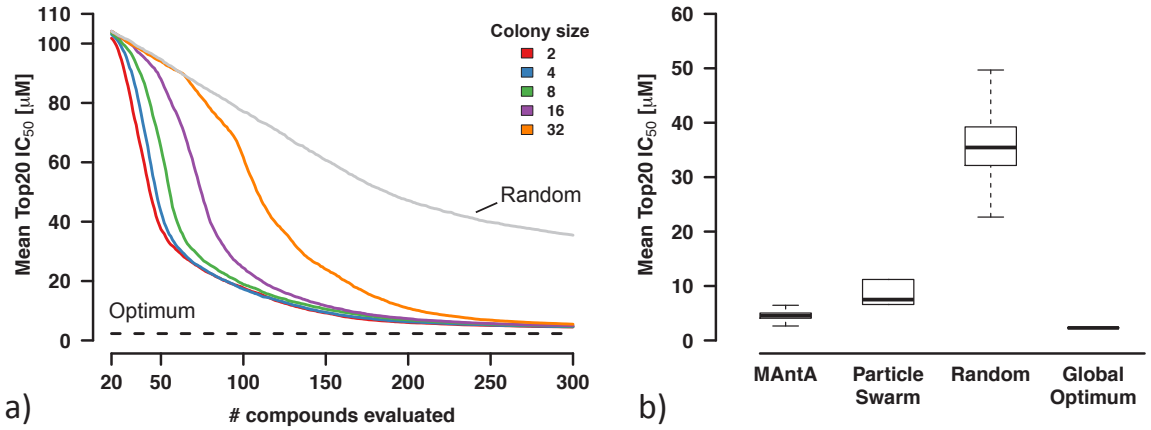


Figure 24. Performance evaluation of the MAntA algorithm for the optimization of factor Xa compounds based on the mean IC_{50} [µM] of the 20 most active compounds. (**a**) Influence of the MAntA colony size on the optimization performance relative to the optimization progress (in terms of evaluated compounds). The grey line shows the result of *pseudo*-random searching, and the dashed line the global optimum. Median values of 1000 repetitions are plotted. (**b**) Comparison of overall performance of MAntA, particle swarm algorithm, random search and the theoretical global optimum. Values for *Particle Swarm* are taken from Ref.[539] (lacking error estimates).

Table 6. MAntA parameters.

	Peptide design	Combinatorial chemistry
Strategy	Explorative sampling	Focused library compilation
Colony size	10	2
ρ	0.95	0.4
p_{best}	0.05	0.95
α	1	1
β	1	1

The pheromone matrix was constructed in analogy to the MHC-I peptide example, and updated in a model-free approach using the measured factor Xa inhibitory activity (IC_{50} value) of compounds stored in the reference database, thereby simulating a direct feedback loop. Figure 25 presents the resulting relative frequencies of building blocks among the generated virtual products. It is immediately apparent that the algorithm focused on few preferred building blocks. Apparently, the three top-ranking benzamidines from the amines set were selected for their ability to bind to the arginine side chain (S1) pocket of trypsin-like serine proteases[540]. This result demonstrates that the algorithm was able to focus on fragments that are well known as preferred building blocks for the S1 position. To evaluate the performance in some more detail, we compared MAntA to a previously reported particle swarm algorithm that had been applied to the identical molecular design task (Figure 24b). The particle swarm algorithm employed by Schüller and Schneider was selected as it emerged from their studies as the best performing algorithm on the Ugi data set[539], compared to random search, simulated annealing and the (1,λ) evolution strategy[537,541,542]. Our preliminary results indicate that MAntA outperforms the particle swarm implementation and might be a preferable computational method for combinatorial library design.

We synthesized compound **16** from the most frequently selected (highest pheromone concentrations) building blocks (amine no. 12, aldehyde no. 25, isocyanide no. 21; Scheme 5) following a slightly modified version of the protocol described by Illgen *et al.*, *i.e.*, using phenylphosphinic acid as catalyst (Scheme 5)[537,538]. LC-MS analysis of the reaction mixture after 24 hours revealed the desired compound in approx. 5% yield – in line with reports on similar reactions[538]. The crude mixture was then tested *in vitro* against factors Xa, VIIa, thrombin and trypsin. While no activity was observed against factors VIIa and thrombin, the reaction mixture presented IC_{50} values of 3.4 ± 0.1 μM (Figure 26) and 23 ± 1 μM (not shown) against factor Xa and trypsin, respectively, assuming complete educt conversion. It is of note that the same reaction mixture had exhibited higher activity (IC_{50} = 0.2 μM) against factor Xa in a previous report[537]. Different assay conditions, actual product concentration, and specifically lower enzyme concentration, can be accounted for the observed difference. Compound **1**, the structure of which we disclose here (Scheme 5), is the most potent Ugi-3CR type factor Xa inhibitor in the reference database.

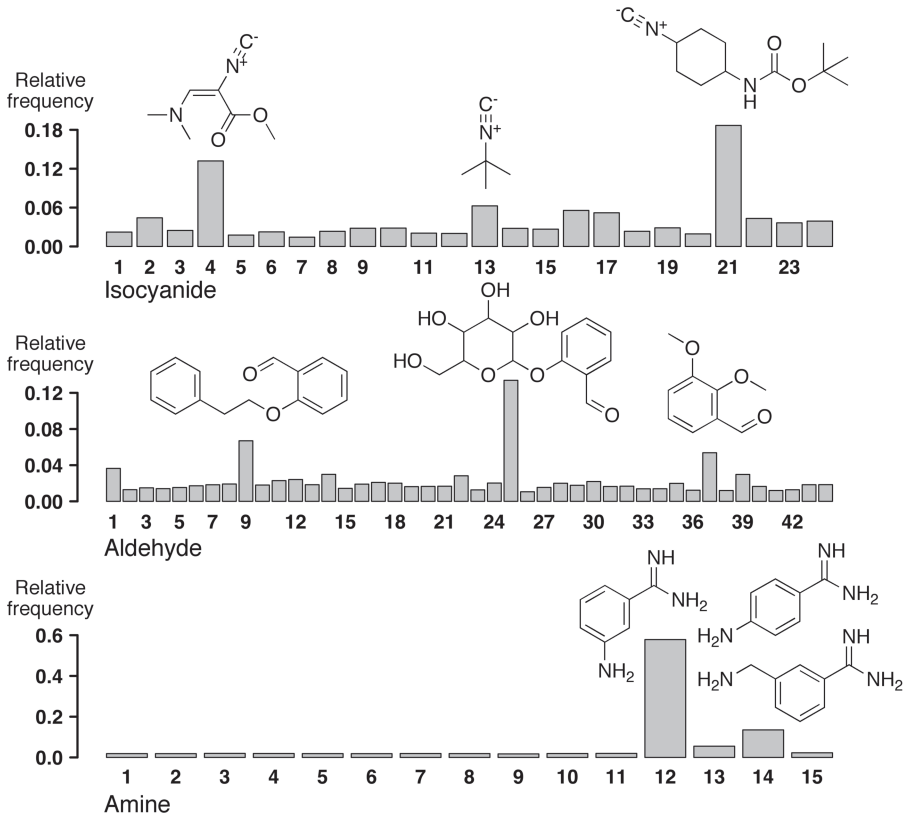
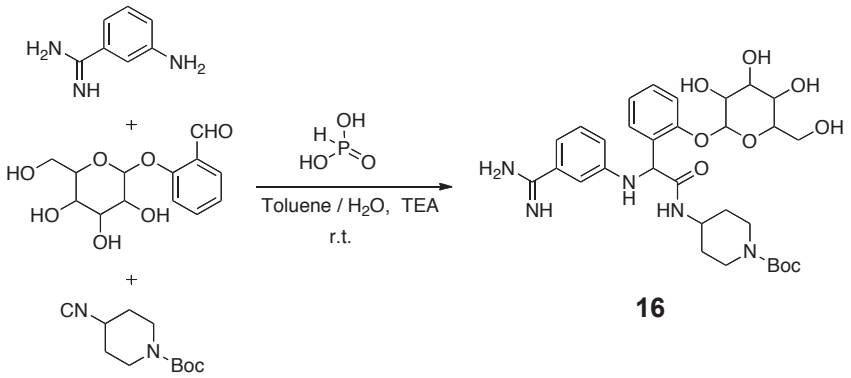


Figure 25. Relative frequencies of building blocks in the designed Ugi-3CR products. Bars give distributions obtained from 1000 independent MAntA runs with 300 evaluations per run.



Scheme 5. Synthesis of compound **16** by Ugi-type three-component reaction.

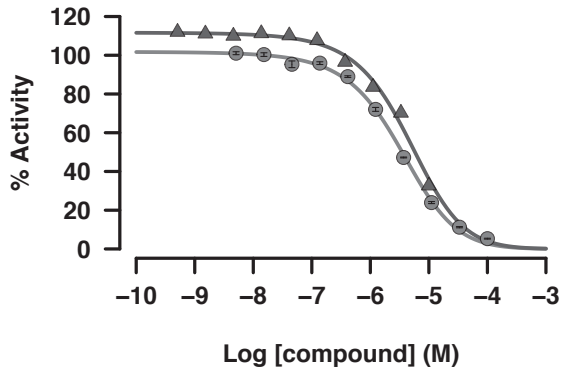


Figure 26. IC_{50} curves obtained for the inhibitory activity of compound **16** (*grey circles*) and gabexate mesilate (IC_{50} = 4.7 μM, *black triangles*) against factor Xa (n = 3).

3.4.5 Conclusion

Ant colony algorithms have been successfully used to solve the travelling salesman problem, resulting in the optimal ordering of a set of objects, *e.g.* the shortest path connecting a group of cities. In contrast to applications, in which an order of objects (e.g. cities) is to be determined, the design of peptides or other combinatorial reaction products aim at the best combination of molecular building blocks. Such a subset selection problem differs from other combinatorial problems because there is no intuitive path concept[543]. We had previously introduced the ACO paradigm to peptide design[544,545], and in the present study extended it to arbitrary combinatorial reactions implementing a modified *MAX-MIN Ant System* (MMAS)[515]. For the peptide experiments, we used a nonlinear projection for visualizing the underlying SAR landscape. Apparently, several local sequence clusters presenting several residue motifs characterize the group of MHC-I stabilizing peptides. While the known canonical motif discovered by Rammensee and coworkers[524] is fully or partially present in the majority of known MHC-I stabilizing peptides, there are additional, yet to explore sequence clusters containing bioactive peptides but lacking this residues pattern. We envisage computational peptide design as one of the key technologies for obtaining peptides with desired properties, e.g. for future personalized vaccine design, and the design of selective antimicrobial and cell-penetrating peptides. Such techniques will be particularly useful in cases where biochemical methods, e.g. phage-display, are prohibitive. Smart combinatorial design approaches might also be coupled to high-throughput peptide synthesis.

Overall, we have demonstrated that ACO-inspired peptide sampling in combination with a machine-learning model serving as fitness function may be used for finding novel bioactive amino acid sequences without the need to provide a fixed residue grouping based on amino acid similarity. It is evident that different prediction models bare the possibility of leading to different local optima in chemical space. We therefore advocate the use of multi-model fitness functions, unless a biochemical experiment is used to serve this purpose. This concept is immediately transferrable to other types of combinatorial chemistry where a compound is considered as constructed from a defined number of molecular building blocks or fragments. Our

MAntA approach is tailored to exploring combinatorial libraries without the need for full product synthesis. As a consequent next development, the algorithm could be coupled to a continuous flow-synthesis system with inline analytics, so that rapid hit prototyping becomes possible in a fully automated fashion.

3.4.6 Publication details and contributions

Authors

Jan A. Hiss,‡[a] Michael Reutlinger,‡[a] Christian P. Koch,‡[a] Anna M. Perna,[a] Petra Schneider,[a] Tiago Rodrigues,[a] Sarah Haller,[a] Gerd Folkers,[a,b] Lutz Weber,[c] Renato B. Baleeiro,[d] Peter Walden,[d] Paul Wrede,[e] Gisbert Schneider [a]

‡ J.A.H, M.R., and C.P.K. contributed equally.

[a] ETH Zurich, Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

[b] Collegium Helveticum, Schmelzbergstr. 25, 8092 Zurich, Switzerland.

[c] OntoChem GmbH, H.-Damerow-Str. 4, 06120 Halle/Saale, Germany.

[d] Charité-Universitätsmedizin Berlin, Department of Dermatology, Venerology and Allergology, Clinical Research Group Tumour Immunology, Charitéplatz 1, 10117 Berlin, Germany.

[e] Charité-Universitätsmedizin Berlin, Molecular Biology and Bioinformatics, Arnimallee 22, 14195 Berlin, Germany.

Author contributions

J.A.H, M.R., and C.P.K. contributed equally.

M.R. designed and executed the ACO research, provided the visualizations, analysed the results, and contributed to the manuscript.

Acknowledgements

The authors thank Morphochem AG for providing the Ugi data collection.

Reference

Hiss JA, Reutlinger M, Koch CP, Perna AM, Schneider P, Rodrigues T, Haller S, Folkers G, Weber L, Baleeiro RB, Walden P, Wrede P, Schneider G (2014) Combinatorial chemistry by ant colony optimization. *Future Med. Chem.* **6**, 267-280.

Licence

Source of funding

The research was financially supported by the ETH Zurich, the Swiss National Science Foundation (grant no. 205321-134783), and the OPO-Foundation Zurich.

3.5 Multi-objective molecular *de novo* design by adaptive fragment prioritization

Encouraged by the promising results obtained with the individual methods, the potential use as an integrated multi-target *de novo* design process (MAntA) was subsequently investigated. Focusing on therapeutic targets relevant for treatment of neuropsychiatric disorders, it was assessed whether the MAntA approach is capable of proposing innovative compounds matching a desired target profile.

3.5.1 Abstract

We present the development and application of a computational molecular de novo design method for obtaining bioactive compounds with desired on- and off-target binding. The approach translates the nature-inspired concept of ant colony optimization to combinatorial building block selection. By relying on publicly available structure-activity data, we developed a predictive quantitative polypharmacology model for 640 human drug targets. By taking reductive amination as an example of a privileged reaction, we obtained novel subtype-selective and multi-target modulating dopamine D_4 antagonists, as well as sigma-1 receptor-selective ligands with accurately predicted affinities. Nanomolar potency of the hits obtained, their high ligand efficiencies, and an overall success rate of 90% demonstrate that this ligand-based computer-aided molecular design method may guide target-focused combinatorial chemistry.

3.5.2 Introduction

Traditional combinatorial chemistry aims at the generation of large diverse compound arrays for bioactivity screening[546]. It has been realized that multiple "adaptive" synthesis-and-test cycles using smaller, focused compound libraries might be better suited, faster, and more economical to find lead-like bioactive compounds[19,356,547]. Computational molecular design methods offer the additional advantage to generate bioactive compounds while considering multiple objectives in

parallel[300,548,549], and combinatorial libraries with desired properties can be obtained by relying on chemistry-oriented computational molecular design[100,550]. Though potentially appealing, these methods have rarely been prospectively applied. Here, we present the comprehensive application of a computational concept for designing combinatorial libraries that exhibit an accurately predicted bioactivity profile. We show that the *Molecular Ant Algorithm* (MAntA)[513,551] effectively transfers a nature-inspired optimization principle to chemistry-driven molecular design. For proof-of-concept we focused on the reductive amination reaction as a scheme for combinatorial synthesis. By automated structure optimization, MAntA generated small compound libraries with lead-like qualities, high hit rates, and nanomolar activities. It implements a new design strategy that is applicable to all kinds of chemistry-driven computational methods[1,273,279], and does neither require prior knowledge about the bioactivity of scaffold classes nor is it limited to privileged scaffolds. In a retrospective study, ant colony optimization turned out to perform better or *en par* with other optimization methods[551]. Here, we pioneer the concept of polypharmacology-based molecular *de novo* design using combinatorial chemistry. We demonstrate that both target-selective, and multi-target modulating members of large combinatorial compound libraries are rapidly identified without the need for full library enumeration and synthesis.

3.5.3 Materials and Methods

Compound data

For machine learning, we used data from the ChEMBL (v14) database containing 1,213,242 distinct compounds with a total of 10,129,256 annotated bioactivities for 9,003 targets. The raw ChEMBL activity data were restricted to half-maximum inhibitory concentration (IC_{50}), K_i or K_d values and the corresponding log transformations. Prior to model building, all data were filtered according to the following ChEMBL criteria:

(i) Confidence score ≥ 7 (direct protein complex subunit, homologous single protein or direct single protein assigned);

(ii) Relationship type D (direct protein target) or H (homologue protein target);

(iii) Target type = Protein;

(iv) Measured activity or annotated inactive.

Protein targets with fewer than 200 annotated bioactivities were excluded. Activity end-points were standardized to p*Affinity* = -$\log_{10}$(*activity*). Only annotated inactive compounds were included in the training data. These define a region in chemical space, which should be avoided for designing potential candidates. To prevent the machine learning algorithms from learning an artificial tested *vs*. untested boundary we considered only experimentally validated inactive compounds for machine learning. End-points annotated as inactive entities were assigned the lowest p*Affinity* value of the corresponding target. Unrealistic entries with p*Affinity* less than 3 or greater than 12 were excluded. For compounds with multiple measurements per target the arithmetic mean and standard deviation were calculated. Compounds with a standard deviation > 0.5 log units were excluded for the corresponding target. The final affinity data set consisted of 279,866 compounds with 569,725 bioactivities for 640 human targets, which were used for training machine learning models. Filtering and post processing was conducted using Python (www.python.org) and Knime (www.knime.org). Prior to descriptor calculation all chemical structures were standardized using the "wash" function in MOE 2012.10 (The Chemical Computing Group Inc., Montreal, Canada). For each molecule, we computed a topological pharmacophore feature descriptor (CATS2, correlation distance: 0-9 bonds, type-sensitive scaling)[486] using in-house software, and an ECFP-like circular fingerprint (Morgan fingerprint, radius: 4, 2048 bits)[36] using RDKit (www.rdkit.org).

Gaussian process models

To model the nonlinear structure-activity relationship *Gaussian Process* (GP) regression models were trained individually for each of the 640 targets[171]. In the following a short introduction to GP modelling is given following the description given by Rasmussen and Williams[171]. A GP defines a distribution over functions $p(f)$, where f is a mapping of an input space $\mathcal{X}$ to $\mathbb{R}$. It is used as a *prior* for Bayesian

inference. The GP is fully specified by its mean function *m* and the covariance function (kernel function) *k*. By defining kernel functions on different representations of the input, one can capture heterogeneous perspectives on the input. We used a linear combination of an isotropic squared exponential kernel k_{SE} (CATS2 descriptor) and a Tanimoto kernel k_{TM} (Morgan fingerprint)[552] (Eq. 29).

$$\begin{aligned} k_{SE}(\mathbf{x},\mathbf{x}') &= \sigma_0^2 exp\left(-\frac{\left|\mathbf{x}-\mathbf{x}'\right|^2}{2\ell^2}\right), \\ k_{TM}(\mathbf{x},\mathbf{x}') &= \sigma_1^2\, T(\mathbf{x},\mathbf{x}'), \\ k\left(\mathbf{x},\mathbf{x}'\right) &= k_{SE}\left(\mathbf{x},\mathbf{x}'\right) + k_{TM}\left(\mathbf{x},\mathbf{x}'\right), \end{aligned} \tag{29}$$

where $T\left(\mathbf{x},\mathbf{x}'\right)$ is the Tanimoto structural similarity for fixed-sized bit vectors, and $\theta = \{\sigma_0, \ell, \sigma_1\}$ are adjustable hyperparameters. Hyperparameter values were directly estimated from the data by optimizing the logarithmic evidence of the data given the hyperparameters θ. For prediction of pAffinity values, we computed a lower confidence bound $\mu_* - \sigma_*^2$ instead of directly using the predictive mean. All Gaussian process models were implemented using Matlab R2012b (The MathWorks Inc., Natick, USA) and the GPML toolbox v3.1.

Combinatorial design

For exploration of combinatorial space we developed the *Molecular Ant Algorithm* (MAntA) as a modified *Ant Colony Optimization* (ACO) algorithm[506,507], based on the *MAX-MIN Ant System* (MMAS) approach introduced by Stützle and Hoos[515]. In ACO artificial ants deposit pheromones on edges of a graph. The amount of pheromones is determined by the fitness of the solution. In MMAS only edges belonging to the best solution are considered and pheromone levels are bound to a min/max interval. The deposited pheromones evaporate over time, enabling explorative behaviour and avoiding early stagnation in suboptimal solutions. Eventually the optimal solution, combination of edges with the greatest fitness emerges. We modelled the problem of finding the optimal configuration of building blocks as a subset selection problem. Because of the lack of a path concept, ants deposit pheromones directly on the nodes, corresponding to individual molecular building blocks. Per iteration each ant of the

colony builds candidate solutions by selecting exactly one building block per *R*-group position according to probabilities defined in Eq. 30:

$$p_{ij}(t) = \frac{[\tau_{ij}(t)]}{\sum_{l \in \mathcal{N}_i}[\tau_{ij}(t)]} \; j \in \mathcal{N}_i \;, \tag{30}$$

where $\tau_{ij}(t)$ is the pheromone concentration for building block j at position i at iteration t, and $\mathcal{N}_i$ is the set of allowed molecular building blocks at position *i.* Only the best solution found in one iteration contributed to pheromone update (Eq. 31):

$$\tau_{ij}(t+1) = \rho\, \tau_{ij}(t) + \Delta\tau_{ij}^{best} \text{ and } \Delta\tau_{ij}^{best} = 1/\mathcal{F}\left(s^{ib}\right), \tag{31}$$

where ρ is the pheromone persistence, and $\mathcal{F}\left(s^{ib}\right)$ the fitness of the solution. Fitness was computed by coupling the design process to the Gaussian process regression model. Pheromone bounds $[\tau_{min}, \tau_{max}]$ were enforced after each iteration and updated accordingly. To allow for exploration of the local neighborhood of promising solutions additional pheromone was added using the scaled Tanimoto structural distance between building blocks and the global best solution. To allow batch calls to the Gaussian process regression prediction framework a tabu list was introduced, forcing ants to build only novel solutions. Free parameters were set to *colony size* = 100, ρ = 0.8, p_{best} = 0.9, ν = 0.3. The design process was stopped after exploration of 0.1% of the combinatorial space. Results from three individual runs were merged and the best 3,000 designs were kept for further investigation. ReactionMQL was used for virtual synthesis of the selected educts, according to the predefined synthesis scheme[553].

Landscape visualization

Chemical library distribution and activity landscapes were visualized in LiSARD[478]. A two-dimensional molecule distribution was obtained by compressing the 210-dimensional CATS2 descriptor using *Stochastic Neighbor Embedding* (SNE)[382]. The ability to preserve the local neighborhood was evaluated using rank based quality metrics (Table S7)[554]. Local neighborhood influence in SNE was set to 30. By adding the measured activity data for each molecule as an additional dimension a three-

dimensional point distribution was obtained. For fitting a surface to the data points the Nadayara-Watson estimator was applied[438,439]. The value for a given location is estimated as a locally weighted average of the data, using a multivariate Gaussian kernel as weighing function. The smoothing factor *k* was set to 0.5.

Performance assessment

The quality of the predictive models was analysed using 10-fold stratified cross-validation. For stratification the training labels were split in three p*Affinity* categories (< 5; 5-7; > 7). The performance was evaluated in regard to regression and early enrichment potential, and compared to binary kernel discrimination and random forest regression methods (Supplementary Table S4 and Figure S12-S13). To evaluate the regression performance the cross-validated squared correlation coefficient (Q^2, Eq. 32), as a measure for the overall fit, and the *Mean Absolute Error* (MAE, Eq. 33) were calculated[51,52,501]. To assess the ability of the models to distinguish between active compounds and an assumed inactive background the early enrichment performance was analysed. The *Boltzmann-enhanced discrimination of ROC* (BEDROC, Eq. 34)[53] (α = 56, top 3% contribute 80% to the score) and the percentage of active compounds retrieved in the top 3% (Recall 3%) were used to quantify to performance. As a negative control the target labels (*y*-scrambling) were randomly permutated and the validation procedure was repeated (Supplementary Table S5). The overall ranking performance was visualized using ROC curves based on the combined cross-validation folds (Supplementary Figure S1-S2). For rank analysis the training data was enriched with a random inactive background of 50,000 compounds. For labelling the ChEMBL data a threshold of p*Affinity* = 5 (10 µM) was used (≤ 5 = inactive; > 5 = active). Let *N* be the total number of compounds in the dataset and *n* is the number of actives.

$$Q^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{32}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{33}$$

with y_i denoting the target label for the i^{th} compound, $\hat{y}_i$ the prediction and N being the total number of compounds.

$$BEDROC = RIE \times \frac{\frac{n}{N}\sinh\left(\frac{\alpha}{2}\right)}{cosh\left(\frac{\alpha}{2}\right) - cosh\left(\frac{\alpha}{2} - \alpha\frac{n}{N}\right)} + \frac{1}{1 - e^{\alpha\left(1-\frac{n}{N}\right)}}, \quad (34)$$

where $RIE = \sum_{i=1}^{n} e^{-\alpha x_i} \Big/ \frac{n}{N}\left(\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}\right)$, x_i is the relative rank of the i^{th} active in the ordered list, and α is the early recognition tuning parameter.

Random Forest regression

The Random Forest[183] approach belongs to the class of ensemble machine learning methods, in contrast to Kernel methods *e.g.* Gaussian processes. A random forest is an ensemble of tree predictors $h(\boldsymbol{x}, \theta)$ where $\mathbf{x}$ is the descriptor vector of length p and θ a numerical random vector. Individual trees are grown to the maximum size, but only a randomly selected subset of m_{try} descriptors out of the total p is used for tree definition. For the ensemble of n_{tree} trees, the prediction is given by the unweighted average over the ensemble[189]. The models were developed using the Random Forest package for Matlab (www.mathworks.com) implemented by Abhishek Jaiantilal (https://code.google.com/p/randomforest-matlab). Free parameters were set to m_{try} = sqrt(p) = 47 and n_{tree} = 500.

Multiple-molecule query virtual screening

Binary Kernel Discrimination (BKD) was implemented using the Harper kernel function[555] in combination with the Tanimoto similarity coefficient according to Chen *et al.*[556] with bandwidth parameter λ = 0.6 and β = 1.0. The BKD method was implemented using Matlab R2012b (www.mathworks.com).

Synthesis

We compiled chemical building blocks from three vendors (Chembridge, Sigma-Aldrich, Maybridge). Undesired structural motifs were eliminated (Table S3),

resulting in 7062 amines and 2879 aldehydes and ketones. The required aldehyde or ketone (1 molar eq.) and amine (1 molar eq.) starting materials were dissolved or suspended in 1,2-dichloroethane (5 mL/mmol). $NaBH(OAc)_3$ (1.5-2.5 molar eq.) was added and the mixture left reacting at room temperature until completion (overnight). In the case of ketones, acetic acid (1.5 molar eq.) was used as catalyst. The crude product was washed with water (10 mL) and brine (10 mL) before evaporating the organic solvent under reduced pressure. Products **17-32** were purified from reverse-phase flash chromatography using an 5-50% acetonitrile:H_2O (+0.1% formic acid) gradient run over 20 minutes.

Binding assays

Binding assays were performed at Cerep (Celle l'Evescault, France, www.cerep.fr) on a fee-for-service basis. K_i values for $D_{1\text{-}5}$ (Cerep assays ref. 0044, 0046, 0048, 0049, 0050)[557-561], and sigma-1 (ref. 0889)[562] receptors, as well as histamine H_3 (ref. 1332)[563], serotonin 5-HT_{1A} (ref. 0131)[564], δ (ref. 0114)[565], κ (ref. 1971)[566], and μ (ref. 0118)[567] receptors were determined by measuring scintillation of suitable reference ligands upon 60 or 120 minutes of incubation.

3.5.4 Results and Discussion

The molecular design method requires (i) a compound synthesis scheme, (ii) an affinity prediction method for the virtual products, and (iii) a technique for building block optimization. For our concept study, we chose the reductive amination reaction working with aldehydes/ketones and amines as building blocks. We applied MAntA to single-step reductive amination products accessible from commercially available building blocks. Its reaction products have a high likelihood of possessing desirable druglike features, as visualized in Figure 27, which presents a map of the known bioactivity space. Virtual reaction products cluster in a densely populated area, and the reductive amination may be regarded as a preferred reaction for drug discovery.

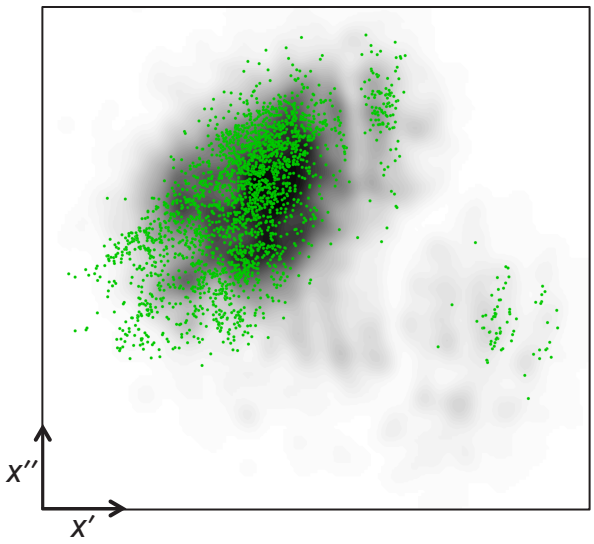


Figure 27. The distribution of 5000 virtual reductive amination products (green dots) in a druglike chemical space. The two-dimensional landscape was calculated from the density of 10,000 drug-like molecules sampled from the ChEMBL database. The intensity of grey color indicates the density of known bioactive substances (white: sparsely populated; black: highest local density). The compounds were represented by topological pharmacophores ("CATS2" descriptor)[486] and projected to the plane (*x*', *x*'') by stochastic neighbor embedding (SNE), which led to a local-neighborhood preserving map of chemical space. The axes represent nonlinear combinations of the original molecular descriptors. The image was generated with LiSARD[478].

For affinity prediction we trained individual *Gaussian Process* (GP) regression models[171] for 640 human targets annotated in ChEMBL (v14)[72], based on 279,866 compounds with 569,725 measured bioactivities. Molecules were represented by topological pharmacophore ("CATS2")[486] and substructure (circular Morgan fingerprints)[36] descriptors. The choice of GP regression was motivated by extensive comparison to other modeling techniques using the same training data, where the GP approach performed best (Supplementary Table S4-S5). In addition, GP models compute a data density dependent confidence estimate, which we combined with the quantitative bioactivity prediction (p*Affinity*) to obtain a single robust prediction score for each compound.

Equipped with this quantitative affinity prediction model, MAntA performs an adaptive search for optimal building block combinations for the given reaction scheme (Figure 28). The search space consists of all possible educts labeled with *pseudo*-probabilities ("pheromones"), according to their contributions to the computed predictive score. Individual ants traverse the search space following pheromone trails and assemble virtual products. These are scored and the *pseudo*-probabilities on their respective molecular building blocks are adjusted accordingly.

Over simulation time, high scoring building block combinations emerge from the ant colony's optimal path-finding capability.

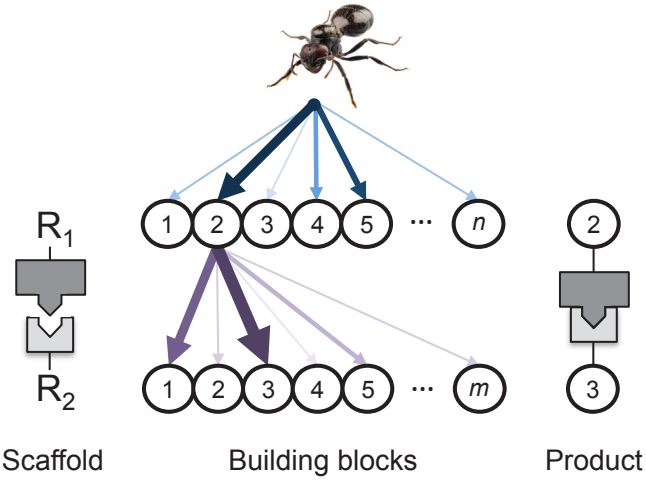


Figure 28. Molecular building block selection by combinatorial ant colony optimization (MAntA). The arrows represent artificial ant paths for this two-component combinatorial library. Their widths correspond to *pseudo*-probabilities ("pheromone concentrations") that influence the choice made by the ants and thereby determine the actual product spectrum. The pheromone concentrations are adaptive and subject to evaporation.

We employed MAntA for the multi-objective design of novel ligands for high-profile macromolecular drug targets that are involved in neuropsychiatric disorders – sigma-1 and dopamine D_4 receptors. The choice of D_4 receptor was also made to allow for a direct comparison to a recent publication by Hopkins and coworkers[100]. In our study, the task was to select a small number of preferred products from a total of approximately 20 million. First, we discarded all designed molecules with undesired structural motifs[568], and poor predicted absorption, distribution, metabolism and excretion ("negative design")[569]. From the 3529 remaining molecules we selected candidates having different aims in mind ("positive design"):

i) Potent and selective (sigma-1) or multi-target modulating (dopamine D_4) ligands;
ii) Target subtype-selective ligands;
iii) Exploratory molecules, lying outside the training domain as expressed by Morgan fingerprint Tanimoto similarities < 0.20;
iv) Inactive compounds as nearest neighbors to known high-affinity ligands in ChEMBL bioactivity space.

For the sigma-1 receptor, we selected compounds **17**–**19** according to the high affinity criterion. Molecules **20**–**21** were designed as receptor-selective ligands (Figure 29a). In fact, the MAntA designs were experimentally validated for their specific goals with accurately predicted pK_i values (Table 7). Compounds **17**–**19** exhibited K_i values of 1.1–2.2 nM, and designs **20**–**21** yielded more than 2500-fold selectivity over the δ, κ, and μ opioid receptors. Noteworthy, **17**–**20** are equipotent to their nearest neighbor counterparts from ChEMBL, despite being structurally dissimilar (Tanimoto similarity ~ 0.45), thereby endorsing the exploratory potential of MAntA. Furthermore, the low molecular weight of **17**–**21**, coupled to low nanomolar K_i values warrant these compounds high ligand efficiency. Additionally, we synthesized and tested compounds **22**–**24** as scaffold hops from known ChEMBL chemical space (structural Tanimoto similarity to nearest neighbors ~ 0.20), without critical loss of affinity (sigma-1 K_i = 10–210 nM, ΔpK_i ~ 0.5, Table 7, Figure 29b) with exception of compound **24**. Furthermore, compounds **17-24** contain scaffolds that were not present in the training data used for model building (Supplementary Table S6). Apparently, the low structural resemblance to known small molecules did not considerably affect the algorithm's performance. Finally, compound **25** was designed and validated as a low affinity sigma-1 ligand (K_i > 2,500 nM) despite having a highly potent nearest neighbor (K_i ~ 6 nM, Tanimoto similarity = 0.45, Table 7), pinpointing the adaptive design capabilities of MAntA that go beyond structural similarity analysis. Altogether, the experimental results are in agreement with the landscape projection of the preferred sigma-1 activity islands (Figure 30a; individual target landscapes are shown in Supplementary Figure S17). Furthermore, as an *off*-target for the synthesized compounds, MAntA predicted moderate histamine H_3 receptor affinities, which were partly confirmed experimentally.

Next, we designed antagonists for the dopamine D_4 receptor. D_4 receptors are especially implicated in attention deficit hyperactivity disorder, mood disorders, and Parkinson's disease, among other neuropsychiatric illnesses[570]. From the top 1600 prioritized small molecules with predicted pK_i > 7 for the D_4 receptor, we selected compounds **26** and **27** as high-affinity ligands. Although **26** (K_i = 2.0 nM) features an already known scaffold[571], **27** represents a notably different, and more ligand efficient entity than its ChEMBL nearest neighbor (Tanimoto similarity ~ 0.6, Table 7). While the selected ligands were primarily designed as high-affinity D_4 receptor

antagonists, a polypharmacology profile was not precluded. Accordingly, promiscuous dopamine $D_{1\text{-}5}$ and 5-HT_{1A} receptor binding was predicted. Subsequent binding tests confirmed the multi-target modulating profiles of **26** and **27** in agreement with the MAntA-predicted bioactivity spectra and landscape projections (Table 7, Figure 29b, preferred design zones Figure 30b; individual target landscapes are shown in Supplementary Figure S18). Conversely, compounds **28** and **29** were designed to meet the D_4 receptor selectivity goal. Selective D_4 antagonists are equally relevant in clinics, as they can prevent stress-induced cognitive dysfunction without extrapyramidal motor symptoms or neuroendocrine side effects[571]. Weak binding affinities of **28** and **29** were predicted for the *off*-targets in the assay panel. Structural simplicity and low nanomolar affinities (K_i = 10−12 nM) designate these compounds. Their selectivity for the D_4 receptor is particularly significant, given the high structural similarity to promiscuous molecules **26** and **27**. Of note, 1,4-disubstituted aromatic piperazines have previously been recognized as predominant in promiscuous biogenic amine G-protein coupled receptor (GPCR) ligands[572]. The opposing target engagement profiles for the arylpiperazines **26**–**27** and **28**–**29** confirm effective building-block selections. The polypharmacology profile of **30** and **31**, which extend the known chemical diversity of D_4 receptor antagonists, is also in agreement with the pK_i predictions. Remarkably, **30** is one log unit more potent against the D_4 receptor than the closest related reference antagonist, which together with the screening results of the designed inactive **32**, demonstrates the successful application of MAntA to dopamine receptors. Evidently, considerably extended experimental GPCR panel activities will be required for further hit-to-lead progression of the MAntA designs.

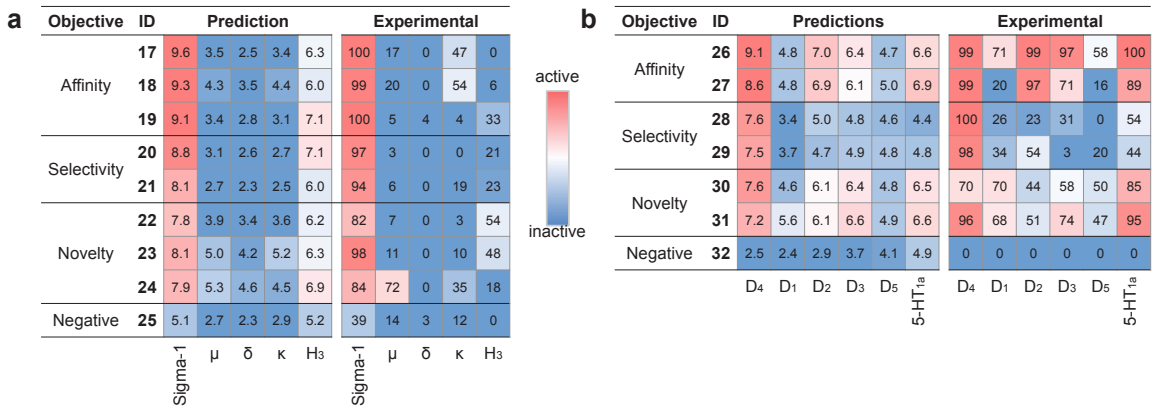


Figure 29. Bioactivity profiles of designed ligands. Comparison of the Gaussian process prediction for sigma-1 (**a**) and dopamine D_4 (**b**) receptors, together with the observed experimental results. Predictions are variance-corrected p*Affinity* values. Experimental data are expressed as % inhibition (competitive radioligand binding assays at 2.5 µM). Colors are linearly interpolated from the predicted p*Affinity* intervals [4, 9] (**a**) and [4, 8] (**b**), and % inhibition interval [20, 100] for the experimental data.

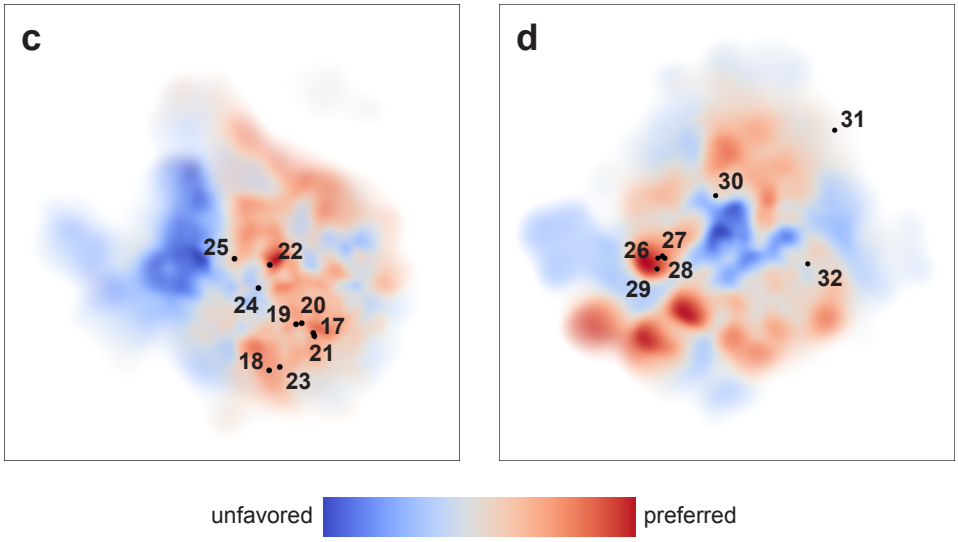
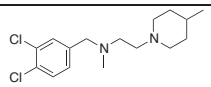
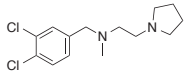
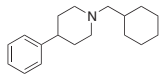
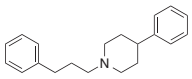
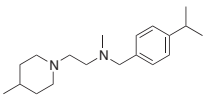
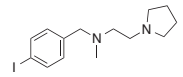
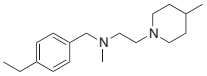
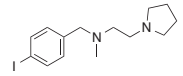
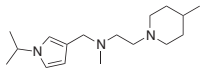
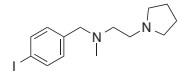
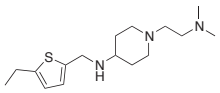
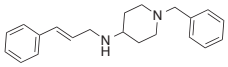
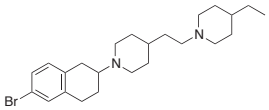
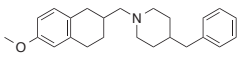
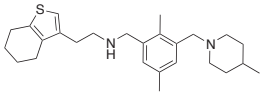
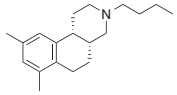
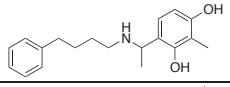
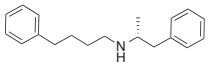
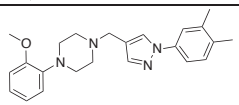
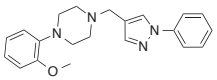
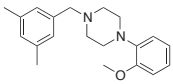
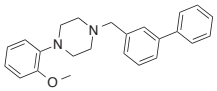
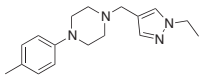
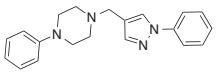
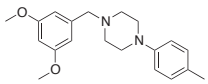
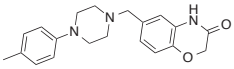
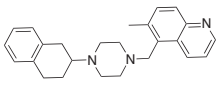
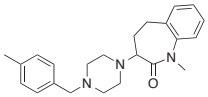
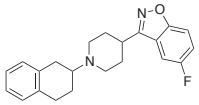
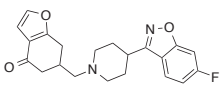
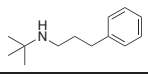
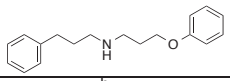


Figure 30. LiSARD multi-target selectivity landscapes. Sigma-1 receptor (**a**) and dopamine D_4 receptor (**b**) as the respective *on*-target.

3.5.5 Conclusion

With regard to the polygenic nature of most major central nervous system diseases and the individual variability of their genetic basis, new drugs with selected polypharmacological activities are desirable[573]. The results of this study suggest a feasible solution for the combinatorial design of new chemical entities with affinity profiles and properties that exceed the average drug-likeness for approved drugs (*Quantitative Estimate of Drug-likeness*, *QED* = 0.72 ± 0.10 *vs*. 0.49)[19,547]. The automated molecular design method should be broadly applicable to other drug target classes and chemistry, provided reliable structure-activity data are available for constructing predictive affinity prediction models. The actual computational design process is fast (within minutes on a desktop computer), so that focused combinatorial library design and synthesis can be realized within a day of work. A particular advantage of MAntA compared to many other approaches, *e.g*. the meticulous work of Besnard *et al*. on adaptive drug design[100], lies in the simultaneous generation of both potent structural analogs and innovative scaffold-hops from known reference compounds. Together with rapid computation, low-cost synthesis, and readily accessible chemical structures, the concept of adaptive building block and fragment prioritization might become widely applicable.

Table 7. The designed molecules and their nearest neighbors from the ChEMBL training data with the predicted and experimentally determined binding affinities of compounds **17**-**32**.

	MAntA designs				Nearest neighbors (training data)			
ID	**Structure**	**Pred. p*Affin.***	**Exp. pK_i**	***LE***[a]	**Structure**	**ChEMBL ID**[b]	**Struct. Sim.**[c]	**pK_i**
17		9.7	9.0	0.63		143089	0.70	9.4
18		9.4	8.9	0.65		112124	0.44	9.0
19		9.3	8.7	0.58		154397	0.44	8.7
20		9.0	8.8	0.61		154397	0.47	8.7
21		8.1	7.9	0.55		154397	0.34	8.7
22		7.9	7.2	0.50		111909	0.21	7.8
23		8.1	8.1	0.42		179530	0.21	7.7
24		8.0	6.7	0.32		544748	0.20	8.7[d]
25		4.1	*n.d.*	*n.d.*		20976	0.45	8.2
26		9.4	8.7	0.44		379602	0.70	9.6
27		9.0	8.3	0.50		285577	0.57	8.1
28		7.9	8.0	0.53		210405	0.41	9.0
29		7.8	7.9	0.46		345552	0.43	8.4
30		7.9	6.6	0.33		305061	0.23	5.4
31		7.3	7.6	0.41		143027	0.27	7.7
32		2.5	*n.d.*	*n.d.*		129931	0.40	8.5

[a] Ligand efficiency (*LE* = -1.4 × pK_i / number of heavy atoms); [b] ChEMBL IDs are quoted without the "CHEMBL" prefix; [c] Tanimoto similarity index (Morgan fingerprints with radius = 3); [d] pIC_{50} value. *n.d.*: not determined.

3.5.6 Publication details and contributions

Authors

Michael Reutlinger,[a] Tiago Rodrigues,[a] Petra Schneider,[a] Gisbert Schneider [a]

[a] ETH Zurich, Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

Author contributions

M.R. designed and conducted the research, developed the *de novo* design software, performed the synthesis, analyzed the results, and contributed the manuscript.

Acknowledgements

We thank Dr. Michael Bieler for computing QED values.

Reference

Reutlinger R, Rodrigues T, Schneider P, Schneider G (2014) Multi-objective molecular *de novo* design by fragment prioritization. *Angew. Chem. Int. Ed.,* doi: 10.1002/anie.201310864.

Licence

Source of funding

This research was financially supported by a research grant from the OPO Foundation, Zurich.

4 Conclusions and Outlook

This thesis presents a holistic approach for the *de novo* design of multi-target modulating compounds (MAntA). It combines a visualization tool for intuitive navigation in chemical space, an ant colony optimization method, designing synthetically accessible compounds, and machine-learning models for multi-target affinity prediction and small-molecule prioritization. The modular components of the approach were individually evaluated for their potential use in early drug discovery. All tools were collectively applied in a prospective study to identify innovative, potent, and ligand-efficient multi-target modulating compounds that can be further developed through medicinal chemistry optimization. The results presented herein demonstrate the capabilities of MAntA for generating new chemical entities with accurately predicted bioactivities for a panel of relevant drug targets.

The visualization component (LiSARD) included in the MAntA process utilizes stochastic neighbor embedding (SNE)[382] for dimensionality reduction of mathematically represented chemical datasets. SNE aims at preserving the local neighborhood observed in the original high-dimensional data space. To quantify the potential benefits of this concept, a panel of quality indices for local neighborhood preservation was calculated for three different compound libraries. SNE performed favorably in comparison to PCA[367], MDS[368] and SPE[381] (*cf*. Table 3). Several recently proposed, nonlinear dimension reduction methods (e.g. Isomap[574], local linear embedding[385], Laplacian eigenmaps[575]) utilize neighborhood graphs to assess the local neighborhood and model the nonlinear geometry of the low-dimensional manifold in close proximity to the observed data. In contrast to SNE, these methods cannot guarantee that all data points are included in the local embedding. Is is of note that SNE has a runtime and memory complexity of $\mathcal{O}(N^2)$ that usually restricts it to datasets of up to 15,000 data points. The computational complexity did not adversely affect the studies conducted within the scope of the thesis, but there are several scenarios (e.g. analyzing HTS results or large corporate databases) that require the ability to handle larger datasets. In these scenarios, the recently proposed Barnes-Hut-SNE algorithm could be employed[576], as it reduces the runtime and memory complexity to $\mathcal{O}(N \log N)$ and $\mathcal{O}(N)$, respectively, which

makes it applicable to learning embeddings of datasets containing millions of data points.

Scattered point distributions produced by dimensionality reduction can be enriched with additional drug-relevant properties. Based on the enriched data, LiSARD calculates a three-dimensional response surface on the fly to obtain an visually accessible map of chemical space. Potential benefits of the LiSARD approach for drug discovery efforts were investigated using a comprehensive dataset obtained from a chemogenomics study performed at Roche Pharmaceuticals, Basel. This study aimed at finding non-peptidic hSST5R antagonists (*cf*. Chapter 2.1). Even though the study was retrospective in nature, the analysis of data evolution over time demonstrated the potential application at several project stages. Importantly, trends in the bioactive space of hSST5R activity were observable already within the first 100 synthesized compounds (Figure 12d). Landscapes generated from subsequent compound generations still resembled the early trend but also revealed further details. This result clearly underlines the advantage of an early visual inspection of chemical space to steer lead optimization. Thus, LiSARD may be regarded as adequate to identify undesired *tabu*-areas, and focus on regions of chemical space containing compounds with desired properties. The bioactivity-focused landscapes are calculated using whole-molecule descriptors, and visualize the global relation between compounds in chemical space. Yet they lack an apparent interpretation regarding the contributions of individual functional groups. Therefore, they could be complemented with methods providing a visualization of feature importance. For example, Hansen *et al*. recently introduced an suitable method that might also be suitable for the machine-learning approach used in this work[577]. While landscape visualization may not provide the required accuracy for prioritizing individual compounds, it often offers a useful perception of chemical space and can aid medicinal chemists in discovering global multi-objective trends in complex structure-activity relationship data[15].

For prioritizing candidate compounds, MAntA applies a quantitative machine-learning method to polypharmacology-focused *de novo* molecule design. A ligand-based strategy was implemented to provide a predictive multi-target framework. It was successfully used to generate innovative compounds fitting the desired target profile, and identifying macromolecular targets for several members of an Ugi-3CR-driven GPCR focused library.

The predictive models were trained using publicly available data obtained from the ChEMBL database[72]. The raw data required substantial pre-processing and data cleaning to be useful for model building. Annotation errors, e.g. wrongly assigned affinity units, were frequently encountered. For the targets included in the prospective design study, the manual inspection of assay data revealed several quality issues including affinities measured for proteins with mutations in the binding site, which we consequently excluded from the training data. The results presented in this thesis indicate that contemporary machine-learning methods are able to handle such errors in the data, provided that sufficient high-quality data is available for model building. For example, the employed Gaussian process regression explicitly learns a noise parameter, which accounts for data errors and the inevitable experimental uncertainty. We found that the observed accuracy of the prospective results is in agreement with theoretic considerations on the experimental uncertainty of heterogeneous assay data[460]. A desirable feature of a predictive model would be the possibility to predict the functional effect in addition to the binding affinity. However, the lack of systematic functional annotation in current chemogenomics databases prevents an immediate inclusion, and further text-mining will be required to extract the required information from unstructured text data.

MAntA utilizes a quantitative nonlinear machine-learning method to evaluate *de novo* designed compounds. This offers the potential advantage of ranking the designed compounds according to estimated binding affinities, which is a desirable property compared to related approaches that only rank according to the probability of compounds being active. The prospective studies included in this thesis confirm the applicability to *de novo* small-molecule design, with close agreement between predicted and experimentally measured binding affinities. In a similar study, Besnard *et al*. employed naïve Bayes classifiers, a qualitative machine-learning approach, to a related set of aminergic GPCRs[100]. The success rates reported by Besnard *et al*. are comparable to the ones reported for MAntA in Chapter 3.5.4. However, the design process implemented in MAntA is not restricted to transformations of a given template or scaffold motifs but rather creates synthetically feasible molecules from scratch, which should be beneficial for exploration of new chemical space and the identification of innovative compounds.

GPCRs have been extensively explored as drug targets in pharmaceutical research[348,578,579], which is also reflected by the fact that over 25% of currently marketed drugs address GPCR targets[580]. Consequently, a considerable amount of ligand binding affinity data is available for GPCRs, including the ones investigated in scope of this thesis. Our retrospective analysis of 640 targets (Table S4, Figure S12-S13) showed that for over 90% of the targets a predictive model could be built. Additionally, the results revealed an overall positive influence of additional training data on the screening performance, which might be a reason for the exceptional performance of the approach in the prospective studies. Further exploratory work including targets with a limited amount of data will help to fully comprehend the potential and also limitations of the proposed method. For biological targets where there is limited knowledge on active ligands, alternative *de novo* design approaches might be more suitable, e.g. utilizing the software DOGS, which requires only a single known active as template to propose new compounds[273].

In order to broaden the scope of MAntA, it might be beneficial to combine several complementary scoring schemes. Each scoring method should be tailored to address a specific range of available information. The appropriate method for the investigated target could then be chosen automatically, either by the number of available training data or by retrospective performance estimates. A closely related approach was introduced recently for computational target fishing by Rognan and coworkers[581]. Another opportunity might also be to consider ligands from closely related targets by incorporating a target similarity metric into the prediction framework. The similarity could be calculated on the basis of protein sequence, as suggested in a recent publication[582], or by relating the targets by binding pocket similarity. With the renaissance of phenotypic screening as a method to identify drug candidates[330], ligand-based machine-learning methods might also be applied to the raw phenotypic readout without explicit knowledge of the ligand-target interaction in order to identify promising drug candidates.

The current MAntA implementation is exclusively ligand-based and does not incorporate structural information about the macromolecular target. For the GPCR target family investigated in this work, there are currently only few crystal structures available. With recent progress in solving GPCR structures, the number of available

crystal structures is likely to increase[583]. Therefore, MAntA could be combined with receptor-based methods, especially when only a limited number of known ligands is available. Such a combination could either be sequential, parallel, or made by directly including receptor-information in the scoring process[193]. However, it has been recognized that ligand-based methods often outperform contemporary receptor-based approaches in regard to hit retrieval[44,584,585], and in the majority of cases the combination of ligand- and receptor-based methods does not improve the performance[584]. Reasons for this observation could be that current ligand-receptor docking methods are able to identify the receptor-relevant binding pose with acceptable accuracy but are unable to accurately rank compounds according to their binding affinity[210,233,586]. The problems mainly stem from the approximative nature of the currently utilized scoring functions[587]. Several problematic areas have been identified that will probably be addressed in the future[208]. A major concern is the inherent protein flexibility, which affects the spatial arrangement of residues in the ligand binding site, and consequently the ligand binding mode[588]. Proteins are often found to exist as a heterogeneous ensemble of conformations without a single preferred bioactive conformation[588]. It has recently been realized that molecular ligand-receptor recognition is a "conformational selection" process where the ligand selects the most favorable conformation out of the pre-existing conformational ensemble upon binding[589-592], rather than being an "induced fit" effect as proposed by Koshland in 1958[593]. How to accurately model the target flexibility in a scoring function is still an unsolved question[208,588,594]. A second issue is the inadequate treatment of entropic effects[210]. Current scoring functions mostly govern the enthalpic contributions to binding affinity and neglect entropic contributions[210]. A related issue which requires further attention in scoring function development is the treatment of water in the binding pocket[595]. Despite these drawbacks, receptor-based methods have the advantage that they do not require any *a priori* ligand data. Therefore, they can be utilized to find ligands for orphan targets[587].

Finally, it should be mentioned that it cannot be expected for a single method to perform equally well for all targets. Rigorous evaluation of the methods' domain of applicability is indispensable for picking the appropriate method(s) for the investigated target(s).

5 Acknowledgments

First and foremost, I would like to thank Prof. Dr. Gisbert Schneider for giving me the opportunity to conduct this fascinating research within his excellent group at ETH Zurich. I am exceedingly grateful for his invaluable guidance during the time of my Ph.D., and continuous scientific and personal support. His highly positive attitude towards science and life in general is inspiring. I would also like to explicitly thank Prof. Dr. Gerd Folkers for taking the time to act as co-referee for my thesis. Further thanks go to our collaborators at Hoffmann-La Roche and Boehringer Ingelheim. In particular Dr. Wolfgang Guba for supporting me in the development of LiSARD, and Michael Bieler and Dr. Jan Kriegl for our fruitful collaboration on multi-target prediction. Many enjoyable meetings with insightful discussions contributed to the final outcome of this work.

It goes without saying that I am grateful to my colleagues in the wonderful Schneider group at ETH Zurich that I deeply enjoyed working with. We had lots of fun, but at the same time shared scientifically challenging and also very fruitful experiences. In particular, I would like to mention Dr. Tiago Rodrigues for his supervision in the lab and support in most of my projects, Christian Koch, Anna Perna, and Daniel Reker for numerous vivid discussions (not only scientifically), Jens Kunze for taking care of my fitness, Dr. Jan Hiss for his work on artificial ants and consultation regarding my projects, and Dr. Petra Schneider for always having an open ear for me and our shared interests. Additionally, I would like to thank Chrissula Chatzidis, Dr. Tim Geppert, Sarah Haller, Dr. Markus Hartenfeller, Dr. Johannes Kirchmair, Yen-Chu "Jimmy" Lin, Max Pillong, Dr. Felix Reisen, Katharina Stutz, and Nickolay Todoroff.

Finally, I would like to thank my family for always believing in me and supporting me unconditionally. Thank you also to my two "girls" Valeria and Varinka for their emotional support. My deepest gratitude goes to my beloved Claudia, who encouraged me in my decision of doing this Ph.D. and always unconditionally supported me. This work would not have been possible without you.

6 Curriculum Vitae

Michael R. Reutlinger
Dipl.-Bioinformatiker

Date of birth: February 7th, 1978

Place of birth: Kronberg, Germany

EDUCATION

01/2011 – 03/2014 **Swiss Federal Institute of Technology (ETH), Zurich, Switzerland**
PhD in Computational Chemistry and Drug Design (Dr. sc. ETH)

Computer-Assisted Drug Design Group, Prof. Dr. Gisbert Schneider

PhD project: Development of a nature-inspired computational design approach for advanced poly-pharmacology based combinatorial drug design, synthesis of designed compounds and experimental validation of poly-pharmacology predictions

10/2005 – 11/2010 **Goethe University, Frankfurt am Main, Germany**
Graduated as "Diplom-Bioinformatiker" (with Honors)

Thesis: "Modellierung von Struktur-Aktivitätsbeziehungen mit selbstorganisierenden Projektionen" in collaboration with F. Hoffmann-La Roche Ltd. (Dr. Wolfgang Guba)

1988 – 1997 **Anna-Schmidt-Schule, Frankfurt am Main, Germany**

School diploma: University entrance diploma (Abitur)

PROFESSIONAL EXPERIENCE

Since 02/2014 **F. Hoffmann-La Roche Ltd., Basel, Switzerland**
Scientist in Computer-Aided Drug Design

11/2008 – 01/2009 **F. Hoffmann-La Roche Ltd., Basel, Switzerland**
Research internship in the Molecular Design Group (Dr. Bernd Kuhn)

01/1999 – 09/2005 **arago Institut für komplexes Datenmanagement AG, Frankfurt am Main, Germany**
Head of Products and Projects

AWARDS

11/2013 Poster Award GCC 2013
Title: "Go with the flow: De-orphaning of focused combinatorial libraries"

11/2008 – 10/2010 MainCampus academicus scholarship,
Stiftung Polytechnische Gesellschaft Frankfurt am Main, Germany

PUBLICATIONS

Reutlinger M, Rodrigues T, Schneider P, Schneider G (2014) Multi-objective molecular de novo design by adaptive fragment prioritization. *Angew. Chem. Int. Ed.*, doi: 10.1002/anie.201310864.

Reutlinger M, Rodrigues T, Schneider P, Schneider G (2014) Combining on-chip synthesis of a focused combinatorial library with in silico target prediction reveals imidazopyridine GPCR ligands. *Angew. Chem. Int. Ed.* **53**, 582-585.

Rupp M, Bauer MR, Wilcken R, Lange A, Reutlinger M, Boeckler FM, Schneider G (2014) Machine learning estimates of natural product conformational energies. *PLoS Comput. Biol.* **10**, e1003400.

Koch CP, Reutlinger M, Todoroff N, Schneider P, Schneider G (2014) *CATS for scaffold-hopping in medicinal chemistry.* In: *Scaffold-hopping in medicinal chemistry* (N. Brown, ed.), Wiley-VCH, Weinheim, 119-130.

Hiss JA, Reutlinger M, Koch CP, Perna AM, Schneider P, Rodrigues T, Haller S, Folkers G, Weber L, Baleeiro RB, Walden P, Wrede P, Schneider G (2014) Combinatorial chemistry by ant colony optimization. *Future Med. Chem.* **6**, 267-280.

Spänkuch B, Keppner S, Lange L, Rodrigues T, Zettl H, Koch CP, Reutlinger M, Hartenfeller M, Schneider P, Schneider G (2013) Drugs by numbers: Reaction-driven de novo design of potent and selective anticancer leads. *Angew. Chem. Int. Ed.* **52**, 4676-4681.

Koch CP, Perna AM, Weissmüller S, Bauer S, Pillong M, Baleeiro RB, Reutlinger M, Folkers G, Walden P, Wrede P, Hiss JA, Waibler Z, Schneider G (2013) Exhaustive proteome mining for functional MHC-I ligands. *ACS Chem. Biol.* **8**, 1876-1881.

Reutlinger M, Koch CP, Reker D, Todoroff N, Schneider P, Rodrigues T, Schneider G (2013) Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for 'orphan' molecules. *Mol. Inf.* **32**, 133-138.

Schneider G, Lin Y-C, Koch CP, Pillong M, Perna AM, Reutlinger M, Hiss JA (2013) Adaptive peptide design. *Chimia* **67**, 859-863.

Geppert T, Bauer S, Hiss JA, Conrad E, Reutlinger M, Schneider P, Weisel M, Pfeiffer B, Altmann KH, Waibler Z, Schneider G (2012) Immunosuppressive small molecule discovered by structure-based virtual screening for protein-protein interaction inhibitors. *Angew. Chem. Int. Ed.* **51**, 258-261.

Reutlinger M, Schneider G (2012) Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J. Mol. Graphics Modell.* **34**, 108-117.

Reutlinger M, Guba W, Martin RE, Alanine AI, Hoffmann T, Klenner A, Hiss JA, Schneider P, Schneider G (2011) Neighborhood-preserving visualization of adaptive structure-activity landscapes and application to drug discovery. *Angew. Chem. Int. Ed.* **50**, 11633-11636.

Schneider G, Geppert T, Hartenfeller M, Reisen F, Klenner A, Reutlinger M, Hähnke V, Hiss JA, Zettl H, Keppner S, Spänkuch S, Schneider P (2011) Reaction-driven de novo design, synthesis and testing of potential type II kinase inhibitors. *Future Med. Chem.* **3**, 415-424.

Schneider P, Stutz K, Kasper L, Haller S, Reutlinger M, Reisen F, Geppert T, Schneider G (2011) Target profile prediction for a Biginelli-type dihydropyrimidine compound library and practical evaluation. *Pharmaceuticals* **4**, 1236-1247.

Kuhn B, Fuchs J, Reutlinger M, Stahl M, Taylor NR (2011) Rationalizing Tight Ligand Binding Through Cooperative Networks. *J. Chem. Inf. Model.* **51**, 3180–3198.

Schneider G, Hartenfeller M, Reutlinger M, Tanrikulu Y, Proschak E, Schneider P (2008) Voyages to the (un)known: Adaptive design of bioactive compounds. *Trends Biotechnol.* **27**, 18-26.

7 References

1. Hartenfeller M, Schneider G (2011) Enabling future drug discovery by *de novo* design. *WIREs Comput. Mol. Sci.* **1**, 742-759.
2. Hare R (1982) New light on the history of penicillin. *Med. Hist.* **26**, 1-24.
3. Keseru GM, Makara GM (2006) Hit discovery and hit-to-lead approaches. *Drug Discov. Today* **11**, 741-748.
4. Roth BL, Sheffler DJ, Kroeze WK (2004) Magic shotguns versus magic bullets: Selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3**, 353-359.
5. Morphy R, Kay C, Rankovic Z (2004) From magic bullets to designed multiple ligands. *Drug Discov. Today* **9**, 641-651.
6. Scannell JW, Blanckley A, Boldon H, Warrington B (2012) Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191-200.
7. Schneider G, Böhm H-J (2002) Virtual screening and fast automated docking methods. *Drug Discov. Today* **7**, 64-70.
8. Schneider G (2012) Designing the molecular future. *J. Comput. Aided Mol. Des.* **26**, 115-120.
9. Schneider G, Baringhaus K-H (2008) *Molecular design: Concepts and applications.* Wiley-VCH, Weinheim.
10. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **11**, 1046-1053.
11. Schneider G (2010) Virtual screening: An endless staircase? *Nat. Rev. Drug Discov.* **9**, 273-276.
12. Kubinyi H, Mannhold R, Timmerman H, Böhm H-J, Schneider G (2008) *Virtual screening for bioactive molecules.* Wiley & Sons, New York.
13. Klebe G (2000) *Virtual screening: An alternative or complement to high throughput screening?* Springer, Dordrecht.
14. Schneider G, Neidhart W, Giller T, Schmid G (1999) "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem. Int. Ed.* **38**, 2894-2896.
15. Schneider G, Baringhaus K-H (2013) *De novo design: From models to molecules.* In: *De Novo Molecular Design* (ed. Schneider G) Wiley-VCH, Weinheim, 1-56.
16. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* **23**, 3-25.
17. Congreve M, Carr R, Murray C, Jhoti H (2003) A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today* **8**, 876-877.
18. Wunberg T, Hendrix M, Hillisch A, Lobell M, Meier H, Schmeck C, Wild H, Hinzen B (2006) Improving the hit-to-lead process: Data-driven assessment of drug-like and lead-like screening hits. *Drug Discov. Today* **11**, 175-180.
19. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90-98.
20. Walters WP, Namchuk M (2003) Designing screens: How to make your hits a hit. *Nat. Rev. Drug Discov.* **2**, 259-266.
21. Roche O, Schneider P, Zuegge J, Guba W, Kansy M, Alanine A, Bleicher K, Danel F, Gutknecht E-M, Rogers-Evans M, Neidhart W, Stalder H, Dillon M, Sjögren E, Fotouhi N, Gillespie P, Goodnow R, Harris W, Jones P, Taniguchi M, Tsujii S, von dS, Wolfgang, Zimmermann G, Schneider G (2002) Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J. Med. Chem.* **45**, 137-142.

22. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**, 2719-2740.
23. Rechenberg I (1973) *Evolutionsstrategie - Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.* Frommann-Holzboog, Stuttgart.
24. Johnson MA, Maggiora GM (1990) *Concepts and applications of molecular similarity.* Wiley, New York.
25. Martin YC, Kofron JL, Traphagen LM (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **45**, 4350-4358.
26. Maggiora GM (2006) On outliers and activity cliffs--why QSAR often disappoints. *J. Chem. Inf. Model.* **46**, 1535.
27. Xue L, Godden JW, Bajorath J (2000) Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **40**, 1227-1234.
28. Medina-Franco JL, Martinez-Mayorga K, Giulianotti MA, Houghten RA, Pinilla C (2008) Visualization of the chemical space in drug discovery. *Curr. Comput.-Aided Drug Des.* **4**, 322-333.
29. Todeschini R, Consonni V (2008) *Handbook of molecular descriptors.* Wiley, New York.
30. Rognan D (2007) Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **152**, 38-52.
31. Wiener H (1947) Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **69**, 17-20.
32. Kier LB (1976) *Molecular connectivity in chemistry and drug research.* Academic Press, New York.
33. Kier LB, Hall LH (1986) *Molecular connectivity in structure-activity analysis.* Research Studies Press, Letchworth.
34. Moreau G, Broto P (1980) The autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.* **4**, 359-360.
35. Morgan HL (1965) The generation of a unique machine description for chemical structures — A technique developed at chemical abstracts service. *J. Chem. Doc.* **5**, 107-113.
36. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742-754.
37. Schneider P, Schneider G (2003) Collection of bioactive reference compounds for focused library design. *QSAR Comb. Sci.* **22**, 713-718.
38. McGregor MJ, Muskal SM (1999) Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **39**, 569-574.
39. Mason JS, Morize I, Menard PR, Cheney DL, Hulme C, Labaudiniere RF (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **42**, 3251-3264.
40. Hemmer MC, Steinhauer V, Gasteiger J (1999) Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectroscopy* **19**, 151-164.
41. Tanrikulu Y, Nietert M, Scheffer U, Proschak E, Grabowski K, Schneider P, Weidlich M, Karas M, Gobel M, Schneider G (2007) Scaffold hopping by “fuzzy” pharmacophores and its application to RNA targets. *ChemBioChem* **8**, 1932-1936.
42. Fechner U, Franke L, Renner S, Schneider P, Schneider G (2003) Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput. Aided Mol. Des.* **17**, 687-698.
43. Hristozov DP, Oprea TI, Gasteiger J (2007) Virtual screening applications: A study of ligand-based methods and different structure representations in four different scenarios. *J. Comput. Aided Mol. Des.* **21**, 617-640.

44. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell WD (2007) Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **47**, 1504-1519.
45. Leach AR, Gillet VJ (2007) *Similarity Methods.* In: *An Introduction To Chemoinformatics* Springer Netherlands, Dordrecht, 99-117.
46. Fechner U, Schneider G (2004) Evaluation of distance metrics for ligand-based similarity searching. *ChemBioChem* **5**, 538-540.
47. Xu H, Agrafiotis D (2002) Retrospect and prospect of virtual screening in drug discovery. *Curr. Trends Med. Chem.* **2**, 1305-1320.
48. Whittle M, Gillet VJ, Willett P, Alex A, Loesel J (2004) Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: A comparison of similarity coefficients. *J. Chem. Inf. Comput. Sci.* **44**, 1840-1848.
49. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **44**, 1177-1185.
50. Sastry M, Lowrie JF, Dixon SL, Sherman W (2010) Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **50**, 771-784.
51. Nicholls A (2008) What do we know and when do we know it? *J. Comput. Aided Mol. Des.* **22**, 239-255.
52. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* **29**, 476-488.
53. Truchon J-F, Bayly CI (2007) Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **47**, 488-508.
54. Hawkins PC, Warren GL, Skillman AG, Nicholls A (2008) How to do an evaluation: Pitfalls and traps. *J. Comput. Aided Mol. Des.* **22**, 179-190.
55. Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO (2005) Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **48**, 2534-2547.
56. Hanely JA, McNeil BJ (1982) The meaning and use of the area under a *Receiver Operating Characteristic* (ROC) curve. *Radiology* **143**, 29-36.
57. Metz CE (1978) Basic principles of ROC analysis. *Semin. Nucl. Med.* **8**, 283-298.
58. Walters WP, Stahl MT, Murcko MA (1998) Virtual screening — an overview. *Drug Discov. Today* **3**, 160-178.
59. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **25**, 64-73.
60. Willett P, Winterman V, Bawden D (1986) Implementation of nearest-neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* **26**, 36-41.
61. Vogt M, Bajorath J (2011) Predicting the performance of fingerprint similarity searching. *Methods Mol. Biol.* **672**, 159-173.
62. Schneider G, Schneider P, Renner S (2006) Scaffold-hopping: How far can you jump? *QSAR Comb. Sci.* **25**, 1162-1171.
63. Clark DE (2008) What has virtual screening ever done for drug discovery? *Expert Opin. Drug Discov.* **3**, 841-851.
64. Kearsley SK, Sallamack S, Fluder EM, Andose JD, Mosley RT, Sheridan RP (1996) Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **36**, 118-127.
65. Ginn CMR, Willett P, Bradshaw J (2002) *Combination of molecular similarity measures using data fusion.* In: *Virtual Screening: An Alternative or Complement to High Throughput Screening?* Springer, New York, 1-16.

66. Wang R, Wang S (2001) How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **41**, 1422-1426.
67. Xue L, Stahura FL, Godden JW, Bajorath J (2001) Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **41**, 746-753.
68. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2**, 3256-3266.
69. Nasr RJ, Swamidass SJ, Baldi PF (2009) Large scale study of multiple-molecule queries. *J. Cheminform.* **1**, 1-7.
70. Harper G, Bradshaw J, Gittins JC, Green DVS, Leach AR (2001) Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **41**, 1295-1300.
71. Aitchison J, Aitken CGG (1976) Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413-420.
72. 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. *Nucl. Acids. Res.* **40**, D1100-D1107.
73. Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA (1998) Glossary of terms used in medicinal chemistry. *Pure & Appl. Chem.* **70**, 1129-1143.
74. Wolber G, Langer T (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **45**, 160-169.
75. Horvath D (2010) *Chemoinformatics and computational chemical biology (Methods in molecular biology).* Humana Press.
76. Kotsiantis SB, Zaharakis ID, Pintelas PE (2007) Supervised machine learning: A review of classification techniques. *Informatica* **31**, 249-268.
77. Ghahramani Z (2004) *Unsupervised learning.* In: *Advanced Lectures on Machine Learning* Springer, New York, 72-112.
78. Chen B, Harrison RF, Papadatos G, Willett P, Wood DJ, Lewell XQ, Greenidge P, Stiefl N (2007) Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput. Aided Mol. Des.* **21**, 53-62.
79. Schwaighofer A, Schroeter T, Mika S, Blanchard G (2009) How wrong can we get? A review of machine learning approaches and error bars. *Comb. Chem. High Throughput Screening* **12**, 453 - 468.
80. Melville JL, Burke EK, Hirst JD (2009) Machine learning in virtual screening. *Comb. Chem. High Throughput Screening* **12**, 332-343.
81. Gertrudes JC, Maltarollo VG, Silva RA, Oliveira PR, Honorio KM, da S, A.B.F. (2012) Machine learning techniques and drug design. *Curr. Med. Chem.* **19**, 4289-4297.
82. Dietterich TG (1997) Machine-learning research. *AI Mag.* **18**, 97.
83. Angluin D, Laird P (1988) Learning from noisy examples. *Mach. Learn.* **2**, 343-370.
84. Glick M, Jenkins JL, Nettles JH, Hitchings H, Davies JW (2006) Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J. Chem. Inf. Model.* **46**, 193-200.
85. Alpaydin E (2004) *Introduction to machine learning.* MIT Press, Cambridge.
86. Dietterich TG (2003) *Machine learning.* In: *Nature Encyclopedia of Cognitive Science* Macmillan, London,
87. Bishop CM, Nasrabadi NM (2006) *Pattern recognition and machine learning.* Springer, New York.
88. Shen M, LeTiran A, Xiao Y, Golbraikh A, Kohn H, Tropsha A (2002) Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.* **45**, 2811-2823.

89. Shen M, Beguin C, Golbraikh A, Stables JP, Kohn H, Tropsha A (2004) Application of predictive QSAR models to database mining: Identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **47**, 2356-2364.
90. Oloff S, Muegge I (2007) kScore: A novel machine learning approach that is not dependent on the data structure of the training set. *J. Comput. Aided Mol. Des.* **21**, 87-95.
91. Comparative evaluation of prediction algorithms (CoEPrA). <www.coepra.org> acc. 20. Nov. 2013.
92. Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Phil. Trans. R. Soc. London* **53**, 370-418.
93. Labute P (1999) Binary QSAR: A new method for the determination of quantitative structure activity relationships. *Proc. Pac. Symp. Biocomput.* **4**, 444-455.
94. Gao H, Williams C, Labute P, Bajorath J (1999) Binary quantitative structure-activity relationship (QSAR) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **39**, 164-168.
95. Makarenkov V, Zentilli P, Kevorkov D, Gagarin A, Malo N, Nadon R (2007) An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics* **23**, 1648-1657.
96. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **24**, 167-175.
97. Liu Y (2004) A comparative study on feature selection methods for drug discovery. *J. Chem. Inf. Comput. Sci.* **44**, 1823-1828.
98. Bender A, Mussa HY, Glen RC, Reiling S (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* **44**, 1708-1718.
99. Glem RC, Bender A, Arnby CH, Carlsson L, Boyer S, Smith J (2006) Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **9**, 199-204.
100. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang XP, Norval S, Sassano MF, Shin AI, Webster LA, Simeons FR, Stojanovski L, Prat A, Seidah NG, Constam DB, Bickerton GR, Read KD, Wetsel WC, Gilbert IH, Roth BL, Hopkins AL (2012) Automated design of ligands to polypharmacological profiles. *Nature* **492**, 215-220.
101. Pearl J (1988) *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufmann Publishers Inc., Burlington.
102. Abdo A, Salim N (2009) Similarity-based virtual screening with a Bayesian inference network. *ChemMedChem*
103. Chen B, Mueller C, Willett P (2009) Evaluation of a Bayesian inference network for ligand-based virtual screening. *J. Cheminform.* **1**, 5.
104. Gasteiger J, Zupan J (1993) Neural networks in chemistry. *Angew. Chem. Int. Ed.* **32**, 503-527.
105. Sadowski J, Kubinyi H (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **41**, 3325-3329.
106. Schneider G, Wrede P (1998) Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* **70**, 175-222.
107. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* **323**, 533-536.
108. Kohonen T (2000) *Self-organizing maps.* Springer, New York.
109. Kohonen T (1990) The self-organizing map. *Proc. IEEE* **78**, 1464-1480.
110. Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems* **2**, 303-314.
111. Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**, 359-366.
112. Devillers J (1996) *Neural networks in QSAR and drug design.* Academic Press, London.
113. Lobanov V (2004) Using artificial neural networks to drive virtual screening of combinatorial libraries. *Drug Discov. Today Biosilico* **2**, 149-156.

114. Gasteiger J, Teckentrup A, Terfloth L, Spycher S (2003) Neural networks as data mining tools in drug design. *J. Phys. Org. Chem.* **16**, 232-245.
115. ACD: Available Chemicals Directory, Version 2/96, MDL Information Systems, 1996.
116. WDI: World Drug Index, Version 2/96, Derwent Information, 1996.
117. Ghose AK, Crippen GM (1986) Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships 1. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **7**, 565-577.
118. Ghose AK, Crippen GM (1987) Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships 2. Modeling dispersive and hydrophobic interactions. *J. Comput. Chem.* **27**, 21-35.
119. Ghose AK, Pritchett A, Crippen GM (1988) Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships 3. Modeling hydrophobic interactions. *J. Comput. Chem.* **9**, 80-90.
120. Zuegge J, Fechner U, Roche O, Parrott NJ, Engkvist O, Schneider G (2002) A fast virtual screening filter for cytochrome P450 3A4 inhibition liability of compound libraries. *Quant. Struct.-Act. Relat.* **21**, 249-256.
121. Ajay, Bemis GW, Murcko MA (1999) Designing libraries with CNS activity. *J. Med. Chem.* **42**, 4942-4951.
122. SSKEYS, MDL Information Systems Inc., San Leandro, CA.
123. Buntine WL, Weigend AS (1991) Bayesian back-propagation. *Complex Systems* **5**, 603-643.
124. Ajay, Walters WP, Murcko MA (1998) Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **41**, 3314-3324.
125. Manallack DT, Pitt WR, Gancia E, Montana JG, Livingstone DJ, Ford MG, Whitley DC (2002) Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks. *J. Chem. Inf. Comput. Sci.* **42**, 1256-1262.
126. Balakin KV, Tkachenko SE, Lang SA, Okun I, Ivashchenko AA, Savchuk NP (2002) Property-based design of GPCR-targeted library. *J. Chem. Inf. Comput. Sci.* **42**, 1332-1342.
127. Specht DF (1990) Probabilistic neural networks. *Neural Netw.* **3**, 109-118.
128. Derksen S, Rau O, Schneider P, Schubert-Zsilavecz M, Schneider G (2006) Virtual screening for PPAR modulators using a probabilistic neural network. *ChemMedChem* **1**, 1346-1350.
129. Mueller R, Dawson ES, Meiler J, Rodriguez AL, Chauder BA, Bates BS, Felts AS, Lamb JP, Menon UN, Jadhav SB, Kane AS, Jones CK, Gregory KJ, Niswender CM, Conn PJ, Olsen CM, Winder DG, Emmitte KA, Lindsley CW (2012) Discovery of 2-(2-benzoxazoyl amino)-4-aryl-5-cyanopyrimidine as negative allosteric modulators (NAMs) of metabotropic glutamate receptor 5 (mGlu(5)): From an artificial neural network virtual screen to an *in vivo* tool compound. *ChemMedChem* **7**, 406-414.
130. Schwab CH, Hristozov D, Gasteiger J (2006) ADRIANA.Code: Algorithms for the encoding of molecular structures (version 2.0), Molecular Networks GmbH, Erlangen, Germany.
131. Cherkasov A (2005) Inductive descriptors: 10 successful years in QSAR. *Curr. Comput.-Aided Drug Des.* **1**, 21-42.
132. Cherkasov A, Shi Z, Fallahi M, Hammond GL (2005) Successful *in silico* discovery of novel nonsteroidal ligands for human sex hormone binding globulin. *J. Med. Chem.* **48**, 3203-3213.
133. Cortes C, Vapnik V (1995) Support-vector networks. *Mach. Learn.* **20**, 273-297.
134. Byvatov E, Fechner U, Sadowski J, Schneider G (2003) Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **43**, 1882-1889.
135. Krogh A, Sollich P (1997) Statistical mechanics of ensemble learning. *Phys. Rev. E* **55**, 811-825.

136. Cristianini N, Shawe-Taylor J (2000) *An introduction to support vector machines and other kernel-based learning methods.* Cambridge University Press, Cambridge.
137. Michielan L, Moro S (2010) Pharmaceutical perspectives of nonlinear QSAR strategies. *J. Chem. Inf. Model.* **50**, 961-978.
138. Czermiński R, Yasri A, Hartsough D (2001) Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat.* **20**, 227-240.
139. Burbidge R, Trotter M, Buxton B, Holden S (2001) Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* **26**, 5-14.
140. Vapnik VN (1982) *Estimation of dependences based on empirical data.* Springer, New-York, Addendum 1.
141. Thorburn WM (1915) Occam's razor. *Mind* **24**, 287-288.
142. Hoffmann R, Minkin VI, Carpenter BK (1997) Ockham's razor and chemistry. *HYLE* **3**, 3-28.
143. Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **50**, 205-216.
144. Burges CJC (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2**, 121-167.
145. Azencott C-A, Ksikes A, Swamidass SJ, Chen JH, Ralaivola L, Baldi P (2007) One-to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties. *J. Chem. Inf. Model.* **47**, 965-974.
146. Rupp M, Schneider G (2010) Graph kernels for molecular similarity. *Mol. Inf.* **29**, 266-273.
147. Hinselmann G, Rosenbaum L, Jahn A, Fechner N, Ostermann C, Zell A (2011) Large-scale learning of structure-activity relationships using a linear support vector machine and problem-specific metrics. *J. Chem. Inf. Model.* **51**, 203-213.
148. Warmuth MK, Liao J, Rätsch G, Mathieson M, Putta S, Lemmen C (2003) Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **43**, 667-673.
149. Jorissen RN, Gilson MK (2005) Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **45**, 549-561.
150. Franke L, Byvatov E, Werz O, Steinhilber D, Schneider P, Schneider G (2005) Extraction and visualization of potential pharmacophore points using support vector machines: Application to ligand-based virtual screening for COX-2 inhibitors. *J. Med. Chem.* **48**, 6997-7004.
151. Byvatov E, Schneider G (2004) SVM-based feature selection for characterization of focused compound collections. *J. Chem. Inf. Comput. Sci.* **44**, 993-999.
152. Demiriz A, Bennett KP, Breneman CM, Embrechts MJ (2001) Support vector machine regression in chemometrics. *Proc. UIST* **33**,
153. Agarwal S, Dugar D, Sengupta S (2010) Ranking chemical structures for drug discovery: A new machine learning approach. *J. Chem. Inf. Model.* **50**, 716-731.
154. Rathke F, Hansen K, Brefeld U, Müller K-R (2011) StructRank: A new approach for ligand-based virtual screening. *J. Chem. Inf. Model.* **51**, 83-92.
155. Herbrich R, Graepel T, Obermayer K (1999) Support vector learning for ordinal regression. *Proc. ICANN* **9**,
156. Herbrich R, Graepel T, Obermayer K (2000) *Large margin rank boundaries for ordinal regression.* In: *Advances in Neural Information Processing Systems* MIT Press, Cambridge, 115-132.
157. Joachims T (2002) Optimizing search engines using clickthrough data. *Proc. ACM SIGKDD* 133-142.
158. Mahe P, Ralaivola L, Stoven V, Vert JP (2006) The pharmacophore kernel for virtual screening with support vector machines. *J. Chem. Inf. Model.* **46**, 2003-2014.
159. Byvatov E, Sasse BC, Stark H, Schneider G (2005) From virtual to real screening for D_3 dopamine receptor ligands. *ChemBioChem* **6**, 997-999.

160. Tropsha A, Weifan Z (2001) Identification of the descriptor pharmacophores using variable selection QSAR applications to database mining. *Curr. Pharm. Des.* **7**, 599-612.
161. Tang H, Wang XS, Huang XP, Roth BL, Butler KV, Kozikowski AP, Jung M, Tropsha A (2009) Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation. *J. Chem. Inf. Model.* **49**, 461-476.
162. Bock JR, Gough DA (2005) Virtual screen for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model.* **45**, 1402-1414.
163. Enot DP, Gautier R, Le Marouille JY (2001) Gaussian process: An efficient technique to solve quantitative structure-property relationship problems. *SAR QSAR Environ. Res.* **12**, 461-469.
164. Burden FR (2001) Quantitative structure-activity relationship studies using Gaussian processes. *J. Chem. Inf. Comput. Sci.* **41**, 830-835.
165. O'Hagan A (1978) Curve fitting and optimal design for prediction. *J. Roy. Statist. Soc. Ser. B* **40**, 1-42.
166. Schwaighofer A, Schroeter T, Mika S, Laub J, ter Laak A, Sulzle D, Ganzer U, Heinrich N, Müller K-R (2007) Accurate solubility prediction with error bars for electrolytes: A machine learning approach. *J. Chem. Inf. Model.* **47**, 407-424.
167. Obrezanova O, Csanyi G, Gola JM, Segall MD (2007) Gaussian processes: A method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model.* **47**, 1847-1857.
168. Obrezanova O, Gola JM, Champness EJ, Segall MD (2008) Automatic QSAR modeling of ADME properties: Blood-brain barrier penetration and aqueous solubility. *J. Comput. Aided Mol. Des.* **22**, 431-440.
169. Rupp M, Schroeter T, Steri R, Zettl H, Proschak E, Hansen K, Rau O, Schwarz O, Müller-Kuhrt L, Schubert-Zsilavecz M, Müller K-R, Schneider G (2010) From machine learning to natural product derivatives that selectively activate transcription factor PPARγ. *ChemMedChem* **5**, 191-194.
170. MacKay DJC (2003) *Information theory, inference and learning algorithms.* Cambridge University Press, Cambridge.
171. Rasmussen CE, Williams CKI (2006) *Gaussian processes for machine learning.* The MIT Press, Cambridge.
172. Hansen K, Rathke F, Schroeter T, Rast G, Fox T, Kriegl JM, Mika S (2009) Bias-correction of regression models: A case study on hERG inhibition. *J. Chem. Inf. Model.* **49**, 1486-1496.
173. Obrezanova O, Segall MD (2010) Gaussian processes for classification: QSAR modeling of ADMET and target activity. *J. Chem. Inf. Model.* **50**, 1053-1061.
174. Schroeter TS, Schwaighofer A, Mika S, Ter Laak A, Suelzle D, Ganzer U, Heinrich N, Müller K-R (2007) Estimating the domain of applicability for machine learning QSAR models: A study on aqueous solubility of drug discovery molecules. *J. Comput. Aided Mol. Des.* **21**, 485-498.
175. Schaul T, Sun Y, Wierstra D, Gomez F, Schmidhuber J (2011) Curiosity-driven optimization. *Proc. CEC* 1343-1349.
176. Srinivas N, Krause A, Kakade SM, Seeger M (2010) Gaussian process optimization in the bandit setting: No regret and experimental design. *Proc. ICML*
177. Frean M, Boyle P (2008) *Using Gaussian processes to optimize expensive functions.* Springer, New York, 258-267.
178. De Grave K, Ramon J, De Raedt L (2008) Active learning for high throughput screening. *Proc. Int. Conf. Discov. Sci.* **11**, 185-196.
179. Rupp M, Bauer MR, Wilcken R, Lange A, Reutlinger M, Boeckler FM, Schneider G (2014) Machine learning estimates of natural product conformational energies. *PLoS Comput. Biol.* **10**, e1003400.
180. Wolpert DH (1992) Stacked generalization. *Neural Netw.* **5**, 241-259.
181. Breiman L (1996) Bagging predictors. *Mach. Learn.* **24**, 123-140.

182. Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q (2005) Boosting: An ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **45**, 786-799.
183. Breiman L (2001) Random forests. *Mach. Learn.* **45**, 5-32.
184. DeFries RS, Chan JC-W (2000) Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote Sens. Environ.* **74**, 503-515.
185. Renner S, Hechenberger M, Noeske T, Bocker A, Jatzke C, Schmuker M, Parsons CG, Weil T, Schneider G (2007) Searching for drug scaffolds with 3D pharmacophores and neural network ensembles. *Angew. Chem. Int. Ed.* **46**, 5336-5339.
186. Efron B, Tibshirani R (1993) *An introduction to the bootstrap.* CRC Press, Boca Raton.
187. Freund Y, Schapire RE (1997) A desicion-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst.* **55**, 119-139.
188. Seierstad M, Agrafiotis DK (2006) A QSAR model of hERG binding using a large, diverse, and internally consistent training set. *Chem. Biol. Drug Des.* **67**, 284-296.
189. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947-1958.
190. Ehrman TM, Barlow DJ, Hylands PJ (2007) Virtual screening of Chinese herbs with random forest. *J. Chem. Inf. Model.* **47**, 264-278.
191. Leach AR, Gillet VJ, Lewis RA, Taylor R (2010) Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **53**, 539-558.
192. Muegge I, Oloff S (2006) Advances in virtual screening. *Drug Discov. Today Technol.* **3**, 405-411.
193. Drwal MN, Griffith R (2013) Combination of ligand- and structure-based methods in virtual screening. *Drug Discov. Today Technol.* **10**, e395-e401.
194. Greenidge PA, Carlsson B, Bladh L-G, Gillner M (1998) Pharmacophores incorporating numerous excluded volumes defined by X-ray crystallographic structure in three-dimensional database searching: Application to the thyroid hormone receptor. *J. Med. Chem.* **41**, 2503-2512.
195. Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**, 849-857.
196. Böhm H-J (1992) The computer program LUDI: A new method for the *de novo* design of enzyme inhibitors. *J. Comput. Aided Mol. Des.* **6**, 61-78.
197. Pickett S (2005) *The biophore concept.* In: *Protein-Ligand Interactions: From Molecular Recognition to Drug Design* (eds. Böhm H-J, Schneider G) Wiley-VCH, Weinheim, 73-105.
198. Böhm H-J, Boehringer M, Bur D, Gmuender H, Huber W, Klaus W, Kostrewa D, Kuehne H, Luebbers T, Meunier-Keller N (2000) Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem.* **43**, 2664-2674.
199. Löwer M, Geppert T, Schneider P, Hoy B, Wessler S, Schneider G (2011) Inhibitors of Helicobacter pylori protease HtrA found by 'virtual ligand'screening combat bacterial invasion of epithelia. *PLoS ONE* **6**, e17986.
200. Rella M, Rushworth CA, Guy JL, Turner AJ, Langer T, Jackson RM (2006) Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. *J. Chem. Inf. Model.* **46**, 708-716.
201. Sanders MPA, Verhoeven S, de Graaf C, Roumen L, Vroling B, Nabuurs SB, de Vlieg J, Klomp JPG (2011) Snooker: A structure-based pharmacophore generation tool applied to class A GPCRs. *J. Chem. Inf. Model.* **51**, 2277-2292.
202. Horvath D (2011) *Pharmacophore-based virtual screening.* In: *Chemoinformatics and Computational Chemical Biology* Springer, New York, 261-298.
203. Lyne PD (2002) Structure-based virtual screening: An overview. *Drug Discov. Today* **7**, 1047-1055.

204. Klebe G (2006) Virtual ligand screening: Strategies, perspectives and limitations. *Drug Discov. Today* **11**, 580-594.
205. Stahl M, Guba W, Kansy M (2006) Integrating molecular design resources within modern drug discovery research: The Roche experience. *Drug Discov. Today* **11**, 326-333.
206. Blaney JM, Dixon JS (1993) A good ligand is hard to find: Automated docking methods. *Perspect. Drug Discovery Des.* **1**, 301-319.
207. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **153 Suppl 1**, S7-26.
208. Yuriev E, Ramsland PA (2013) Latest developments in molecular docking: 2010-2011 in review. *J. Mol. Recognit.* **26**, 215-239.
209. Huang S-Y, Grinter SZ, Zou X (2010) Scoring functions and their evaluation methods for protein–ligand docking: Recent advances and future directions. *Phys. Chem. Chem. Phys.* **12**, 12899-12908.
210. Waszkowycz B, Clark DE, Gancia E (2011) Outstanding challenges in protein–ligand docking and structure-based virtual screening. *WIREs Comput. Mol. Sci.* **1**, 229-259.
211. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **43**, 4759-4767.
212. O'Boyle NM, Liebeschuetz JW, Cole JC (2009) Testing assumptions and hypotheses for rescoring success in protein− ligand docking. *J. Chem. Inf. Model.* **49**, 1871-1878.
213. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK (2004) Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739-1749.
214. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **19**, 1639-1662.
215. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein–ligand docking using GOLD. *Proteins: Struct., Funct., Bioinf.* **52**, 609-623.
216. Jain AN (2003) Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **46**, 499-511.
217. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **261**, 470-489.
218. Oshiro CM, Kuntz ID, Dixon JS (1995) Flexible ligand docking using a genetic algorithm. *J. Comput. Aided Mol. Des.* **9**, 113-130.
219. Abagyan R, Totrov M, Kuznetsov D (1994) ICM — a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **15**, 488-506.
220. Corbeil CR, Williams CI, Labute P (2012) Variability in docking success rates due to dataset preparation. *J. Comput. Aided Mol. Des.* **26**, 775-786.
221. Wu G, Robertson DH, Brooks CL, Vieth M (2003) Detailed analysis of grid-based molecular docking: A case study of CDOCKER—A CHARMm-based MD docking algorithm. *J. Comput. Chem.* **24**, 1549-1562.
222. Zsoldos Z, Reid D, Simon A, Sadjad BS, Peter Johnson A (2006) eHiTS: An innovative approach to the docking and scoring function problems. *Curr. Protein Pept. Sci.* **7**, 421-435.
223. Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* **16**, 151-166.
224. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269-288.
225. Fischer D, Norel R, Wolfson H, Nussinov R (1993) Surface motifs by a computer vision technique: Searches, detection, and implications for protein-ligand recognition. *Proteins: Struct., Funct., Bioinf.* **16**, 278-292.

226. Norel R, Fischer D, Wolfson HJ, Nussinov R (1994) Molecular surface recognition by a computer vision-based technique. *Protein Eng.* **7**, 39-46.
227. Hawkins PC, Skillman AG, Nicholls A (2007) Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **50**, 74-82.
228. Lorber DM, Shoichet BK (1998) Flexible ligand docking using conformational ensembles. *Protein Sci.* **7**, 938-950.
229. Metropolis N, Ulam S (1949) The Monte Carlo method. *J. Am. Stat. Assoc.* **44**, 335-341.
230. Goldberg DE (1989) *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley, Boston.
231. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727-748.
232. Michel J, Verdonk ML, Essex JW (2006) Protein-ligand binding affinity predictions by implicit solvent simulations: A tool for lead optimization? *J. Med. Chem.* **49**, 7427-7439.
233. Plewczynski D, Łaźniewski M, Augustyniak R, Ginalski K (2011) Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **32**, 742-755.
234. Schulz-Gasch T, Stahl M (2004) Scoring functions for protein–ligand interactions: A critical perspective. *Drug Discov. Today Technol.* **1**, 231-239.
235. Huang N, Kalyanaraman C, Irwin JJ, Jacobson MP (2006) Physics-based scoring of protein-ligand complexes: Enrichment of known inhibitors in large-scale virtual screening. *J. Chem. Inf. Model.* **46**, 243-253.
236. Meng EC, Shoichet BK, Kuntz ID (1992) Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **13**, 505-524.
237. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765-784.
238. Zwanzig RW (1954) High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* **22**, 1420.
239. Mitchell MJ, McCammon JA (1991) Free energy difference calculations by thermodynamic integration: Difficulties in obtaining a precise value. *J. Comput. Chem.* **12**, 271-275.
240. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **295**, 337-356.
241. Massova I, Kollman PA (2000) Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discovery Des.* **18**, 113-135.
242. The Nobel Prize in chemistry 2013. Nobelprize.org. Nobel Media AB. <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/> acc. 28 Nov. 2013.
243. Warshel A, Levitt M (1976) Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227-249.
244. Menikarachchi LC, Gascón JA (2010) QM/MM approaches in medicinal chemistry research. *Curr. Top. Med. Chem.* **10**, 46-54.
245. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **11**, 425-445.
246. Böhm H-J (1994) On the use of LUDI to search the fine chemicals directory for ligands of proteins of known three-dimensional structure. *J. Comput. Aided Mol. Des.* **8**, 623-632.
247. Böhm H-J (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des.* **8**, 243-256.

248. Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD (1998) Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins: Struct., Funct., Bioinf.* **33**, 367-382.
249. Muegge I, Martin YC (1999) A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **42**, 791-804.
250. Dill KA (1997) Additivity principles in biochemistry. *J. Biol. Chem.* **272**, 701-704.
251. Cheng T, Li X, Li Y, Liu Z, Wang R (2009) Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **49**, 1079-1093.
252. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **49**, 5912-5931.
253. Shoichet BK, McGovern SL, Wei B, Irwin JJ (2002) Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **6**, 439-446.
254. Carlsson J, Coleman RG, Setola V, Irwin JJ, Fan H, Schlessinger A, Sali A, Roth BL, Shoichet BK (2011) Ligand discovery from a dopamine D_3 receptor homology model and crystal structure. *Nat. Chem. Biol.* **7**, 769-778.
255. Irwin JJ, Shoichet BK, Mysinger MM, Huang N, Colizzi F, Wassam P, Cao Y (2009) Automated docking screens: A feasibility study. *J. Med. Chem.* **52**, 5712-5720.
256. Funatsu K, Miyao T, Arakawa M (2011) Systematic generation of chemical structures for rational drug design based on QSAR models. *Curr. Comput.-Aided Drug Des.* **7**, 1-9.
257. Van Drie JH, Weininger D, Martin YC (1989) ALADDIN: An integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Chem. Inf. Comput. Sci.* **3**, 225-251.
258. Moon JB, Howe WJ (1991) Computer design of bioactive molecules: A method for receptor-based *de novo* ligand design. *Proteins: Struct., Funct., Bioinf.* **11**, 314-328.
259. Schneider G, Fechner U (2005) Computer-based *de novo* design of drug-like molecules. *Nat. Rev. Drug Discov.* **4**, 649-663.
260. Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. *Nature* **432**, 855-861.
261. Dobson CM (2004) Chemical space and biology. *Nature* **432**, 824-828.
262. Babaoglu K, Shoichet BK (2006) Deconstructing fragment-based inhibitor discovery. *Nat. Chem. Biol.* **2**, 720-723.
263. Pierce AC, Rao G, Bemis GW (2004) BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease. *J. Med. Chem.* **47**, 2768-2775.
264. Ertl P, Lewis R (2012) IADE: A system for intelligent automatic design of bioisosteric analogs. *J. Comput. Aided Mol. Des.* **26**, 1207-1215.
265. Bergmann R, Linusson A, Zamora I (2007) SHOP: Scaffold HOPping by GRID-based similarity searches. *J. Med. Chem.* **50**, 2708-2717.
266. Hajduk PJ (2006) Fragment-based drug design: How big is too big? *J. Med. Chem.* **49**, 6972-6976.
267. Ichihara O, Barker J, Law RJ, Whittaker M (2011) Compound design by fragment-linking. *Mol. Inf.* **30**, 298-306.
268. Nazare M, Matter H, Will DW, Wagner M, Urmann M, Czech J, Schreuder H, Bauer A, Ritter K, Wehner V (2012) Fragment deconstruction of small, potent factor Xa inhibitors: Exploring the superadditivity energetics of fragment linking in protein-ligand complexes. *Angew. Chem. Int. Ed.* **51**, 905-911.
269. 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.

270. Schneider G, Lee M-L, Stahl M, Schneider P (2000) *De novo* design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput. Aided Mol. Des.* **14**, 487-494.
271. Boda K, Seidel T, Gasteiger J (2007) Structure and reaction based evaluation of synthetic accessibility. *J. Comput. Aided Mol. Des.* **21**, 311-325.
272. Gillet VJ, Myatt G, Zsoldos Z, Johnson AP (1995) SPROUT, HIPPO and CAESA: Tools for *de novo* structure generation and estimation of synthetic accessibility. *Perspect. Drug Discovery Des.* **3**, 34-50.
273. Hartenfeller M, Zettl H, Walter M, Rupp M, Reisen F, Proschak E, Weggen S, Stark H, Schneider G (2012) DOGS: Reaction-driven de novo design of bioactive compounds. *PLoS Comput. Biol.* **8**, e1002380.
274. Vinkers HM, de Jonge MR, Daeyaert FFD, Heeres J, Koymans LMH, van Lenthe JH, Lewi PJ, Timmerman H, Van Aken K, Janssen PAJ (2003) SYNOPSIS: SYNthesize and OPtimize system *in silico*. *J. Med. Chem.* **46**, 2765-2773.
275. Peng Z (2013) Very large virtual compound spaces: Construction, storage and utility in drug discovery. *Drug Discov. Today Technol.*
276. Schneider G, Geppert T, Hartenfeller M, Reisen F, Klenner A, Reutlinger M, Hahnke V, Hiss JA, Zettl H, Keppner S, Spankuch B, Schneider P (2011) Reaction-driven *de novo* design, synthesis and testing of potential type II kinase inhibitors. *Future Med. Chem.* **3**, 415-424.
277. Hu Q, Peng Z, Sutton SC, Na J, Kostrowicki J, Yang B, Thacher T, Kong X, Mattaparti S, Zhou JZ (2012) Pfizer Global Virtual Library (PGVL): A chemistry design tool powered by experimentally validated parallel synthesis information. *ACS Comb. Sci.* **14**, 579-589.
278. Ertl P (2003) Cheminformatics analysis of organic substituents: Identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **43**, 374-380.
279. Schneider G. (ed.) (2013) *De Novo Molecular Design*. Wiley-VCH, Weinheim.
280. Pearlman DA, Murcko MA (1993) CONCEPTS: New dynamic algorithm for *de novo* drug suggestion. *J. Comput. Chem.* **14**, 1184-1193.
281. DeWitte RS, Shakhnovich EI (1996) SMoG: *De novo* design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **118**, 11733-11744.
282. Koza JR (1992) *Genetic Programming: On the programming of computers by means of natural selection.* MIT Press, Cambridge.
283. Glen RC, Payne AWR (1995) A genetic algorithm for the automated generation of molecules within constraints. *J. Comput. Aided Mol. Des.* **9**, 181-202.
284. Globus A, Lawton J, Wipke T (1999) Automatic molecular design using evolutionary techniques. *Nanotechnology* **10**, 290-299.
285. Alig L, Alsenz J, Andjelkovic M, Bendels S, Benardeau A, Bleicher K, Bourson A, David-Pierson P, Guba W, Hildbrand S, Kube D, Lubbers T, Mayweg AV, Narquizian R, Neidhart W, Nettekoven M, Plancher JM, Rocha C, Rogers-Evans M, Rover S, Schneider G, Taylor S, Waldmeier P (2008) Benzodioxoles: Novel cannabinoid-1 receptor inverse agonists for the treatment of obesity. *J. Med. Chem.* **51**, 2115-2127.
286. Visco DP, Pophale RS, Rintoul MD, Faulon J-L (2002) Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor. *J. Mol. Graph. Model.* **20**, 429-438.
287. Waszkowycz B, Clark DE, Frenkel D, Li J, Murray CW, Robson B, Westhead DR (1994) PRO_LIGAND: An approach to de novo molecular design. 2. Design of novel molecules from molecular field analysis (MFA) models and pharmacophores. *J. Med. Chem.* **37**, 3994-4002.
288. Cramer RD, Patterson DE, Bunce JD (1988) *Comparative Molecular Field Analysis* (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959-5967.

289. Pisabarro MT, Ortiz AR, Palomer A, Cabre F, Garcia L, Wade RC, Gago F, Mauleon D, Carganico G (1994) Rational modification of human synovial fluid phospholipase A2 inhibitors. *J. Med. Chem.* **37**, 337-341.
290. Schneider G, Schrödl W, Wallukat G, Müller J, Nissen E, Rönspeck W, Wrede P, Kunze R (1998) Peptide design by artificial neural networks and computer-based evolutionary search. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12179-12184.
291. van der Horst E, Marqués-Gallego P, Mulder-Krieger T, van Veldhoven J, Kruisselbrink J, Aleman A, Emmerich MTM, Brussee J, Bender A, IJzerman AP (2012) Multi-objective evolutionary design of adenosine receptor ligands. *J. Chem. Inf. Model.* **52**, 1713-1721.
292. Loving K, Alberts I, Sherman W (2010) Computational approaches for fragment-based and *de novo* design. *Curr. Top. Med. Chem.* **10**, 14-32.
293. Kutchukian PS, Shakhnovich EI (2010) De novo design: balancing novelty and confined chemical space. *Expert Opin. Drug Discov.* **5**, 789-812.
294. Bieler M, Koeppen H (2012) The Role of Chemogenomics in the Pharmaceutical Industry. *Drug Dev. Res.* **73**, 357-364.
295. Hodgson J (2001) ADMET--turning chemicals into drugs. *Nat. Biotechnol.* **19**, 722-726.
296. Oprea TI (2002) Virtual screening in lead discovery: A viewpoint. *Molecules* **7**, 51-62.
297. Baringhaus K, Matter H (2005) *Efficient strategies for lead optimization by simultaneously addressing affinity, selectivity and pharmacokinetic parameters.* In: *Chemoinformatics in Drug Discovery* (ed. Oprea TI) Wiley-VCH, Weinheim, 333-379.
298. Nicolaou CA, Brown N (2013) Multi-objective optimization methods in drug design. *Drug Discov. Today Technol.*
299. Bottegoni G, Favia AD, Recanatini M, Cavalli A (2012) The role of fragment-based and computational methods in polypharmacology. *Drug Discov. Today* **17**, 23-34.
300. Nicolaou CA, Brown N, Pattichis CS (2007) Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discovery Dev.* **10**, 316.
301. Derringer G, Suich R (1980) Simultaneous optimization of several response variables. *J. Quality Technol.* **12**, 214-219.
302. Cruz-Monteagudo M, Borges F, Cordeiro MNDS, Cagide Fajin JL, Morell C, Ruiz RM, Cañizares-Carmenate Y, Dominguez ER (2008) Desirability-based methods of multiobjective optimization and ranking for global QSAR studies. Filtering safe and potent drug candidates from combinatorial libraries. *J. Comb. Chem.* **10**, 897-913.
303. Cruz-Monteagudo M, Borges F, Cordeiro MNDS (2008) Desirability-based multiobjective optimization for global QSAR studies: Application to the design of novel NSAIDs with improved analgesic, antiinflammatory, and ulcerogenic profiles. *J. Comput. Chem.* **29**, 2445-2459.
304. Gillet VJ, Willett P, Fleming PJ, Green DVS (2002) Designing focused libraries using MoSELECT. *J. Mol. Graph. Model.* **20**, 491-498.
305. Brown N, McKay B, Gasteiger J (2004) The de novo design of median molecules within a property range of interest. *J. Comput. Aided Mol. Des.* **18**, 761-771.
306. Nicolaou CA, Apostolakis J, Pattichis CS (2009) De novo drug design using multiobjective evolutionary graphs. *J. Chem. Inf. Model.* **49**, 295-307.
307. Ekins S, Honeycutt JD, Metz JT (2010) Evolving molecules using multi-objective optimization: Applying to ADME/Tox. *Drug Discov. Today* **15**, 451-460.
308. Cruz-Monteagudo M, D.S. C, M., Tejera E, Dominguez E, Borges F (2012) Desirability-based multi-objective QSAR in drug discovery. *Mini Rev. Med. Chem.* **12**, 920-935.
309. Jalencas X, Mestres J (2013) On the origins of drug polypharmacology. *MedChemComm* **4**, 80-87.
310. Kaufmann SHE (2008) Paul Ehrlich: Founder of chemotherapy. *Nat. Rev. Drug Discov.* **7**, 373-373.
311. Mestres J, Gregori-Puigjane E, Valverde S, Sole RV (2008) Data completeness--the Achilles heel of drug-target networks. *Nat. Biotechnol.* **26(9)**, 983-984.

312. Azzaoui K, Hamon J, Faller B, Whitebread S, Jacoby E, Bender A, Jenkins JL, Urban L (2007) Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* **2**, 874-880.
313. Xie L, Xie L, Kinnings SL, Bourne PE (2012) Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Annu. Rev. Pharmacol. Toxicol.* **52**, 361-379.
314. Knight ZA, Lin H, Shokat KM (2010) Targeting the cancer kinome through polypharmacology. *Nat. Rev. Cancer* **10**, 130-137.
315. Hippius H (1999) A historical perspective of clozapine. *J. Clin. Psychiatry*
316. Roth BL, Sheffler D, Potkin SG (2003) Atypical antipsychotic drug actions: Unitary or multiple mechanisms for 'atypicality'? *Clin. Neurosci. Res.* **3**, 108-117.
317. Hopkins AL (2007) Network pharmacology. *Nat. Biotechnol.* **25**, 1110-1111.
318. Aksamitiene E, Kiyatkin A, Kholodenko BN (2012) Cross-talk between mitogenic Ras/MAPK and survival PI3K/Akt pathways: a fine balance. *Biochem. Soc. Trans.* **40**, 139-146.
319. Jahangiri A, Weiss WA (2013) It takes two to tango: Dual inhbition of PI3K & MAPK in rhabdomyosarcoma. *Clin. Cancer Res.* **19**, 5811-5813.
320. Schiller JH, Harrington D, Belani CP, Langer C, Sandler A, Krook J, Zhu J, Johnson DH (2002) Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. *N. Engl. J. Med.* **346**, 92-98.
321. Croft SL, Sundar S, Fairlamb AH (2006) Drug resistance in leishmaniasis. *Clin. Microbiol. Rev.* **19**, 111-126.
322. Grossberg GT, Edwards KR, Zhao Q (2006) Rationale for combination therapy with galantamine and memantine in Alzheimer's disease. *J. Clin. Pharmacol.* **46**, 17S-26S.
323. Cavalli A, Bolognesi ML, Minarini A, Rosini M, Tumiatti V, Recanatini M, Melchiorre C (2008) Multi-target-directed ligands to combat neurodegenerative diseases. *J. Med. Chem.* **51**, 347-372.
324. Giordano S, Petrelli A (2008) From single- to multi-target drugs in cancer therapy: When aspecificity becomes an advantage. *Curr. Med. Chem.* **15**, 422-432.
325. Schneider P, Stutz K, Kasper L, Haller S, Reutlinger M, Reisen F, Geppert T, Schneider G (2011) Target profile prediction and practical evaluation of a Biginelli-type dihydropyrimidine compound library. *Pharmaceuticals* **4**, 1236-1247.
326. Morphy R, Rankovic Z (2006) The physicochemical challenges of designing multiple ligands. *J. Med. Chem.* **49**, 4961-4970.
327. Achenbach J, Klingler FM, Blöcher R, Moser D, Häfner AK, Rödl CB, Kretschmer S, Krüger B, Löhr F, Stark H, Hofmann B, Steinhilber D, Proschak E (2013) Exploring the chemical space of multi-target ligands using aligned self-organizing maps. *ACS Med. Chem. Lett.*
328. Hopkins AL (2008) Network pharmacology: The next paradigm in drug discovery. *Nat. Chem. Biol.* **4**, 682-690.
329. Wermuth CG (2006) Selective optimization of side activities: The SOSA approach. *Drug Discov. Today* **11**, 160-164.
330. Hart CP (2005) Finding the target after screening the phenotype. *Drug Discov. Today* **10**, 513-519.
331. Jenkins JL, Bender A, Davies JW (2006) *In silico* target fishing: Predicting biological targets from chemical structure. *Drug Discov. Today Technol.* **3**, 413-421.
332. Oprea TI, Mestres J (2012) Drug repurposing: Far beyond new targets for old drugs. *AAPS J.* **14**, 759-763.
333. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Côté S, Shoichet BK, Urban L (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361-367.
334. Schneider G, Tanrikulu Y, Schneider P (2009) Self-organizing molecular fingerprints: A ligand-based view on drug-like chemical space and off-target prediction. *Future Med. Chem.* **1**, 213-218.

335. Schmuker M, De Bruyne M, Hähnel M, Schneider G (2007) Predicting olfactory receptor neuron responses from odorant structure. *Chem. Cent. J.* **1**, 11.
336. Schmuker M, Schneider G (2007) Processing and classification of chemical data inspired by insect olfaction. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20285-20289.
337. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197-206.
338. Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71-84.
339. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KL, Edwards DD, Shoichet BK, Roth BL (2009) Predicting new molecular targets for known drugs. *Nature* **462**, 175-181.
340. Schuffenhauer A, Zimmermann J, Stoop R, van der Vyver J-J, Lecchini S, Jacoby E (2002) An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J. Chem. Inf. Comput. Sci.* **42**, 947-955.
341. Gregori-Puigjané E, Mestres J (2008) A ligand-based approach to mining the chemogenomic space of drugs. *Comb. Chem. High Throughput Screening* **11**, 669-676.
342. Achenbach J, Tiikkainen P, Franke L, Proschak E (2011) Computational tools for polypharmacology and repurposing. *Future Med. Chem.* **3**, 961-968.
343. Filimonov DA, Poroikov VV, Karaicheva EI, Kazarian RK, Budunova AP, Mikhailovskii EM, Rudnitskikh AV, Goncharenko LV, Burov I (1995) The computerized prediction of the spectrum of biological activity of chemical compounds by their structural formula: The PASS system. Prediction of activity spectra for substance. *Eksp. Klin. Farmakol.* **58**, 56-62.
344. Lagunin A, Stepanchikova A, Filimonov D, Poroikov V (2000) PASS: Prediction of activity spectra for biologically active substances. *Bioinformatics* **16**, 747-748.
345. Poroikov V, Filimonov D, Lagunin A, Gloriozova T, Zakharov A (2007) PASS: Identification of probable targets and mechanisms of toxicity. *SAR QSAR Environ. Res.* **18**, 101-110.
346. Nidhi, Glick M, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **46**, 1124-1133.
347. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI (2004) *WOMBAT: World of molecular bioactivity.* In: *Chemoinformatics in drug discovery* (ed. Oprea TI) Wiley-VCH, New York, 223-239.
348. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* **24**, 805-815.
349. Rognan D (2010) Structure-based approaches to target fishing and ligand profiling. *Mol. Inf.* **29**, 176-187.
350. Do Q-T, Renimel I, Andre P, Lugnier C, Muller C, Bernard P (2005) Reverse Pharmacognosy: Application of Selnergy, a New Tool for Lead Discovery. The Example of ε-Viniferin. *Curr. Drug Discovery Technol.* **2**, 161-167.
351. Muller P, Lena G, Boilard E, Bezzine S, Lambeau G, Guichard G, Rognan D (2006) *In silico*-guided target identification of a scaffold-focused library: 1, 3, 5-Triazepan-2, 6-diones as novel phospholipase A2 inhibitors. *J. Med. Chem.* **49**, 6768-6778.
352. Reisen F, Weisel M, Kriegl JM, Schneider G (2010) Self-organizing fuzzy graphs for structure-based comparison of protein pockets. *J. Proteome. Res.* **9**, 6498-6510.
353. Rollinger JM, Schuster D, Danzl B, Schwaiger S, Markt P, Schmidtke M, Gertsch J, Raduner S, Wolber G, Langer T (2009) In silico target fishing for rationalized ligand discovery exemplified on constituents of Ruta graveolens. *Planta Med.* **75**, 195.
354. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. *Science* **321**, 263-266.

355. Iorio F, Isacchi A, di Bernardo D, Brunetti-Pierri N (2010) Identification of small molecules enhancing autophagic function from drug network analysis. *Autophagy* **6**, 1204-1205.
356. Schneider G, Hartenfeller M, Reutlinger M, Tanrikulu Y, Proschak E, Schneider P (2009) Voyages to the (un)known: Adaptive design of bioactive compounds. *Trends Biotechnol.* **27**, 18-26.
357. Ahlberg C (1999) Visual exploration of HTS databases: Bridging the gap between chemistry and biology. *Drug Discov. Today* **4**, 370-376.
358. Bajorath J, Peltason L, Wawer M, Guha R, Lajiness MS, Van Drie JH (2009) Navigating structure-activity landscapes. *Drug Discov. Today* **14**, 698-705.
359. Maniyar DM, Nabney IT, Williams BS, Sewing A (2006) Data visualization during the early stages of drug discovery. *J. Chem. Inf. Model.* **46**, 1806-1818.
360. Hassan M, Bielawski JP, Hempel JC, Waldman M (1996) Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Divers.* **2**, 64-74.
361. Valler MJ, Green D (2000) Diversity screening versus focussed screening in drug discovery. *Drug Discov. Today* **5**, 286-293.
362. Akella LB, DeCaprio D (2010) Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **14**, 325-330.
363. Givehchi A, Dietrich A, Wrede P, Schneider G (2003) ChemSpaceShuttle: A tool for data mining in drug discovery by classification, projection, and 3D visualization. *QSAR Comb. Sci.* **22**, 549-559.
364. Klenner A, Hähnke V, Geppert T, Schneider P, Zettl H, Haller S, Rodrigues T, Reisen F, Hoy B, Schaible AM, Werz O, Wessler S, Schneider G (2012) From virtual screening to bioactive compounds by visualizing and clustering of chemical space. *Mol. Inf.* **31**, 21-26.
365. Balakin KV, Ivanenkov YA, Savchuk NP, Ivashchenko AA, Ekins S (2005) Comprehensive computational assessment of ADME properties using mapping techniques. *Curr. Drug Discovery Technol.* **2**, 99-113.
366. Feher M, Schmidt JM (2003) Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **43**, 218-227.
367. Jolliffe I (1986) *Principal component analysis.* Springer, New York.
368. Shepard RN (1962) The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* **27**, 219-246.
369. Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325-338.
370. Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59-69.
371. Zupan J, Gasteiger J (1999) *Neural networks in chemistry and drug design, 2nd edition.* Wiley-VCH, Weinheim.
372. Digles D, Ecker GF (2011) Self-organizing maps for *in silico* screening and data visualization. *Mol. Inf.* **30**, 838-846.
373. Stahl M, Taroni C, Schneider G (2000) Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng. Des. Sel.* **13**, 83-88.
374. Bauknecht H, Zell A, Bayer H, Levi P, Wagener M, Sadowski J, Gasteiger J (1996) Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: Dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **36**, 1205-1213.
375. Sadowski J, Wagener M, Gasteiger J (1996) Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem. Int. Ed.* **34**, 2674-2677.
376. Schneider P, Tanrikulu Y, Schneider G (2009) Self-organizing maps in drug discovery: Compound library design, scaffold-hopping, repurposing. *Curr. Med. Chem.* **16**, 258-266.

377. Schneider G, Nettekoven M (2003) Ligand-based combinatorial design of selective purinergic receptor (A_{2A}) antagonists using self-organizing maps. *J. Comb. Chem.* **5**, 233-237.
378. Bishop CM, Svensén M, Williams CKI (1998) GTM: The generative topographic mapping. *Neural Comput.* **10**, 215-234.
379. Tino P, Nabney I (2002) Hierarchical GTM: Constructing localized nonlinear projection manifolds in a principled way. *IEEE Trans. Pattern Anal. Machine Intell.* **24**, 639-656.
380. Sammon JW (1969) A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **C-18**, 401-409.
381. Agrafiotis DK (2003) Stochastic proximity embedding. *J. Comput. Chem.* **24**, 1215-1221.
382. Hinton GE, Roweis ST (2002) Stochastic neighbor embedding. *Proc. NIPS* **15**, 833-840.
383. Mika S, Schölkopf B, Smola AJ, Müller K-R, Scholz M, Rätsch G (1998) Kernel PCA and de-noising in feature spaces. *Proc. NIPS* **11**, 536-542.
384. Bunte K, Hammer B, Villmann T, Biehl M, Wismüller A (2010) Exploratory observation machine (XOM) with Kullback-Leibler divergence for dimensionality reduction and visualization. *Proc. ESANN* **10**, 87-92.
385. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323-2326.
386. Van der Maaten LJP, Hinton GE (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579-2605.
387. Bajorath J (2012) Modeling of activity landscapes for drug discovery. *Expert Opin. Drug Discov.* **7**, 463-473.
388. Wassermann AM, Wawer M, Bajorath J (2010) Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.* **53**, 8209-8223.
389. Guha R (2012) Exploring structure-activity data using the landscape paradigm. *WIREs Comput. Mol. Sci.* **2**,
390. Agrafiotis DK, Shemanarev M, Connolly PJ, Farnum M, Lobanov VS (2007) SAR maps: A new SAR visualization technique for medicinal chemists. *J. Med. Chem.* **50**, 5926-5937.
391. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE (1996) Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **39**, 3049-3059.
392. Shanmugasundaram V, Maggiora GM (2001) Characterizing property and activity landscapes using an information-theoretic approach. *Proc. 222nd ACS Meeting*
393. Medina-Franco JL (2012) Scanning structure-activity relationships with structure-activity similarity and related maps: From consensus activity cliffs to selectivity switches. *J. Chem. Inf. Model.* **52**, 2485-2493.
394. Lajiness M (1991) *The evaluation of the performance of dissimilarity selection.* In: *QSAR: Rational approaches to the design of bioactive compounds* (eds. Silipo C, Vittoria A) Elsevier, Amsterdam, 201-204.
395. Leung CS, Leung SS, Tirado-Rives J, Jorgensen WL (2012) Methyl effects on protein-ligand binding. *J. Med. Chem.* **55**, 4489-4500.
396. Barratt E, Bronowska A, Vondrasek J, Cerny J, Bingham R, Phillips S, Homans SW (2006) Thermodynamic penalty arising from burial of a ligand polar group within a hydrophobic pocket of a protein receptor. *J. Mol. Biol.* **362**, 994-1003.
397. Rücker C, Scarsi M, Meringer M (2006) 2D QSAR of PPARγ agonist binding and transactivation. *Bioorg. Med. Chem.* **14**, 5178-5195.
398. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry: Miniperspective. *J. Med. Chem.* **55**, 2932-2942.
399. Sisay MT, Peltason L, Bajorath J (2009) Structural interpretation of activity cliffs revealed by systematic analysis of structure− activity relationships in analog series. *J. Chem. Inf. Model.* **49**, 2179-2189.
400. Stumpfe D, Hu Y, Dimova D, Bajorath J (2013) Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *J. Med. Chem.*

401. Tebby C, Mombelli E (2013) Modelling structure activity landscapes with cliffs: A kernel regression-based approach. *Mol. Inf.* **32**, 609-623.
402. Kauffman SA, Johnsen S (1991) Coevolution to the edge of chaos: Coupled fitness landscapes, poised states, and coevolutionary avalanches. *J. Theor. Biol.* **149**, 467-505.
403. Maggiora GM, Shanmugasundaram V, Lajiness MS, Doman TN, Schulz MW, Oprea TI (2005) *A practical strategy for directed compound acquisition.* Wiley-VCH, Weinheim.
404. Peltason L, Iyer P, Bajorath J (2010) Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J. Chem. Inf. Model.* **50**, 1021-1033.
405. Johnson SR (2008) The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **48**, 25-26.
406. Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **18**, 326-332.
407. De Jong H (2002) Modeling and simulation of genetic regulatory systems: A literature review. *J. Comput. Biol.* **9**, 67-103.
408. Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks and connectivity indices. *Proteomics* **8**, 750-778.
409. Miyashita Y, Takahashi Y, Yotsui Y, Abe H, Sasaki S-I (1981) Application of pattern recognition to structure-activity problems. *Anal. Chim. Acta* **133**, 615-624.
410. Bienfait B, Gasteiger J (1997) Checking the projection display of multivariate data with colored graphs. *J. Mol. Graph. Model.* **15**, 203-215.
411. Johnson M (1993) Structure-activity maps for visualizing the graph variables arising in drug design. *J. Biopharm. Stat.* **3**, 203-236.
412. Gifford E, Johnson M, Smith D, Tsai C- (1996) Structure-reactivity maps as a tool for visualizing xenobiotic structure-reactivity relationships. *Network Sci.* **2**, 1-33.
413. Wollenhaupt S, Baumann K (2012) INSARA: A new method for the analysis and visualization of Structure-Activity-Relationships. *J. Cheminform.* **4**, P44.
414. Lepp Z, Huang C, Okada T (2009) Finding key members in compound libraries by analyzing networks of molecules assembled by structural similarity. *J. Chem. Inf. Model.* **49**, 2429-2443.
415. Wawer M, Peltason L, Weskamp N, Teckentrup A, Bajorath J (2008) Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J. Med. Chem.* **51**, 6075-6084.
416. Lounkine E, Wawer M, Wassermann AM, Bajorath J (2010) SARANEA: A freely available program to mine structure-activity and structure-selectivity relationship information in compound data sets. *J. Chem. Inf. Model.* **50**, 68-78.
417. Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK (2008) Quantifying the relationships among drug classes. *J. Chem. Inf. Model.* **48**, 755-765.
418. Schneider G, Schuchhardt J, Wrede P (1995) Development of simple fitness landscapes for peptides by artificial neural filter systems. *Biol. Cybern.* **73**, 245-254.
419. Oprea TI, Gottfries J (2001) Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **3**, 157-166.
420. Guha R, Van Drie JH (2008) Structure-activity landscape index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **48**, 646-658.
421. Iyer P, Hu Y, Bajorath J (2011) SAR monitoring of evolving compound data sets using activity landscapes. *J. Chem. Inf. Model.* **51**, 532-540.
422. Guha R (2011) The ups and downs of structure-activity landscapes. *Methods Mol. Biol.* **672**, 101-117.
423. Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Phil. Trans.* **2**, 559-572.
424. Ståhle L, Wold S (1986) On the use of some multivariate statistical methods in pharmacological research. *J. Pharmacol. Methods* **16**, 91-110.

425. Schneider G, So S-S (2003) *Adaptive systems in drug design.* Landes Bioscience, Georgetown.
426. Ekins S, Balakin KV, Savchuk N, Ivanenkov Y (2006) Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques. *J. Med. Chem.* **49**, 5059-5071.
427. Torgerson WS (1952) Multidimensional scaling: I. Theory and method. *Psychometrika* **17**, 401-419.
428. Agrafiotis DK, Xu H (2003) A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.* **43**, 475-484.
429. Agrafiotis DK, Xu H, Zhu F, Bandyopadhyay D, Liu P (2010) Stochastic proximity embedding: Methods and applications. *Mol. Inf.* **29**, 758-770.
430. Reisine T, Bell GI (1995) Molecular biology of somatostatin receptors. *Endocr. Rev.* **16**, 427-442.
431. Ösapay G, Ösapay K (1998) Therapeutic applications of somatostatin analogues. *Expert Opin. Ther. Pat.* **8**, 855-870.
432. Martin RE, Green LG, Guba W, Kratochwil N, Christ A (2007) Discovery of the first nonpeptidic, small-molecule, highly selective somatostatin receptor subtype 5 antagonists: A chemogenomics approach. *J. Med. Chem.* **50**, 6291-6294.
433. Guba W, Green LG, Martin RE, Roche O, Kratochwil N, Mauser H, Bissantz C, Christ A, Stahl M (2007) From astemizole to a novel hit series of small-molecule somatostatin 5 receptor antagonists via GPCR affinity profiling. *J. Med. Chem.* **50**, 6295-6298.
434. Martin RE, Mohr P, Maerki HP, Guba W, Kuratli C, Gavelle O, Binggeli A, Bendels S, Alvarez-Sánchez R, Alker A (2009) Benzoxazole piperidines as selective and potent somatostatin receptor subtype 5 antagonists. *Bioorg. Med. Chem. Lett.* **19**, 6106-6113.
435. Alker A, Binggeli A, Christ AD, Green L, Maerki HP, Martin RE, Mohr P (2010) Piperidinyl-nicotinamides as potent and selective somatostatin receptor subtype 5 antagonists. *Bioorg. Med. Chem. Lett.* **20**, 4521-4525.
436. Fechner U, Schneider G (2004) Optimization of a pharmacophore-based correlation vector descriptor for similarity searching. *QSAR Comb. Sci.* **23**, 19-22.
437. Van der Maaten LJP, Postma EO, Van Den Herik HJ (2009) Dimensionality reduction: A comparative review. *J. Mach. Learn. Res.* **10**, 1-41.
438. Nadaraya EA (1964) On estimating regression. *Theory Probab. Appl.* **9**, 141-142.
439. Watson GS (1964) Smooth regression analysis. *Sankhya Ser. A* 359-372.
440. Wand MP, Jones MC (1993) Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Am. Stat. Assoc.* **88**, 520-528.
441. Scott DW (2009) *Multivariate density estimation: Theory, practice, and visualization.* Wiley & Sons, New York.
442. Sain SR (1994) Adaptive kernel density estimation.
443. Loftsgaarden DO, Quesenberry CP (1965) A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.* **36**, 1049-1051.
444. Lee JJA, Verleysen M (2007) *Nonlinear dimensionality reduction.* Springer, New York.
445. Chen L, Buja A (2006) Local multidimensional scaling for nonlinear dimension reduction, graph layout and proximity analysis.
446. Sprecher U, Mohr P, Martin RE, Maerki HP, Sanchez RA, Binggeli A, Kunnecke B, Christ AD (2010) Novel, non-peptidic somatostatin receptor subtype 5 antagonists improve glucose tolerance in rodents. *Regul. Pept.* **159**, 19-27.
447. Kansy M, Senner F, Gubernator K (1998) Physicochemical high throughput screening: Parallel artificial membrane permeation assay in the description of passive absorption processes. *J. Med. Chem.* **41**, 1007-1010.
448. Alsenz J, Kansy M (2007) High throughput solubility measurement in drug discovery and development. *Adv. Drug. Deliv. Rev.* **59**, 546-567.
449. Peters JU, Schnider P, Mattei P, Kansy M (2009) Pharmacological promiscuity: Dependence on compound properties and target specificity in a set of recent Roche compounds. *ChemMedChem* **4**, 680-686.

450. Pammolli F, Magazzini L, Riccaboni M (2011) The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.* **10**, 428-438.
451. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (2010) How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203-214.
452. Willett P (2011) Similarity searching using 2D structural fingerprints. *Methods Mol. Biol.* **672**, 133-158.
453. Gardiner EJ, Holliday JD, O'Dowd C, Willett P (2011) Effectiveness of 2D fingerprints for scaffold hopping. *Future Med. Chem.* **3**, 405-414.
454. Renner S, Schneider G (2006) Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* **1**, 181-185.
455. Garcia-Serna R, Mestres J (2010) Anticipating drug side effects by comparative pharmacology. *Expert Opin. Drug Metab. Toxicol.* **6**, 1253-1263.
456. Areias FM, Brea J, Gregori-Puigjané E, Zaki MEA, Carvalho MA, Domínguez E, Gutiérrez-de-Terán H, Proença MF, Loza MI, Mestres J (2010) *In silico* directed chemical probing of the adenosine receptor family. *Bioorg. Med. Chem.* **18**, 3043-3052.
457. Gregori-Puigjané E, Mestres J (2008) Coverage and bias in chemical library design. *Curr. Opin. Chem. Biol.* **12**, 359-365.
458. Koutsoukas A, Simms B, Kirchmair J, Bond PJ, Whitmore AV, Zimmer S, Young MP, Jenkins JL, Glick M, Glen RC (2011) From *in silico* target prediction to multi-target drug design: Current databases, methods and applications. *J. Proteomics* **74**, 2554-2574.
459. Wassermann AM, Bajorath J (2011) BindingDB and ChEMBL: Online compound databases for drug discovery. *Expert Opin. Drug Discov.* **6**, 683-687.
460. Kramer C, Kalliokoski T, Gedeck P, Vulpetti A (2012) The experimental uncertainty of heterogeneous public K_i data. *J. Med. Chem.* **55**, 5165-5173.
461. MACCS-II; MDL Information Systems/Symyx: Santa Clara, CA. acc.
462. Fingerprint Toolkit, Chapter 6: *Fingerprints - screening and similarity*. Daylight Chemical Information Systems Inc. <http://www.daylight.com/products/finger_kit.html> acc. 28 Nov. 2013.
463. RDKit: Open-source cheminformatics. <http://www.*rdkit*.org> acc.
464. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887-2893.
465. Holliday JD, Kanoulas E, Malim N, Willett P (2011) Multiple search methods for similarity-based virtual screening: Analysis of search overlap and precision. *J. Cheminform.* **3**, 29.
466. Wilcoxon F (1945) Individual comparisons by ranking methods. *Biometrics Bull.* **1**, 80-83.
467. Ward Jr JH (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236-244.
468. Klenner A, Hartenfeller M, Schneider P, Schneider G (2010) 'Fuzziness' in pharmacophore-based virtual screening and *de novo* design. *Drug Discov. Today Technol.* **7**, e237-e244.
469. Schneider G (2013) *De novo* design - hop(p)ing against hope. *Drug Discov. Today Technol.* **10**, e453-e460.
470. Bienayme H, Bouzid K (1998) A new heterocyclic multicomponent reaction for the combinatorial synthesis of fused 3-aminoimidazoles. *Angew. Chem. Int. Ed.* **37**, 2234-2237.
471. Hähnke V, Todoroff N, Rodrigues T, Schneider G (2012) Significance estimation for sequence-based chemical similarity searching (PhAST) and application to AuroraA kinase inhibitors. *Future Med. Chem.* **4**, 1897-1906.
472. Kim O, Jeong Y, Lee H, Hong S-S, Hong S (2011) Design and synthesis of imidazopyridine analogues as inhibitors of phosphoinositide 3-kinase signaling and angiogenesis. *J. Med. Chem.* **54**, 2455-2466.

473. Yu H, Ai Y, Yu L, Zhou X, Liu J, Li J, Xu X, Liu S, Chen J, Liu F (2008) Phosphoinositide 3-kinase/Akt pathway plays an important role in chemoresistance of gastric cancer cells against etoposide and doxorubicin induced cell death. *Int. J. Cancer* **122**, 433-443.
474. Baviskar AT, Madaan C, Preet R, Mohapatra P, Jain V, Agarwal A, Guchhait SK, Kundu CN, Banerjee UC, Bharatam PV (2011) N-fused imidazoles as novel anticancer agents that inhibit catalytic activity of topoisomerase IIα and induce apoptosis in G1/S phase. *J. Med. Chem.* **54**, 5013-5030.
475. Coan KED, Shoichet BK (2008) Stoichiometry and physical chemistry of promiscuous aggregate-based inhibitors. *J. Am. Chem. Soc.* **130**, 9606-9612.
476. Noeske T, Jirgensons A, Starchenkovs I, Renner S, Jaunzeme I, Trifanova D, Hechenberger M, Bauer T, Kauss V, Parsons CG, Schneider G, Weil T (2007) Virtual screening for selective allosteric mGluR1 antagonists and structure-activity relationship investigations for coumarine derivatives. *ChemMedChem* **2**, 1763-1773.
477. Steri R, Achenbach J, Steinhilber D, Schubert-Zsilavecz M, Proschak E (2012) Investigation of imatinib and other approved drugs as starting points for antidiabetic drug discovery with FXR modulating activity. *Biochem. Pharmacol.* **83**, 1674-1681.
478. Reutlinger M, Guba W, Martin RE, Alanine AI, Hoffmann T, Klenner A, Hiss JA, Schneider P, Schneider G (2011) Neighborhood-preserving visualization of adaptive structure-activity landscapes: Application to drug discovery. *Angew. Chem. Int. Ed.* **50**, 11633-11636.
479. Sur C, Mallorga PJ, Wittmann M, Jacobson MA, Pascarella D, Williams JB, Brandish PE, Pettibone DJ, Scolnick EM, Conn PJ (2003) N-desmethylclozapine, an allosteric agonist at muscarinic 1 receptor, potentiates N-methyl-D-aspartate receptor activity. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13674-13679.
480. SciFinder 2012. Chemical Abstracts Service; RN 58-08-02. <http://www.cas.org/products/scifinder> acc. Nov. 9, 2012.
481. Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, Green DVS, Hertzberg RP, Janzen WP, Paslay JW (2011) Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* **10**, 188-195.
482. Bryan MC, Hein CD, Gao H, Xia X, Eastwood H, Bruenner BA, Louie SW, Doherty EM (2013) Disubstituted 1-Aryl-4-Aminopiperidine Library Synthesis Using Computational Drug Design and High-Throughput Batch and Flow Technologies. *ACS Comb. Sci.* **15**, 503-511.
483. Ugi I (1962) The α-addition of immonium ions and anions to isonitriles accompanied by secondary reactions. *Angew. Chem. Int. Ed.* **1**, 8-21.
484. Beck B, Srivastava S, Khoury K, Herdtweck E, Dömling A (2010) One-pot multicomponent synthesis of two novel thiolactone scaffolds. *Mol. Divers.* **14**, 479-491.
485. Kalinski C, Lemoine H, Schmidt J, Burdack C, Kolb J, Umkehrer M, Ross G (2008) Multicomponent reactions as a powerful tool for generic drug synthesis. *Synthesis* **2008**, 4007-4011.
486. Reutlinger M, Koch CP, Reker D, Todoroff N, Schneider P, Rodrigues T, Schneider G (2013) Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for 'orphan' molecules. *Mol. Inf.* **32**, 133-138.
487. Hieke M, Rödl CB, Wisniewska JM, la Buscató E, Stark H, Schubert-Zsilavecz M, Steinhilber D, Hofmann B, Proschak E (2012) SAR-study on a new class of imidazo [1, 2-a] pyridine-based inhibitors of 5-lipoxygenase. *Bioorg. Med. Chem. Lett.* **22**, 1969-1975.
488. Meister S, Plouffe DM, Kuhen KL, Bonamy GMC, Wu T, Barnes SW, Bopp SE, Borboa R, Bright AT, Che J (2011) Imaging of Plasmodium liver stages to drive next-generation antimalarial drug discovery. *Science* **334**, 1372-1377.
489. Wu T, Nagle A, Kuhen K, Gagaring K, Borboa R, Francek C, Chen Z, Plouffe D, Goh A, Lakshminarayana SB (2011) Imidazolopiperazines: Hit to lead optimization of new antimalarial agents. *J. Med. Chem.* **54**, 5116-5130.
490. Wisniewska JM, Rodl CB, Kahnt AS, Buscato E, Ulrich S, Tanrikulu Y, Achenbach J, Rorsch F, Grosch S, Schneider G, Cinatl JJ, Proschak E, Steinhilber D, Hofmann B (2012)

Molecular characterization of EP6 - a novel imidazo[1,2-a]pyridine based direct 5-lipoxygenase inhibitor. *Biochem. Pharmacol.* **83**, 228-240.
491. Evans BE, Rittle KE, Bock MG, DiPardo RM, Freidinger RM, Whitter WL, Lundell GF, Veber DF, Anderson PS (1988) Methods for drug discovery: Development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **31**, 2235-2246.
492. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2007) *KNIME: The Konstanz information miner.* In: *Classification, Data Analysis, and Knowledge Organization (GfKL)* Springer, Berlin, 319-326.
493. Chen JH, Linstead E, Swamidass SJ, Wang D, Baldi P (2007) ChemDB update — full-text search and virtual chemical space. *Bioinformatics* **23**, 2348-2351.
494. Cordeaux Y, Ijzerman AP, Hill SJ (2004) Coupling of the human A_1 adenosine receptor to different heterotrimeric G proteins: Evidence for agonist-specific G protein activation. *Br. J. Pharmacol.* **143**, 705-714.
495. Cooper J, Hill SJ, Alexander SP (1997) An endogenous A2B adenosine receptor coupled to cyclic AMP generation in human embryonic kidney (HEK 293) cells. *Br. J. Pharmacol.* **122**, 546-550.
496. Vicentic A (2002) Biochemistry and pharmacology of epitope-tagged alpha 1-adrenergic receptor subtypes. *J. Pharmacol. Exp. Ther.* **302**, 58-65.
497. Schwinn DA, Johnston GI, Page SO, Mosley MJ, Wilson KH, Worman NP, Campbell S, Fidock MD, Furness LM, Parry-Smith DJ (1995) Cloning and pharmacological characterization of human alpha-1 adrenergic receptors: Sequence corrections and direct comparison with other species homologues. *J. Pharmacol. Exp. Ther.* **272**, 134-142.
498. Townsend-Nicholson A, Schofield PR (1994) A threonine residue in the seventh transmembrane domain of the human A_1 adenosine receptor mediates specific agonist binding. *J. Biol. Chem.* **269**, 2373-2376.
499. Antolín AA, Mestres J (2012) *Knowledge base for nuclear receptor drug discovery.* In: *Therapeutic Targets: Modulation, Inhibition, and Activation* John Wiley & Sons, New York, 309-326.
500. Hu Y, Bajorath J (2013) What is the likelihood of an active compound to be promiscuous? Systematic assessment of compound promiscuity on the basis of PubChem confirmatory bioassay data. *AAPS J.* **15**, 808-815.
501. Roy K, Mitra I (2011) On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb. Chem. High Throughput Screening* **14**, 450-474.
502. Hann MM, Keserü GM (2012) Finding the sweet spot: The role of nature and nurture in medicinal chemistry. *Nat. Rev. Drug Discov.* **11**, 355-365.
503. Nge PN, Rogers CI, Woolley AT (2013) Advances in microfluidic materials, functions, integration, and applications. *Chem. Rev.* **113**, 2550-2583.
504. Lombardi D, Dittrich PS (2010) Advances in microfluidics for drug discovery. *Expert Opin. Drug Discov.* **5**, 1081-1094.
505. Tsukahara T, Mawatari K, Kitamori T (2010) Integrated extended-nano chemical systems on a chip. *Chem. Soc. Rev.* **39**, 1000-1013.
506. Dorigo M, Di Caro G (1999) Ant colony optimization: A new meta-heuristic. *Proc. CEC* **2**, 1470-1477.
507. Dorigo M, Blum C (2005) Ant colony optimization theory: A survey. *Theor. Comput. Sci.* **344**, 243-278.
508. Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M (2010) Synthetic therapeutic peptides: Science and market. *Drug Discov. Today* **15**, 40-56.
509. Audie J, Boyd C (2010) The synergistic use of computation, chemistry and biology to discover novel peptide-based drugs: The time is right. *Curr. Pharm. Des.* **16**, 567-582.
510. Fjell CD, Hiss JA, Hancock REW, Schneider G (2011) Designing antimicrobial peptides: Form follows function. *Nat. Rev. Drug Discov.* **11**, 37-51.

511. Rentero I, Heinis C (2011) Screening of large molecule diversities by phage display. *Chimia* **65**, 843-845.
512. Pande J, Szewczyk MM, Grover AK (2010) Phage display: Concept, innovations, applications and future. *Biotechnol. Adv.* **28**, 849-858.
513. Hiss JA, Hartenfeller M, Schneider G (2010) Concepts and applications of natural computing techniques in *de novo* drug and peptide design. *Curr. Pharm. Des.* **16**, 1656-1665.
514. Daeyaert F, De Jonge M, Koymans L, Vinkers M (2007) An ant algorithm for the conformational analysis of flexible molecules. *J. Comput. Chem.* **28**, 890-898.
515. Stützle T, Hoos HH (2000) MAX-MIN ant system. *Fut. Gener. Comput. Sys.* **16**, 889-914.
516. Dorigo M, Maniezzo V, Colorni A (1996) Ant system: Optimization by a colony of cooperating agents. *IEEE Trans. Syst. Man. Cybern. B Cybern.* **26**, 29-41.
517. Dorigo M, Gambardella LM (1997) Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE Trans. Evol. Computat.* **1**, 53-66.
518. Koch CP, Perna AM, Weissmüller S, Bauer S, Pillong M, Baleeiro RB, Reutlinger M, Folkers G, Walden P, Wrede P, Schneider G (2013) Exhaustive proteome mining for functional MHC-I ligands. *ACS Chem. Biol.* **8**, 1876-1881.
519. Koch CP, Perna AM, Pillong M, Todoroff NK, Wrede P, Folkers G, Hiss JA, Schneider G (2013) Scrutinizing MHC-I binding peptides and their limits of variation. *PLoS Comput. Biol.* **9**, e1003088.
520. Van Der Maaten LJP, Postma EO, Van Den Herik HJ (2007) Matlab toolbox for dimensionality reduction. *Proc. BNAIC* 439-440.
521. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188-1190.
522. Udaka K, Wiesmüller K-H, Kienle S, Jung G, Walden P (1995) Tolerance to amino acid variations in peptides binding to the major histocompatibility complex class I protein H-2K^{b}. *J. Biol. Chem.* **270**, 24130-24134.
523. Ljunggren HG, Kärre K (1985) Host resistance directed selectively against H-2-deficient lymphoma variants. Analysis of the mechanism. *J. Exp. Med.* **162**, 1745-1759.
524. Rammensee HG, Bachmann J, Emmerich NPN, Bachor OA, Stevanović S (1999) SYFPEITHI: Database for MHC ligands and peptide motifs. *Immunogenetics* **50**, 213-219.
525. Reutlinger M, Schneider G (2012) Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J. Mol. Graph. Model.* **34**, 108-117.
526. Schneider TD, Stephens RM (1990) Sequence logos: A new way to display consensus sequences. *Nucl. Acids. Res.* **18**, 6097-6100.
527. Brock R, Wiesmüller K-H, Jung G, Walden P (1996) Molecular basis for the recognition of two structurally different major histocompatibility complex/peptide complexes by a single T-cell receptor. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13108-13113.
528. Rötzschke O, Falk K, Stevanovic S, Jung G, Walden P, Rammensee H (1991) Exact prediction of a natural T cell epitope. *Eur. J. Immunol.* **21**, 2891-2894.
529. Wisniewska JM, Jäger N, Freier A, Losch FO, Wiesmüller K-H, Walden P, Wrede P, Schneider G, Hiss JA (2010) MHC I stabilizing potential of computer-designed octapeptides. *J. Biomed. Biotechnol.* **2010**, 396847.
530. Ugi I, Meyr R, Fetzer U, Steinbrückner C (1959) Versuche mit Isonitrilen. *Angew. Chem.* **71**, 386.
531. Pan SC, List B (2008) Catalytic three-component Ugi reaction. *Angew. Chem. Int. Ed.* **47**, 3622-3625.
532. Yokobayashi Y, Ikebukuro K, McNiven S, Karube I (1996) Directed evolution of trypsin inhibiting peptides using a genetic algorithm. *J. Chem. Soc., Perkin Trans.* **1**, 2435-2437.
533. Weber L, Wallbaum S, Broger C, Gubernator K (1995) Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm. *Angew. Chem. Int. Ed.* **34**, 2280-2282.

534. Fechner U, Schneider G (2006) Flux (1): A virtual synthesis scheme for fragment-based *de novo* design. *J. Chem. Inf. Model.* **46**, 699-707.
535. Riester D, Wirsching F, Salinas G, Keller M, Gebinoga M, Kamphausen S, Merkwirth C, Goetz R, Wiesenfeldt M, Sturzebecher J, Bode W, Friedrich R, Thurk M, Schwienhorst A (2005) Thrombin inhibitors identified by computer-assisted multiparameter design. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8597-8602.
536. Sinauridze EI, Romanov AN, Gribkova IV, Kondakova OA, Surov SS, Gorbatenko AS, Butylin AA, Monakov MY, Bogolyubov AA, Kuznetsov YV, Sulimov VB, Ataullakhanov FI (2011) New synthetic thrombin inhibitors: Molecular design and experimental verification. *PLoS ONE* **6**, e19969.
537. Illgen K, Enderle T, Broger C, Weber L (2000) Simulated molecular evolution in a full combinatorial library. *Chem. Biol.* **7**, 433-441.
538. Illgen K, Nerdinger S, Fuchs T, Friedrich C, Weber L, Herdtweck E (2004) A versatile synthesis of 6-oxo-1,4,5,6-tetrahydro-pyrazine-2-carboxylic acid methyl esters via MCR chemistry. *Synlett* **15**, 53-56.
539. Schüller A, Schneider G (2008) Identification of hits and lead structure candidates with limited resources by adaptive optimization. *J. Chem. Inf. Model.* **48**, 1473-1491.
540. de Candia M, Lopopolo G, Altomare C (2009) Novel factor Xa inhibitors: A patent review. *Expert Opin. Ther. Pat.* **19**, 1535-1580.
541. Agrafiotis DK (1997) Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* **37**, 841-851.
542. Zheng W, Cho SJ, Waller CL, Tropsha A (1999) Rational combinatorial library design. 3. Simulated annealing guided evaluation (SAGE) of molecular diversity: a novel computational tool for universal library design and database mining. *J. Chem. Inf. Comput. Sci.* **39**, 738-746.
543. Engelbrecht AP (2005) *Fundamentals of computational swarm intelligence.* Wiley & Sons, New York.
544. Hiss JA, Bredenbeck A, Losch FO, Wrede P, Walden P, Schneider G (2007) Design of MHC I stabilizing peptides by agent-based exploration of sequence space. *Protein Eng. Des. Sel.* **20**, 99-108.
545. Jäger N, Wisniewska JM, Hiss JA, Freier A, Losch FO, Walden P, Wrede P, Schneider G (2010) Attractors in sequence space: Agent-based exploration of MHC I binding peptides. *Mol. Inf.* **29**, 65-74.
546. Long A (2012) Parallel chemistry in the 21st century. *Curr. Protoc. Pharmacology* **58**, 9.16. 1-9.16. 16.
547. Hann MM, Oprea TI (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **8**, 255-263.
548. Gillet VJ (2004) Designing combinatorial libraries optimized on multiple objectives. *Methods Mol. Biol.* **275**, 335-354.
549. Fang G, Xue M, Su M, Hu D, Li Y, Xiong B, Ma L, Meng T, Chen Y, Li J (2012) CCLab — a multi-objective genetic algorithm based combinatorial library design software and an application for histone deacetylase inhibitor design. *Bioorg. Med. Chem. Lett.* **22**, 4540-4545.
550. Reutlinger R, Rodrigues T, Schneider P, Schneider G (2014) Combining on-chip synthesis of a focused combinatorial library with *in silico* target prediction reveals imidazopyridine GPCR ligands. *Angew. Chem. Int. Ed.* **53**, 582-585.
551. Hiss JA, Reutlinger M, Koch CP, Perna AM, Schneider P, Rodrigues T, Haller S, Folkers G, Weber L, Baleeiro RB, Walden P, Wrede P, Schneider G (2014) Combinatorial chemistry by ant colony optimization. *Future Med. Chem.* **6**, 267-280.
552. Swamidass SJ, Chen J, Bruand J, Phung P, Ralaivola L, Baldi P (2005) Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics* **21**, i359-i368.
553. Reisen FH, Schneider G, Proschak E (2009) Reaction-MQL: Line notation for functional transformation. *J. Chem. Inf. Model.* **49**, 6-12.

554. Lee JA, Verleysen M (2009) Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing* **72**, 1431-1443.
555. Harper G (1999) The selection of compounds for screening in pharmaceutical research.
556. Chen B, Harrison RF, Pasupa K, Willett P, Wilton DJ, Wood DJ, Lewell XQ (2006) Virtual screening using binary kernel discrimination: Effect of noisy training data and the optimization of performance. *J. Chem. Inf. Model.* **46**, 478-486.
557. Zhou QY, Grandy DK, Thambi L, Kushner JA, Van Tol HH, Cone R, Pribnow D, Salon J, Bunzow JR, Civelli O (1990) Cloning and expression of human and rat D_1 dopamine receptors. *Nature* **347**, 76-80.
558. Grandy DK, Marchionni MA, Makam H, Stofko RE, Alfano M, Frothingham L, Fischer JB, Burke-Howie KJ, Bunzow JR, Server AC (1989) Cloning of the cDNA and gene for a human D_2 dopamine receptor. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 9762-9766.
559. MacKenzie RG, VanLeeuwen D, Pugsley TA, Shih Y-H, Demattos S, Tang L, Todd RD, O'Malley KL (1994) Characterization of the human dopamine D_3 receptor expressed in transfected cell lines. *Eur. J. Pharmacol.* **266**, 79-85.
560. Van Tol HHM, Wu CM, Guan H-C, Ohara K, Bunzow JR, Civelli O, Kennedy J, Seeman P, Niznik HB, Jovanovic V (1992) Multiple dopamine D_4 receptor variants in the human population. *Nature* **358**, 149-152.
561. Sunahara RK, Guan H-C, O'Dowd BF, Seeman P, Laurier LG, Ng G, George SR, Torchia J, Van Tol HHM, Niznik HB (1991) Cloning of the gene for a human dopamine D_5 receptor with higher affinity for dopamine than D_1. *Nature* **350**, 614-619.
562. Ganapathy ME, Prasad PD, Huang W, Seth P, Leibach FH, Ganapathy V (1999) Molecular and ligand-binding characterization of the ç-receptor in the jurkat human T lymphocyte cell line. *J. Pharmacol. Exp. Ther.* **289**, 251-260.
563. Lovenberg TW, Roland BL, Wilson SJ, Jiang X, Pyati J, Huvar A, Jackson MR, Erlander MG (1999) Cloning and functional expression of the human histamine H3 receptor. *Mol. Pharmacol.* **55**, 1101-1107.
564. Mulheron JG, Casanas SJ, Arthur JM, Garnovskaya MN, Gettys TW, Raymond JR (1994) Human 5-HT_{1A} receptor expressed in insect cells activates endogenous G (o)-like G protein(s). *J. Biol. Chem.* **269**, 12954-12962.
565. Simonin F, Befort K, Gaveriaux-Ruff C, Matthes H, Nappey V, Lannes B, Micheletti G, Kieffer B (1994) The human delta-opioid receptor: Genomic organization, cDNA cloning, functional expression, and distribution in human brain. *Mol. Pharmacol.* **46**, 1015-1021.
566. Meng F, Xie G-X, Thompson RC, Mansour A, Goldstein A, Watson SJ, Akil H (1993) Cloning and pharmacological characterization of a rat kappa opioid receptor. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 9954-9958.
567. Wang J-B, Johnson PS, Persico AM, Hawkins AL, Griffin CA, Uhl GR (1994) Human μ opiate receptor: cDNA and genomic clones, pharmacologic characterization and chromosomal assignment. *FEBS Lett.* **338**, 217-222.
568. Saubern S, Guha R, Baell JB (2011) KNIME workflow to assess PAINS filters in SMARTS format. Comparison of RDKit and indigo cheminformatics libraries. *Mol. Inf.* **30**, 847-850.
569. Lagorce D, Sperandio O, Galons H, Miteva MA, Villoutreix BO (2008) FAF-Drugs2: Free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinf.* **9**, 396.
570. Tarazi F, Baldessarini RJ (1999) Dopamine D_4 receptors: Significance for molecular psychiatry at the millennium. *Mol. Psychiatry* **4**, 529-538.
571. Arnsten AFT, Murphy B, Merchant K (2000) The selective dopamine D_4 receptor antagonist, PNU-101387G, prevents stress-induced cognitive deficits in monkeys. *Neuropsychopharmacol.* **23**, 405-410.
572. Löber S, Hübner H, Tschammer N, Gmeiner P (2011) Recent advances in the search for D_3-and D_4-selective drugs: Probes, models and candidates. *Trends Pharmacol. Sci.* **32**, 148-157.

573. Giusti-Rodríguez P, Sullivan PF (2013) The genomics of schizophrenia: update and implications. *J. Clin. Invest.* **123**, 4557-4563.
574. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319-2323.
575. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**, 1373-1396.
576. van der Maaten L (2013) Barnes-Hut-SNE. *arXiv preprint arXiv:1301.3342*
577. Hansen K, Baehrens D, Schroeter T, Rupp M, Müller K-R (2011) Visual interpretation of kernel-based prediction models. *Mol. Inf.* **30**, 817-826.
578. Nambi P, Aiyar N (2003) G protein-coupled receptors in drug discovery. *Assay Drug Dev. Technol.* **1**, 305-310.
579. Drews J, Ryser S (1997) Classic drug targets. *Nat. Biotechnol.* **15**, 1350-1350.
580. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? *Nat. Rev. Drug Discov.* **5**, 993-996.
581. Meslamani J, Bhajun R, Martz F, Rognan D (2013) Computational profiling of bioactive compounds using a target-dependent composite workflow. *J. Chem. Inf. Model.* **53**, 2322-2333.
582. Wassermann AM, Geppert H, Bajorath J (2009) Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. *J. Chem. Inf. Model.* **49**, 2155-2167.
583. Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF, Babu MM (2013) Molecular signatures of G-protein-coupled receptors. *Nature* **494**, 185-194.
584. Tan L, Geppert H, Sisay MT, Gutschow M, Bajorath J (2008) Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. *ChemMedChem* **3**, 1566-1571.
585. von Korff M, Freyss J, Sander T (2009) Comparison of ligand- and structure-based virtual screening on the DUD data set. *J. Chem. Inf. Model.* **49**, 209-231.
586. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH (2012) Structure-based virtual screening for drug discovery: A problem-centric review. *AAPS J.* **14**, 133-141.
587. Nicholls A (2012) The character of molecular modeling. *J. Comput. Aided Mol. Des.* **26**, 103-105.
588. Teague SJ (2003) Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* **2**, 527-541.
589. Monod J, Wyman J, Changeux J-P (1965) On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12**, 88-118.
590. Lange OF, Lakomek NA, Fares C, Schroder GF, Walter KF, Becker S, Meiler J, Grubmuller H, Griesinger C, de Groot BL (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **320**, 1471-1475.
591. Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789-796.
592. Changeux JP, Edelstein S (2011) Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol. Rep.* **3**, 19.
593. Koshland DE (1958) Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **44**, 98-104.
594. Tuffery P, Derreumaux P (2012) Flexibility and binding affinity in protein-ligand, protein-protein and multi-component protein interactions: limitations of current computational approaches. *J. R. Soc. Interface.* **9**, 20-33.
595. Homans SW (2007) Water, water everywhere--except where it matters? *Drug. Discov. Today* **12**, 534-539.

8 Appendix

8.1 Perspective – Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery

8.1.1 Abstract

Visualization of "chemical space" and compound distributions has received much attraction by medicinal chemists, as it may help to intuitively comprehend pharmaceutically relevant molecular features. It has been realized that for meaningful feature extraction from complex multivariate chemical data, such as compound libraries represented by many molecular descriptors, nonlinear projection techniques are required. Recent advances in machine-learning and artificial intelligence have resulted in a transfer of such methods to chemistry. We provide an overview of prominent visualization methods based on nonlinear dimensionality reduction, and highlight applications in drug discovery. Emphasis is on neural network techniques, kernel methods and stochastic embedding approaches, which have been successfully used for ligand-based virtual screening, SAR landscape analysis, combinatorial library design, and screening compound selection.

8.1.2 Introduction

The first modern atlas of the world, the "*Typvs Orbis Terrarvm*", was published in 1570 employing the Mercator projection of the globe (Figure S1). There is no doubt that a two-dimensional (2D) map of the three-dimensional (3D) surface of the earth not only facilitated traveling from one place to the other, but more generally shaped our modern perception of the world. Similarly, it can be helpful to visualize chemical data in two dimensions, so that visual "navigation in chemical space" becomes possible. Visualization of compound distributions presents complex data in a simpler form[S1,S2]. By focusing on intrinsic dimensions of chemical data, relationships between compounds may be graphically displayed to inspire medicinal chemists and support hit finding and lead structure prioritization in drug discovery[S1,S3-S5]. By "compound library" we here refer to a defined set of compounds, *e.g.* drug-like

molecules or a combinatorial compound collection under investigation, rather than the whole universe of stable chemical structures. In this review, we present some of the essential mathematical concepts and motivate the use of nonlinear projection methods for vectorial numerical chemical data, which is obtained from representations of compounds by molecular descriptors like structural fingerprints, pharmacophoric features, or physicochemical properties. For an extensive overview of applications and practical examples of visualization in early drug discovery we refer to an excellent review article by Balakin and coworkers [S4].

Figure S1. World Map "*Typvs Orbis Terrarvm*" (A. Ortelius, 1570; source: The Library of Congress, Washington DC, USA).

Often, the number of descriptors *d* used to encode molecular structure and properties exceeds the number of uncorrelated features by far, and dimensionality reduction and feature extraction methods are applied so that fewer "meaningful" descriptors or descriptor combinations are found. In other words, most multivariate compound data in $\mathbb{R}^d$ are not truly *d*-dimensional but form patterns on a lower-dimensional manifold[S6]. In the context of this study, we refer to such a lower-dimensional molecular representation as a "projection" of data from a high-dimensional pattern space to a low-dimensional feature space $X \rightarrow X'$.

This concept of low-dimensional virtual screening and chemical data analysis is further motivated by several observations that can be made for high-dimensional chemical descriptor spaces[S7]. With d approaching infinity, one encounters:

1. The *empty space phenomenon*: An exponential number of samples is needed to cover $\mathbb{R}^d$ in the sense that each dimension contains at least two compounds. In typical drug discovery scenarios, the chemical space spanned by a descriptor will be empty in terms of dataset coverage. For example, the maximum dimension that could be covered by a compound collection of one million compounds is as low as $\left\lfloor \log_2\left(10^6\right) \right\rfloor = 20$.
2. *Vanishing sphere volumes*: The volume of a d-dimensional Euclidean sphere with radius r becomes zero for $d \to \infty$. As a practical consequence for compound data from $\mathbb{R}^d$, there is a dimension d after which a sphere of radius r centered on compound x_i contains only x_i and no other sample. In other words, with increasing dimension of the molecular representation, the probability mass contained in a sphere with fixed radius around a compound decreases rapidly.
3. *Distance concentration*: Sample norms tend to concentrate and as a consequence, all distances are similar, samples lie on a hyper sphere, and, each compound is nearest neighbor of all other compounds.

Consequences for virtual screening and the analysis of multivariate chemical data, including compound ranking and clustering, bear the danger of leading to erroneous results and consequently misinterpretation. We therefore motivate data visualization and dimensionality reduction methods as potentially very useful for hit finding and hit-to-lead optimization in early drug discovery.

8.1.3 Results and Discussion

Principal Component Analysis

Principal Component Analysis (PCA) is a linear dimension reduction method and belongs to the class of spectral dimension reduction methods. The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of their variation

(variance). This is achieved by transforming the data to a new set of uncorrelated variables, the *Principle Components* (PCs), which are ordered, such that the first few retain most of the variation present in all of the original variables[S8]. PCs are linear combinations of the original descriptor axes and represent those directions in data space along which the scatter of the data is greatest. PCA has found widespread application in molecular modeling and drug design, and it is common practice to visualize compound distributions in graphical displays using the first two or three principal components[S9].

PCA is performed by determining the eigenvectors and eigenvalues of the covariance matrix, or approximated values in case of large data matrices. The covariance of two random variables is their tendency to vary together. Suppose we have *n* independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of a *p*-dimensional random variable (feature vector, molecular descriptor vector). The sample covariance matrix $\mathbf{S}$ is given by Eq. S1, with $\bar{\mathbf{x}}$ being the sample mean and the superscript T denotes the matrix transpose.

$$\mathbf{S} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T . \quad \text{(S1)}$$

In the case of centered data with $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ the covariance matrix can be rewritten (Eq. S2):

$$\tilde{\mathbf{S}} = \frac{1}{n-1}\sum_{k=1}^{n}\tilde{\mathbf{x}}_k\tilde{\mathbf{x}}_k{}^T \quad \text{(S2)}$$

By solving the covariance matrix $\tilde{\mathbf{S}}$ for eigenvectors $\mathbf{a} \in \mathbb{R}^p$ and eigenvalues $\lambda \in \mathbb{R}_{\geq 0}$, subject to the constraint that $\mathbf{a}^T\mathbf{a} = 1$,we can find the projection directions (Eq. S3).

$$\tilde{\mathbf{S}}\mathbf{a} = \lambda\mathbf{a} . \quad \text{(S3)}$$

Data points $\tilde{\mathbf{x}}_i$ are transformed into the new PC coordinate system by orthogonal projection of the data points on each of the eigenvectors (Eq. S4).

$$\tilde{\boldsymbol{z}}_i = \tilde{\mathbf{x}}_i\mathbf{A} , \quad \text{(S4)}$$

where $\mathbf{A}$ is the orthogonal ($p \times p$) matrix with the eigenvectors as columns. Eigenvectors are sorted by decreasing eigenvalues, which can be interpreted as their "significance". To obtain a lower dimensional projection of the original data eigenvectors with small eigenvalues are omitted from $\mathbf{A}$.

Ten years ago, Oprea and coworkers presented the *Chemical Global Positioning System* (ChemGPS) in combination with PCA for the linear projection of chemical data[S10]. Its main feature is a set of "satellite" compounds that are placed outside druglike space and thereby define outer data borders, and consequently, the applicability domain of the projection. The original application employed a total of 423 satellites and representative drugs ("core structures"). This concept of defining the borders of a chemical space by extreme-valued compounds can help identify projection artifacts and prevent unjustified conclusions from inspection and interpretation of chemical space maps and is not limited to PCA. In fact, it is recommended to use a basis set of reference cores and satellites for any projection of compound distributions. With the advent of large open access repositories and searchable databases of bioactive compounds – e.g. ChEMBL[S11], PubChem[S12], ChemBank[S13], ChEBI[S14], ChemDB[S15] – the known bioactive chemical space is continuously extended and refined[S16]. This huge body of chemical structures and literature data will help in defining appropriate boundaries of druglike chemical space[S17]. Despites its appeal there are certain limitations of PCA that motivate nonlinear projection techniques to be used complementarily, *e.g.* the requirement for normal-distributed data, susceptibility to outliers, and issues with data manifolds and large data sets, to just name some prominent examples. In drug discovery one is mainly interested in the structure of local neighborhoods of known bioactive compounds or reference molecules,[S18] and the *Chemical Similarity Principle*[S19] is grounded on the neighborhood concept[S20,S21]. PCA *per se* does not preserve local structure of the input data in the projection. In contrast, nonlinear embedding techniques do not assume global linearity but make a weaker local linearity assumption. In high-dimensional input space the Euclidian (L_2 norm) distance is assumed to be a good measure of geodesic distance (*vide infra*) only for nearby points, which is also observed for chemical data suffering from the "curse of dimensionality"[S22].

We often face nonlinearity in structure-activity relationship (SAR) modeling, which manifest as perceived "activity cliffs", that is, when structurally similar (nearby) compounds exhibit a significantly different pharmacological or other measured effect. Seemingly, the *Chemical Similarity Principle* does not hold in these regions of chemical

space. This assumption may not be true, as the measured effect of a considered small change of chemical structure, *e.g.* an exchange of methyl by ethyl, can be dramatic if an essential, function-determining molecular feature (pharmacophore) is destroyed[S23,S24]. In order to account for nonlinearity, specifically its relation to some observable function or property, one can either conceive appropriate, context-dependent molecular descriptors that restore global linearity, or apply nonlinear, local neighborhood preserving embedding methods that are able to capture manifolds in high-dimensional chemical data[S25].

Numerous nonlinear projection methods – expressly manifold learning techniques such as *Local Linear Embedding* (LLE)[S26], the IsoMap[S27] approach and its derivatives Laplacian IsoMap[S28] and Kernel IsoMap[S29] – have found widespread application in natural sciences, in particular bioinformatics[S30-S33], but mainly outside of chemistry. Some of these methods may be considered as specific instances of Kernel PCA[S34,S35], which employs the "kernel trick" to perform conventional linear PCA not in the original input space *X* (*i.e.* the original molecular descriptors), but in a virtual, very high-dimensional Hilbert space *V*, so that nonlinear relationships in *X* will gain linear meaning in *V*. In other words, the kernel trick virtually increases the dimensionality of the input data so that, in this higher dimension space, they become linearly related. The particular appeal of kernel methods is that the Hilbert space is never explicitly generated by transforming the original data into that space, but implicitly computed using a kernel function Φ. The *Support Vector Machine* (SVM) represents a prominent machine-learning concept using the kernel trick. SVMs are most successfully employed for SAR modeling and classification of chemical data[S36]. More recently, kernel-based Gaussian Process modeling has been introduced to chemistry and drug discovery[S37,S38]. While kernel methods provide an elegant approach for nonlinear SAR modeling and classification, they do not explicitly provide a means for data visualization and interpretation of complex nonlinear models. With few exceptions of chemical feature extraction, visualization and interpretation published[S39-S41], this might be a reason why kernel techniques have not found excessive appreciation in medicinal chemistry yet.

Encoder network

Special types of feed-forward artificial networks were among the first nonlinear methods employed for dimensionality reduction in chemistry[S42-S44]. One such system of particular interest is the symmetric encoder network (Figure 2)[S45,S46]. Here, the idea is to simultaneously compute a nonlinear forward- and a back-projection of chemical data. Within the applicability domain and under certain conditions[S47], this concept would allow one to navigate in the low-dimensional projection, and for each position of this map the coordinates (substructures, properties, or pharmacophoric features – depending on the molecular descriptor used) of compound in the original space are provided, thereby solving the "inverse QSAR problem"[S48,S49]. Despite its appeal only few applications of the encoder network approach have been published. One reason for its limited use might be the small number of ready-to-use software tools implementing such a system (one such free tool is the software *ChemSpaceShuttle*[S50]). One also needs to take into account long training times and the requirement for an optimization of network architecture, specifically the number of hidden neurons.

An encoder network must be trained in a supervised fashion, *i.e.* a forward (encoder) and a backward (decoder) function must be found, by using sets of reference compounds represented by a vectorial molecular descriptor **x**. Having defined the number of hidden neurons and the desired projection (number of central layer neurons), the weights of the encoder-decoder function defined by the network's architecture ($\mathbf{w}^{\text{Hidden}}$, $\mathbf{w}^{\text{Central}}$, $\mathbf{v}^{\text{Hidden}}$, $\mathbf{v}^{\text{Central}}$) are optimized so that for every input vector **x**, an identical output vector **y** is computed (Figure S2A).

Typically, variations of the delta rule in combination with gradient-based or stochastic optimizers are employed for minimizing $\|\mathbf{x} - \mathbf{y}\|$. After successful training, the network "compresses" the input data when smaller numbers of neurons are used in the central layer than in the input layer, and can be used for generating projections of **x** as outputs of the central layer neurons **x'** (Figure S2B). Eq. S5 gives the simplified network function computing the actual projection, where $\boldsymbol{\vartheta}$ and $\boldsymbol{\theta}$ are hidden and central neuron bias values.

$$x_i' = f(\mathbf{x}) = \sum_j^{Hidden}(w_j^{Central}(\sum_k^{Input} w_k^{Hidden}) + \vartheta_k) + \theta_i \,. \qquad \text{(S5)}$$

It must be kept in mind that there is an infinite number of input vectors that are projected to the same point in **x'**, and therefore, particular care must be taken to clearly define the applicability domain of encoder network-based QSAR models[S51,S52]. Only recently, this topic has been re-visited and appropriate techniques for applicability domain estimation for various machine learning models have been proposed[S53-S55].

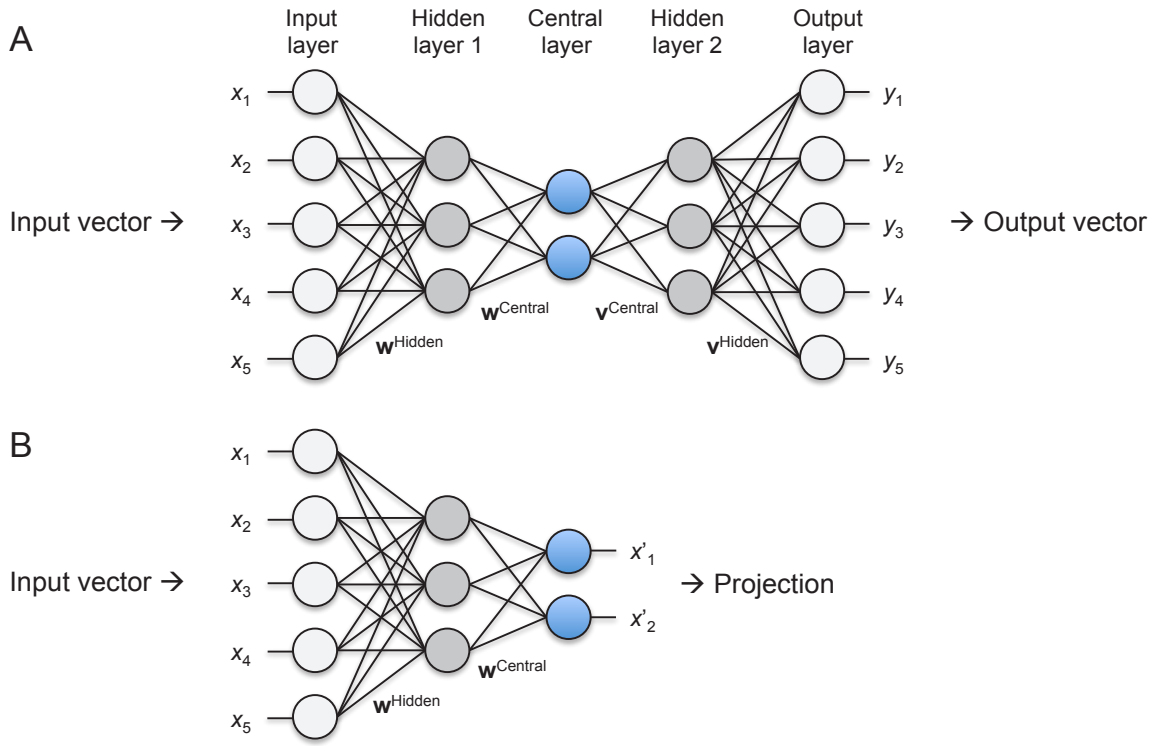


Figure S2. Topology of an encoder network for projection of multivariate data. Artificial neurons are drawn as circles, connection weights as lines between neurons in the different network layers. The symmetric network is trained so that any output vector **y** ideally is identical to its corresponding input vector **x** (A). After successful network training the central layer neurons (blue circles) compute the desired projection of **x** (B). The network structure shown accepts a five-dimensional input vector and computes a two-dimensional projection **x'**.

Self-Organizing Map

The concept of self-organizing mapping of high-dimensional input was conceived by Kohonen in the early 1980s[S56], and introduced to chemistry and drug design by Gasteiger and coworkers in the 1990s[S57]. The *Self-Organizing Map* (SOM) (or "Kohonen net") has been extensively applied in drug discovery ever since[S58,S59]. The SOM architecture consists of a regular array of so-called "neurons", which essentially are vectors that are arranged in a topological structure (typically a 2D array) and have the same dimension as the input data. During the SOM training process – an optimization procedure following the principles of unsupervised, associative Hebbian

learning[S60] - the original high-dimensional space is tessellated, resulting in as many data clusters as there are neurons in the SOM. The neurons represent centroids of each cluster (Voronoi field). Data points within a cluster are more similar to "their" neuron than to any other neuron of the SOM. In this regard, SOM training may be considered a variant of *k*-means clustering, similar to vector quantization[S61,S62]. The resulting prototype vectors capture features in the input space that are unique for each data cluster. Molecular feature analysis can be done, *e.g.*, by comparing adjacent neuron vectors. In analogy to the Mercator projection shown in Figure S1, Figure S3A presents a SOM projection from 3D coordinates of points on the earth's surface to a 2D map. The SOM grid contains 2400 neurons arranged as a toroidal 60×40 grid. Some major city locations are highlighted as reference points. Note that although the overall distribution and shapes of continents and oceans is not metric, local neighborhoods are preserved, *e.g*. London and Zurich are close to each other on the 2D map.

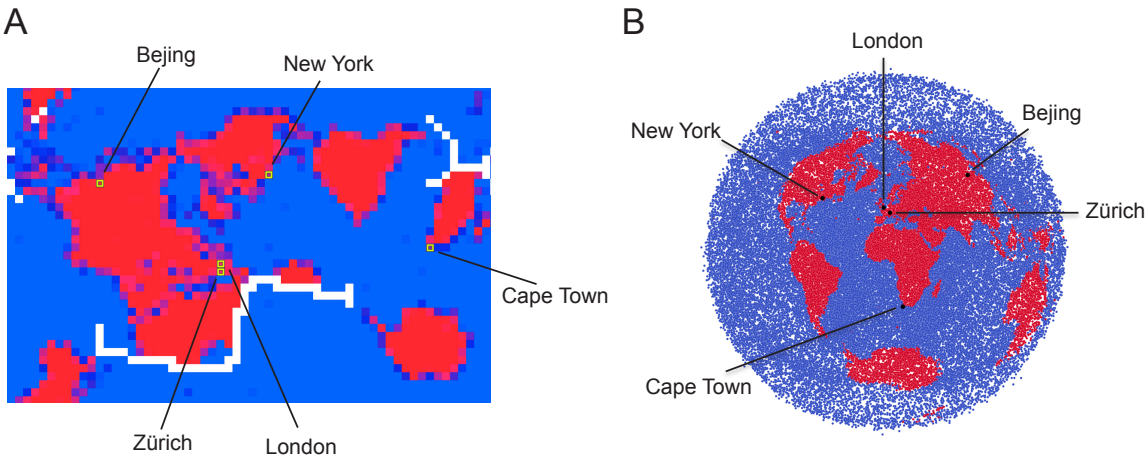


Figure S3. Projections of the land-water distribution on the surface of the earth (*red*: land, *blue*: water). The left panel (A) presents a 2D SOM (60×40 neurons) projection of these data. In B) Isometric SPE (ISPE) (r_c = 0.2) was used to generate a 2D map. White color in A) indicates "empty space", *i.e.* neurons without data points assigned. Earth data were obtained from the NASA NEO data set "Blue Marble: Next Generation (Terra/MODIS)". The 2D latitude/longitude data, as defined by the position in the image, were transformed to Cartesian coordinates. Random sampling from the sphere surface was applied to select subsets of 15,007 (SOM) and 65,167 points (ISPE), equally distributed over the sphere. The binary land/water descriptor was calculated from the color information in HSV color space.

Kohonen's algorithm represents an efficient way for mapping vectors that are close to each other in input space onto contiguous locations in the output space. Preservation of *local* neighborhood is achieved by introducing a topology to the layout of the SOM neurons. The simplest topology is a chain of neurons, followed by a 2D grid. Molecules that are located in adjacent clusters on the neuron grid are also close to

each other in the original high-dimensional space. Topological mapping can be achieved by two simple rules that guide the training process:

1. For input vector **x** locate the best-matching neuron ("winner" neuron, **w***).
2. Move this neuron and its topological neighbors toward **x**.

For the first rule vector distances between **x** and the SOM neurons **w** have to be computed (Eq. S6). The number of comparisons needed depends linearly on the size of the self-organizing system *S*, which can be expressed by its number of neurons.

$$\| \mathbf{x} - \mathbf{w_i} \| \rightarrow min. (\forall\, i \in S) \tag{S6}$$

The second rule requires an updating procedure to adapt the vector elements of the winner neuron **w*** and its topological neighbors (Eq. S7), where ε is a learning rate depending on both the topological distance between **w*** and neuron $\mathbf{w}_i$, and on the training time passed. A toroidal neuron topology can be used to avoid some boundary problems inherent to a planar topology. For a full description of the SOM algorithm see the literature[S63].

$$\mathbf{w}_i = \mathbf{w}_i + \varepsilon \, \| \mathbf{x} - \mathbf{w_i} \| . \tag{S7}$$

Figure 4 presents an application of SOM-based virtual screening for new kinase inhibitors[S64]. The idea in this study was to map and cluster known drugs and lead compounds and a virtual combinatorial library on a 2D SOM. All compounds were represented by 150-dimensional CATS descriptors, a topological pharmacophore feature representation[S65,S66]. The SOM was then colored according to the prevalence of combinatorial compounds (Figure S4A) and known kinase inhibitors (COBRA data collection, Figure S4B). The neuron containing the most combinatorial compounds coincides with a kinase "activity island"[S67] This particular cluster holds many of the reference inhibitors (seven-fold overrepresentation of kinase inhibitors compare to the background distribution), and it was therefore reasonable to assume similar targets for the combinatorial compounds. One candidate (compound **1**) from the virtual library was actually synthesized and successfully tested in a CDK2 inhibition assay.

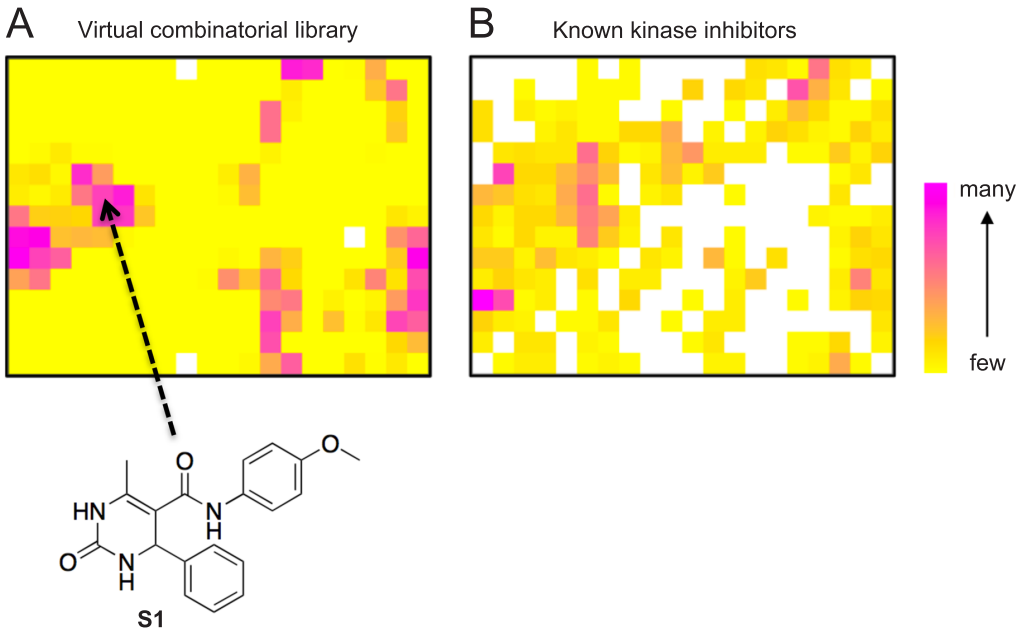


Figure S4. 2D SOM projection of a virtual combinatorial library containing Biginelli-type dihydropyrimidones (A) and drug-like bioactive compounds with kinase inhibitors highlighted (B). The SOM grid contains (15×20) neurons. By SOM analysis, compound **S1** was identified as an inhibitor of cyclin-dependent kinase 2 (CDK2).

The SOM virtual screening approach presented in Figure S4 belongs to the class of ligand-based similarity searching methods[S24,S68,S69]. In contrast to using reference compounds as queries and ranking the combinatorial screening compounds by some pharmacophore similarity index, the SOM offers the potential advantage of performing similarity searching using a "common pharmacophore" model (*i.e.* the neuron vector) as query. This avoids the necessity for comparing and merging ranked lists of candidate compounds[S70]. Despite its appeal, the SOM approach used in this study has several disadvantages compared to other ligand-based virtual screening techniques. A major limitation of the original SOM algorithm is that the dimension of the output space and the number of neurons must be predefined prior to SOM training[S71]. A disadvantage of SOMs can be the comparatively long training time needed, especially if large data sets (*e.g.* HTS screening data, combinatorial compound libraries) are used. Different training runs bear the additional danger of delivering slightly different results due to the stochastic nature of SOM optimization. Several variations and extensions of Kohonen's original SOM algorithm have been published and applied to drug discovery[S72]. Such developments include self-organizing networks with an adapting grid size[S73], cascaded SOMs[S74], and hybrid neural networks[S75,S76]. These systems might provide alternative approaches to virtual

compound screening, although their practical usefulness and applicability to hit and lead finding still needs to be rigorously assessed.

Stochastic Proximity Embedding

Stochastic Proximity Embedding (SPE) is a self-organizing dimensionality reduction algorithm that aims at preserving the pairwise proximities in the lower dimensional embedding. It was introduced to drug discovery by Agrafiotis and coworkers in 2002[S77]. In its first version, SPE was used to find a stochastic approximation of multidimensional scaling that preserved the metric structure. The mapping procedure uses a pairwise refinement strategy that does not require the complete distance or proximity matrix and scales linear with the size of the data set. This allows dimension reduction even for large data sets.

If the data forms a lower dimensional nonlinear manifold, conventional similarity measures, such as the Euclidean distance, tend to underestimate the proximity of points and lead to erroneous embedding. To properly reconstruct the manifold, the geodesic distance, the proximity of two points measured on the manifold itself, needs to be preserved. Algorithms like IsoMap[S27] or LLE[S26] estimate the geodesic distance from the local neighborhood. Agrafiotis observed that the geodesic distance is always greater or equal to the input proximity. If two points are close, the input proximity provides a good approximation to their geodesic distance; when they are further away, the input proximity provides a lower bound[S78]. *Isometric SPE* (ISPE) circumvents the calculation of estimated geodesic distances by incorporating this observation[S6]. It forces embedding distances of nearby points to match their input proximities, while points whose input proximities are larger than a defined threshold are forced to stay further apart. ISPE preserves the distances of the local neighborhood and views the distances between remote points as lower bounds of their true geodesic distances and uses them to impose the global structure (Figure 2B). A similar approach has been applied to SOM training, where neurons adjacent to the winner neuron $\mathbf{w}^*$ are attracted to the data point presented, while neurons outside a defined topological neighborhood on the SOM grid are pushed away. Such a procedure can also be used to enhance contrast on SOMs[S79]. The stress function S minimized stochastically by ISPE is given by Eq. S8.

$$S = \sum_{i<j} \frac{f(d_{ij}, r_{ij})}{r_{ij}} \Big/ \sum_{i<j} r_{ij} \rightarrow min., \tag{S8}$$

where r_{ij} is the input proximity between the i^{th} and j^{th} point, d_{ij} is their Euclidean distance in the low-dimensional embedding space and $f(d_{ij}, r_{ij})$ is the pairwise stress function defined as $f(d_{ij}, r_{ij}) = (d_{ij} - r_{ij})^2$ if $r_{ij} < r_c$ or $d_{ij} < r_{ij}$ and $f(d_{ij}, r_{ij}) = 0$ if $r_{ij} > r_c$ and $d_{ij} > r_{ij}$. r_c is the neighborhood radius.

ISPE minimizes the stress function with a stochastic steepest descent approach. It iteratively refines a starting configuration of the data points by repeatedly selecting two points at random and adjusting their coordinates so that their embedding distance d_{ij} matches more closely their input proximity r_{ij}. The correction is proportional to the disparity $\lambda\, |r\_ij - d\_ij| \,/\, d\_ij$, where λ is a learning rate. To avoid oscillation the learning rate is decreased during the course of refinement. If the points are not neighbors, if $r_{ij} > r_c$ and $d_{ij} > r_{ij}$, their coordinates remain unmodified. The result of ISPE strongly depends on the choice of the neighborhood radius for learning the embedded manifold. If r_c is too large, shortcuts to other branches of the manifold are possible, whereas if it is too small it may lead to fragmented clusters (Figure S5).

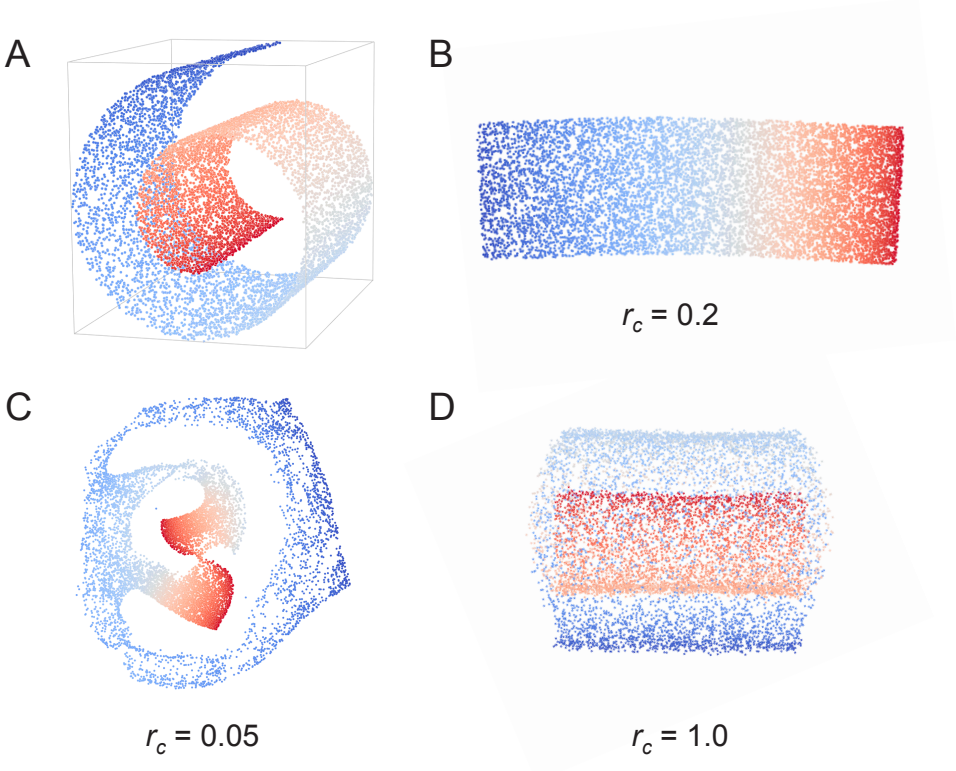


Figure S5. ISPE projections of the 3D "Swiss roll" manifold (A) to two dimensions (B-D) using different cut-off distances r_c. The data points for the Swiss roll were obtained by generating coordinate triplets ($x = i \cos(i)$, $y = i \sin(i)$, j), where i and j are random numbers from the intervals [5, 15] and [0, 30], respectively. The color corresponds to the angle i.

As an illustrative example, Figure S6 presents the application of PCA and ISPE to producing a 2D chemical structure depiction from 3D molecular atom coordinates. The first compound is PPAR-gamma agonist pioglitazone, for which a receptor-bound conformation served as "high-dimensional" input. Apparently, both PCA and ISPE are able to produce a 2D visualization lacking great distortion. This can be explained by the extended shape of the pioglitzone conformer, which is properly projected on the first two PCs. The second example provides 2D projections computed for epothilone D bound to cytochrome P450epoK. Here, ISPE generates a more appealing, less compact 2D mapping than PCA, probably due to the greater degree of 3D conformational folding of the reference structure.

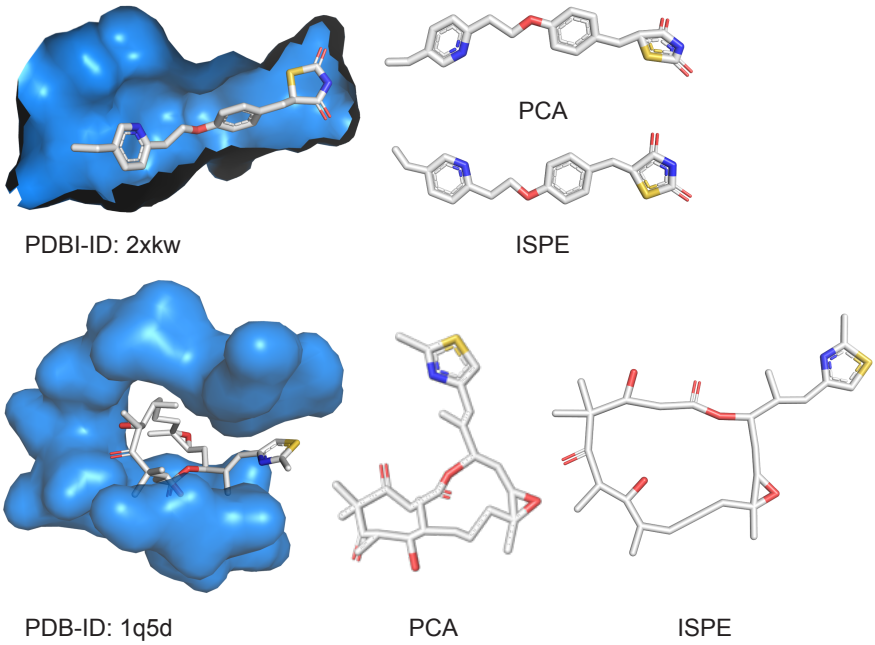


Figure S6. Generation of 2D projections (right) for 3D molecular conformations (left). The top panel shows the results obtained by PCA and ISPE (r_c = 0.2) for the example of PPAR-gamma agonist pioglitazone bound to the receptor (PDB-ID[S80] 2xkw[S81]). The bottom panel provides the projections computed for epothilone D bound to cytochrome P450epoK (PDB-ID 1q5d[S82]).

Stochastic Neighbor Embedding

In contrast to SPE, *Stochastic Neighbor Embedding* (SNE) does not try to preserve pairwise *distances* but instead the *probabilities* of points being neighbors[S83]. The pairwise distances in the input and output space are used to calculate the probability distributions that point *i* is a neighbor of point *j*. The aim of the embedding is to approximate the neighbor probability distribution as close as possible in the low-dimensional embedding.

The probability of point *i* being neighbor to point *j* in the input space is defined as (Eq. S9):

$$p_{ij} = \frac{\exp\left(-d_{ij}^2\right)}{\sum_{k \neq i} \exp\left(-d_{ik}^2\right)} \ . \tag{S9}$$

The proximities d_{ij}^2 may be given as a proximity matrix or calculated using the scaled squared Euclidean distance between the high-dimensional input data points $\mathbf{x}_i$ and $\mathbf{x}_j$ (Eq. S10).

$$d_{ij}^2 = \frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma_i^2}. \tag{S10}$$

The scaling factor σ_i, the variance of the Gaussian at point $\mathbf{x}_i$, is either set manually or by fixing the entropy of the distribution. Setting the entropy to log *k* sets the "effective number of local neighbors" to *k*[S83].

The induced probability q_{ij} that point *i* picks point *j* as its neighbor in the embedding space, with $\mathbf{y}_i$ being the low-dimensional images, is defined as (Eq. S11):

$$q_{ij} = \frac{\exp\left(-\left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2\right)}{\sum_{k \neq i} \exp\left(-\left\|\mathbf{y}_i - \mathbf{y}_k\right\|^2\right)} . \tag{S11}$$

The aim is to match the probabilities as close as possible, as measured by the sum of *Kullback-Leibler* (KL) divergences between the original and induced distributions over neighbors for each object (Eq. S12).

$$C = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} = \sum_i KL(P_i || Q_i) \ . \tag{S12}$$

Hinton *et al.* already showed that the probabilistic framework can easily be extended to allow multiple low-dimensional images for each high-dimensional object through a mixture of Gaussians[S83]. In the past few years several extensions to the classical SNE algorithm have been published. *t-distributed Stochastic Neighbor Embedding* (t-SNE)[S84] uses a symmetric cost function, which is easier to optimize, and a Student-t distribution instead of the Gaussian to model similarity in low-dimensional space. It improves visualization because natural clusters tend to be more separated in the low-dimensional embedding, thus it simplifies the visual perception of clusters. The application was extended to visualize data together with class labels[S85], incorporate multiple similarity matrices[S86], or multiple views of the input data[S87]. SNE has also

been analyzed within the framework of information retrieval and a variant has been introduced, which optimizes the retrieval quality, quantified by precision and recall[S88]. A limitation of SNE is its computational demand for projecting large datasets due to the calculation of the complete pairwise probability matrix and steepest-decent optimization. To some extent this has been addressed in recent publications using trust-regions to speed-up convergence[S89], or landmark sampling to reduce memory consumption[S83].

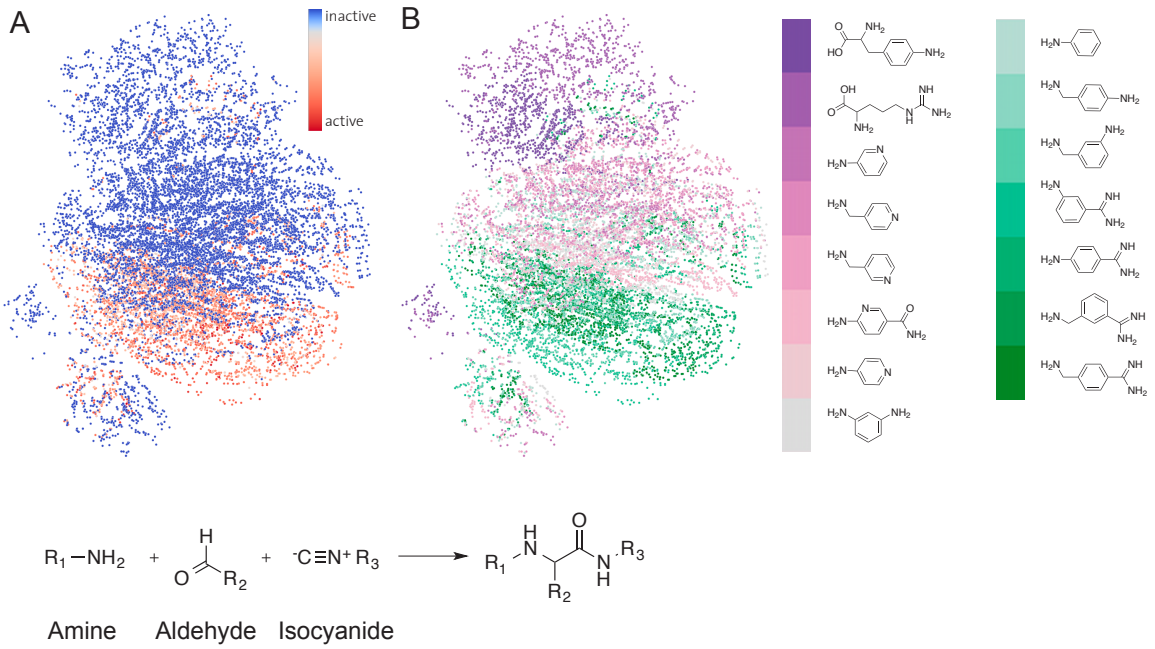


Figure S7. 2D SNE projection of a combinatorial library containing 15,840 three-component Ugi reaction products (α-aminoacyl amide derivatives) formed from the condensation of an amine, aldehyde and isocyanide. In (A) the compounds' measured inhibitory activity against Tryptase is shown by color-coding each compound (dot) from blue (inactive) to red (active). In (B) compound clusters containing the same amine building block (R_1) are highlighted in a different color. Ugi data courtesy of Dr. Lutz Weber (Morphochem AG). Compound structures were standardized using the "wash" method in MOE v2010.10 and explicit hydrogen atoms were removed prior to descriptor calculation. Topological CATS descriptors were computed using the *speedcats* software (0-9 bonds, type-sensitive scaling)[S91].

A typical application of SNE is the visualization of the distribution of combinatorial compound libraries in some high-dimensional pattern space spanned by pharmacophore descriptors (Figure S7). As an example, we encoded a library of 15,840 three-component Ugi reaction products by the topological CATS descriptor. Figure S7A presents a 2D projection of these 250-dimensional data that was obtained by ISPE. Color corresponds to measure inhibitory activity of the compound against tryptase. An "activity island" containing many inhibitors is highlighted in red color (low IC_{50} values). Visualization compound distributions can also help to graphically investigate preliminary SARs. In Figure S7B, the same ISPE projection is colored

according to the primary amine identity used in the multi-component Ugi reaction. It becomes evident from looking at the color distribution that certain positively charged benzamidine derivatives seem to be preferred for blocking the serine protease tryptase, which has been long known from numerous medicinal chemistry projects[S90].

8.1.4 Conclusion

Maps of chemical space have demonstrated their usefulness for analyzing the diversity and complementarity of compound libraries, with particular emphasis on combinatorial compound collections and QSAR modeling[S92-S94]. Efforts to develop open platform for QSAR model generation and data analysis are ongoing[S95,S96], with visualization techniques playing an important role in activity prediction, extraction of ligand-target networks, structural diversity analysis, and cluster visualization. Structure-activity landscapes have received much attention recently[S97,S98], mainly driven by innovative visualization methods that allow for online monitoring of dynamic landscapes and fast and efficient embedding of chemical structures and picturing response surfaces[S99,S100]. Such methods, including our own visualization tool LiSARD (Ligand Structure Activity Relationship Display)[S101], might become a valuable addition to the drug designer's toolbox. Latest developments include a study by Soto *et al.* who compared several mapping algorithms and suggest *Correlative Matrix Mapping* (CMM) as a potential method of choice for target-driven subspace mapping[S102]. We are also witnessing continuing amalgamation of methods. For example, *Neighbor Embedding XOM* (NE-XOM), an extension to the *Exploratory Observation Machine*[S103], is based on minimizing the Kullback-Leibler divergence of neighborhood functions in data and embedding space[S104]. In this it is comparable to SNE combined with principles first encountered in SOM modeling. Numerous related concepts are being developed mainly in the context of machine-learning applications. We expect such innovative data mapping approaches to be studied for their transferability and practical usefulness in molecular modeling and drug discovery, thereby complementing automated virtual screening protocols for rapid focused library design and compound prioritization[S105,S106].

8.1.5 References

S1. Maniyar DM, Nabney IT, Williams BS, Sewing A (2006) Data visualization during the early stages of drug discovery. *J. Chem. Inf. Model.* **46**, 1806-1818.

S2. Howe TJ, Mahieu G, Marichal P, Tabruyn T, Vugts P (2007) Data reduction and representation in drug discovery. *Drug Discov. Today* **12**, 45-53.

S3. Bienfait B, Gasteiger J (1997) Checking the projection display of multivariate data with colored graphs. *J. Mol. Graph. Model.* **15**, 203-215.

S4. Ivanenkov YA, Savchuk NP, Ekins S, Balakin KV (2009) Computational mapping tools for drug discovery. *Drug Discov. Today* **14**, 767-775.

S5. Medina-Franco JL, Martinez-Mayorga K, Giulianotti MA, Houghten RA, Pinilla C (2008) Visualization of the chemical space in drug discovery. *Curr. Comput.-Aided Drug Des.* **4**, 322-333.

S6. Agrafiotis DK, Xu H (2003) A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.* **43**, 475-484.

S7. Rupp M, Schneider P, Schneider G (2009) Distance phenomena in high-dimensional chemical descriptor spaces: Consequences for similarity-based approaches. *J. Comput. Chem.* **30**, 2285-2296.

S8. Jolliffe I (1986) *Principal component analysis.* Springer, New York.

S9. Linusson A, Elofsson M, Andersson IE, Dahlgren MK (2010) Statistical molecular design of balanced compound libraries for QSAR modeling. *Curr. Med. Chem.* **17**, 2001-2016.

S10. Oprea TI, Gottfries J (2001) Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **3**, 157-166.

S11. 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. *Nucl. Acids. Res.* **40**, D1100-D1107.

S12. Li Q, Cheng T, Wang Y, Bryant SH (2010) PubChem as a public resource for drug discovery. *Drug Discov. Today* **15**, 1052-1057.

S13. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M, Ferraiolo P, Tolliday NJ, Schreiber SL, Clemons PA (2008) ChemBank: A small-molecule screening and cheminformatics resource database. *Nucl. Acids. Res.* **36**, D351-D359.

S14. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C (2010) Chemical entities of biological interest: An update. *Nucl. Acids. Res.* **38**, D249-D254.

S15. Chen J, Swamidass SJ, Dou Y, Bruand J, Baldi P (2005) ChemDB: A public database of small molecules and related chemoinformatics resources. *Bioinformatics* **21**, 4133-4139.

S16. Gozalbes R, Pineda-Lucena A (2011) Small molecule databases and chemical descriptors useful in chemoinformatics: An overview. *Comb. Chem. High Throughput Screening* **14**, 548-558.

S17. Bellis LJ, Akhtar R, Al-Lazikani B, Atkinson F, Bento AP, Chambers J, Davies M, Gaulton A, Hersey A, Ikeda K, Kruger FA, Light Y, McGlinchey S, Santos R, Stauch B, Overington JP (2011) Collation and data-mining of literature bioactivity data for drug discovery. *Biochem. Soc. Trans.* **39**, 1365-1370.

S18. Martin YC, Kofron JL, Traphagen LM (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **45**, 4350-4358.

S19. Johnson MA, Maggiora GM (1990) *Concepts and applications of molecular similarity.* Wiley, New York.

S20. Barbosa F, Horvath D (2004) Molecular similarity and property similarity. *Curr. Trends Med. Chem.* **4**, 589-600.

S21. Oprea TI (2002) Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.* **6**, 384-389.

S22. Bellman RE (1961) *Adaptive control processes.* Princeton University Press, Princeton, NJ.
S23. Güner OF (2002) History and evolution of the pharmacophore concept in computer-aided drug design. *Curr. Top. Med. Chem.* **2**, 1321-1332.
S24. Willett P (2011) Similarity searching using 2D structural fingerprints. *Methods Mol. Biol.* **672**, 133-158.
S25. Schneider G, So S-S (2003) *Adaptive systems in drug design.* Landes Bioscience, Georgetown.
S26. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323-2326.
S27. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319-2323.
S28. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**, 1373-1396.
S29. Choi H, Choi S (2004) Kernel isomap. *Electron. Lett.* **40**, 1612.
S30. Hibbs MA, Dirksen NC, Li K, Troyanskaya OG (2005) Visualization methods for statistical analysis of microarray clusters. *BMC Bioinf.* **6**, 115.
S31. Law MH, Jain AK (2006) Incremental nonlinear dimensionality reduction by manifold learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 377-391.
S32. Lee G, Rodriguez C, Madabhushi A (2008) Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **5**, 368-384.
S33. Higgs BW, Weller J, Solka JL (2006) Spectral embedding finds meaningful (relevant) structure in image and microarray data. *BMC Bioinf.* **7**, 74.
S34. Ham J, Lee DD, Mika S, Schölkopf B (2004) A kernel view of the dimensionality reduction of manifolds. *Proc. ICML* **47**,
S35. Schölkopf B, Smola A, Müller K-R (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **10**, 1299-1319.
S36. Sakiyama Y (2009) The use of machine learning and nonlinear statistical tools for ADME prediction. *Expert Opin. Drug Metab. Toxicol.* **5**, 149-169.
S37. Obrezanova O, Csanyi G, Gola JM, Segall MD (2007) Gaussian processes: A method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model.* **47**, 1847-1857.
S38. Rupp M, Schroeter T, Steri R, Zettl H, Proschak E, Hansen K, Rau O, Schwarz O, Müller-Kuhrt L, Schubert-Zsilavecz M, Müller K-R, Schneider G (2010) From machine learning to natural product derivatives that selectively activate transcription factor PPARγ. *ChemMedChem* **5**, 191-194.
S39. Franke L, Byvatov E, Werz O, Steinhilber D, Schneider P, Schneider G (2005) Extraction and visualization of potential pharmacophore points using support vector machines: Application to ligand-based virtual screening for COX-2 inhibitors. *J. Med. Chem.* **48**, 6997-7004.
S40. Hansen K, Baehrens D, Schroeter T, Rupp M, Müller K-R (2011) Visual interpretation of kernel-based prediction models. *Mol. Inf.* **30**, 817-826.
S41. Rosenbaum L, Hinselmann G, Jahn A, Zell A (2011) Interpreting linear support vector machine models with heat map molecule coloring. *J. Cheminform.* **3**, 11.
S42. Schneider G, Wrede P (1998) Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* **70**, 175-222.
S43. Zupan J, Gasteiger J (1999) *Neural networks in chemistry and drug design, 2nd edition.* Wiley-VCH, Weinheim.
S44. Livingstone DJ, Manallack DT, Tetko IV (1997) Data modelling with neural networks: Advantages and limitations. *J. Comput. Aided Mol. Des.* **11**, 135-142.
S45. Livingstone DJ (1996) *Multivariate data display using neural networks.* In: *Neural networks in QSAR and drug design* (ed. Devillers J) Academic Press, London, 157-176.
S46. Livingstone DJ, Hesketh G, Clayworth D (1991) Novel method for the display of multivariate data using neural networks. *J. Mol. Graph.* **9**, 115-118.

S47. Reibnegger G, Werner-Felmayer G, Wachter H (1993) A note on the low-dimensional display of multivariate data using neural networks. *J. Mol. Graph.* **11**, 129-133.
S48. Brown N, Lewis RA (2006) Exploiting QSAR methods in lead optimization. *Curr. Opin. Drug Discovery Dev.* **9**, 419.
S49. Visco DP, Pophale RS, Rintoul MD, Faulon J-L (2002) Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor. *J. Mol. Graph. Model.* **20**, 429-438.
S50. Givehchi A, Dietrich A, Wrede P, Schneider G (2003) ChemSpaceShuttle: A tool for data mining in drug discovery by classification, projection, and 3D visualization. *QSAR Comb. Sci.* **22**, 549-559.
S51. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set descriptor space: A review. *ATLA.* **33**, 445.
S52. Tropsha A, Golbraikh A (2007) Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **13**, 3494-3504.
S53. Melville JL, Burke EK, Hirst JD (2009) Machine learning in virtual screening. *Comb. Chem. High Throughput Screening* **12**, 332-343.
S54. Schwaighofer A, Schroeter T, Mika S, Blanchard G (2009) How wrong can we get? A review of machine learning approaches and error bars. *Comb. Chem. High Throughput Screening* **12**, 453 - 468.
S55. Schroeter TS, Schwaighofer A, Mika S, Ter Laak A, Suelzle D, Ganzer U, Heinrich N, Müller K-R (2007) Estimating the domain of applicability for machine learning QSAR models: A study on aqueous solubility of drug discovery molecules. *J. Comput. Aided Mol. Des.* **21**, 485-498.
S56. Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59-69.
S57. Gasteiger J, Zupan J (1993) Neural networks in chemistry. *Angew. Chem. Int. Ed.* **32**, 503-527.
S58. Kirew DB, Chretien JR, Bernard P, Ros F (1998) Application of Kohonen Neural Networks in classification of biologically active compounds. *SAR QSAR Environ. Res.* **8**, 93-107.
S59. Schneider P, Tanrikulu Y, Schneider G (2009) Self-organizing maps in drug discovery: Compound library design, scaffold-hopping, repurposing. *Curr. Med. Chem.* **16**, 258-266.
S60. Hebb DO (1949) *The organization of behavior.* Wiley & Sons, New York.
S61. Hertz JA, Krogh AS, Palmer RG (1991) *Introduction to the theory of neural computacion.* Addison-Wesley, Redwood City CA.
S62. Duda RO, Hart PE, Stork DG (2001) *Pattern classification.* Wiley & Sons, New York.
S63. Kohonen T (1984) *Self-organization and associative memory.* Springer-Verlag, Heidelberg.
S64. Schneider P, Stutz K, Kasper L, Haller S, Reutlinger M, Reisen F, Geppert T, Schneider G (2011) Target profile prediction and practical evaluation of a Biginelli-type dihydropyrimidine compound library. *Pharmaceuticals* **4**, 1236-1247.
S65. Schneider G, Neidhart W, Giller T, Schmid G (1999) "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem. Int. Ed.* **38**, 2894-2896.
S66. Schüller A, Schneider G (2008) Identification of hits and lead structure candidates with limited resources by adaptive optimization. *J. Chem. Inf. Model.* **48**, 1473-1491.
S67. Schneider G, Schneider P (2004) *Navigation in chemical space: Ligand-based design of focused compound libraries.* In: *Chemogenomics in Drug Discovery* (eds. Kubinyi H, Müller G) Wiley-VCH, Weinheim, 341-376.
S68. Yan A (2006) Application of self-organizing maps in compounds pattern recognition and combinatorial library design. *Comb. Chem. High Throughput Screening* **9**, 473-480.
S69. Digles D, Ecker GF (2011) Self-organizing maps for *in silico* screening and data visualization. *Mol. Inf.* **30**, 838-846.

S70. Holliday JD, Kanoulas E, Malim N, Willett P (2011) Multiple search methods for similarity-based virtual screening: Analysis of search overlap and precision. *J. Cheminform.* **3**, 29.
S71. Ultsch A (2003) Maps for the visualization of high-dimensional data spaces. *Proc. WSOM* 225-230.
S72. Selzer P, Ertl P (2006) Applications of self-organizing neural networks in virtual screening and diversity selection. *J. Chem. Inf. Model.* **46**, 2319-2323.
S73. Wu Z, Yen GG (2003) A SOM projection technique with the growing structure for visualizing high-dimensional data. *Int. J. Neural Syst.* **13**, 353-365.
S74. Furukawa T (2009) SOM of SOMs. *Neural Netw.* **22**, 463-478.
S75. Tetko IV (2002) Associative neural network. *Neural Process. Lett.* **16**, 187-199.
S76. Gupta S, Matthew S, Abreu PM, Aires-de-Sousa J (2006) QSAR analysis of phenolic antioxidants using MOLMAP descriptors of local properties. *Bioorg. Med. Chem.* **14**, 1199-1206.
S77. Agrafiotis DK (2003) Stochastic proximity embedding. *J. Comput. Chem.* **24**, 1215-1221.
S78. Agrafiotis DK, Xu H, Zhu F, Bandyopadhyay D, Liu P (2010) Stochastic proximity embedding: Methods and applications. *Mol. Inf.* **29**, 758-770.
S79. Schmuker M, Schneider G (2007) Processing and classification of chemical data inspired by insect olfaction. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20285-20289.
S80. Berman HM (2000) The protein data bank. *Nucl. Acids. Res.* **28**, 235-242.
S81. Mueller JJ, Schupp M, Unger T, Kintscher U, Heinemann U (article to be published) Binding diversity of pioglitazone by peroxisome proliferator-activated receptor-gamma. *downloaded from URL: http://www.pdb.org*
S82. Nagano S, Li H, Shimizu H, Nishida C, Ogura H, Ortiz de Montellano PR, Poulos TL (2003) Crystal structures of epothilone D-bound, epothilone B-bound, and substrate-free forms of cytochrome P450epoK. *J. Biol. Chem.* **278**, 44886-44893.
S83. Hinton GE, Roweis ST (2002) Stochastic neighbor embedding. *Proc. NIPS* **15**, 833-840.
S84. Van der Maaten LJP, Hinton GE (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579-2605.
S85. Iwata T, Saito K, Ueda N, Stromsten S, Griffiths TL, Tenenbaum JB (2007) Parametric embedding for class visualization. *Neural Comput.* **19**, 2536-2556.
S86. Memisevic R, Hinton GE (2005) Multiple relational embedding. *Proc. NIPS* **17**, 913-920.
S87. Xie B, Mu Y, Tao D, Huang K (2011) m-SNE: Multiview stochastic neighbor embedding. *IEEE Trans. Syst. Man. Cybern. B Cybern.* **41**, 1088-1096.
S88. Venna J, Peltonen J, Nybo K, Aidos H, Kaski S (2010) Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *J. Mach. Learn. Res.* **11**, 451-490.
S89. Kijoeng N, Hongmo J, Seungjin C (2004) Fast stochastic neighbor embedding: A trust-region algorithm. *Proc. IJCNN* 123-128.
S90. Stürzebecher J, Vieweg H, Wikström P, Turk D, Bode W (1992) Interactions of thrombin with benzamidine-based inhibitors. *Biol. Chem. Hoppe-Seyler* **373**, 491-496.
S91. Fechner U, Schneider G (2004) Optimization of a pharmacophore-based correlation vector descriptor for similarity searching. *QSAR Comb. Sci.* **23**, 19-22.
S92. Akella LB, DeCaprio D (2010) Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **14**, 325-330.
S93. Walker T, Grulke CM, Pozefsky D, Tropsha A (2010) Chembench: A cheminformatics workbench. *Bioinformatics* **26**, 3000-3001.
S94. Singh N, Guha R, Giulianotti MA, Pinilla C, Houghten RA, Medina-Franco JL (2009) Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J. Chem. Inf. Model.* **49**, 1010-1024.

S95. Sushko I, Novotarskyi S, Korner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV (2011) Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* **25**, 533-554.
S96. Backman TW, Cao Y, Girke T (2011) ChemMine tools: An online service for analyzing and clustering small molecules. *Nucl. Acids. Res.* **39**, W486-W491.
S97. Guha R, Van Drie JH (2008) Structure-activity landscape index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **48**, 646-658.
S98. Guha R (2011) The ups and downs of structure-activity landscapes. *Methods Mol. Biol.* **672**, 101-117.
S99. Iyer P, Hu Y, Bajorath J (2011) SAR monitoring of evolving compound data sets using activity landscapes. *J. Chem. Inf. Model.* **51**, 532-540.
S100. Peltason L, Iyer P, Bajorath J (2010) Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J. Chem. Inf. Model.* **50**, 1021-1033.
S101. Reutlinger M, Guba W, Martin RE, Alanine AI, Hoffmann T, Klenner A, Hiss JA, Schneider P, Schneider G (2011) Neighborhood-preserving visualization of adaptive structure-activity landscapes: Application to drug discovery. *Angew. Chem. Int. Ed.* **50**, 11633-11636.
S102. Soto AJ, Vazquez GE, Strickert M, Ponzoni I (2011) Target-driven subspace mapping methods and their applicability domain estimation. *Mol. Inf.* **30**, 779-789.
S103. Wismüller A (2009) The exploration machine: A novel method for analyzing high-dimensional data in computer-aided diagnosis. *Proc. SPIE* **7260**, 72600G.
S104. Bunte K, Hammer B, Villmann T, Biehl M, Wismüller A (2011) Neighbor embedding XOM for dimension reduction and visualization. *Neurocomputing* **74**, 1340-1350.
S105. Irwin JJ (2008) Using ZINC to acquire a virtual screening library. *Curr. Protoc. Bioinformatics* **22**, 14.6.1-14.6.23.
S106. Campbell SJ, Gaulton A, Marshall J, Bichko D, Martin S, Brouwer C, Harland L (2010) Visualizing the drug target landscape. *Drug Discov. Today* **15**, 3-15.

8.1.6 Publication details and contributions

Authors

Michael Reutlinger[a] & Gisbert Schneider[a]

[a] ETH Zurich, Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, Wolfgang-Pauli-Str. 10, 8093 Zurich, Switzerland.

Author contributions

M.R. designed and conducted the research, and contributed to the manuscript.

Acknowledgments

The authors are grateful to Dr. Jan Hiss for stimulating discussion, and Dr. Petra Schneider and Dr. Lutz Weber for providing data sets. The Chemical Computing Group Inc. (Montreal, Canada) provided a research license of MOE software.

Reference

Reutlinger M, Schneider G (2012) Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J. Mol. Graph. Model.* **34**, 108-117.

Licence

Source of funding

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, FOR1406TP4)

8.2 Supporting Information – Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for "orphan" molecules

8.2.1 General information

All starting materials and solvents were obtained from ABCR Chemicals, Aldrich, Fluka, Alfa Aesar or Acros, and were used without further purification.

Combinatorial library was built using J-KEM® Scientific robot for mixing the building blocks. Microwave-assisted synthesis was then carried out in a Biotage Initiator reactor.

Analytical HPLC-MS was carried out in a Shimadzu LC-MS2020 system, equipped with a Nucleodur C_{18} HTec column, under an appropriate gradient of acetonitrile : H_2O (+ 0.1% trifluoroacetic acid in each phase), and a total flow rate of 0.5 mL/min. The mass spectrometer was operated in positive-ion mode with ESI. Preparative HPLC was carried out on a Shimadzu LC-8A system, coupled to a Nucleodur 100-5 C_{18} HTec column, and a SPD-20A UV/Vis detector.

Proton and carbon nuclear magnetic resonance spectra (^{1}H and ^{13}C NMR, respectively) were recorded on Bruker Avance 400 spectrometer. Chemical shifts (δ) are reported in units of parts per million (ppm) downfield from $SiMe_4$ (δ 0.0) and relative to the respective solvent's peak. Multiplicities are given as: s (singlet), d (doublet), t (triplet), dd (double of doublet) td (triplet of doublet) or m (multiplet). ^{1}H-^{1}H Coupling constants (*J*) are reported in Hertz (Hz).

8.2.2 Virtual combinatorial library – building blocks

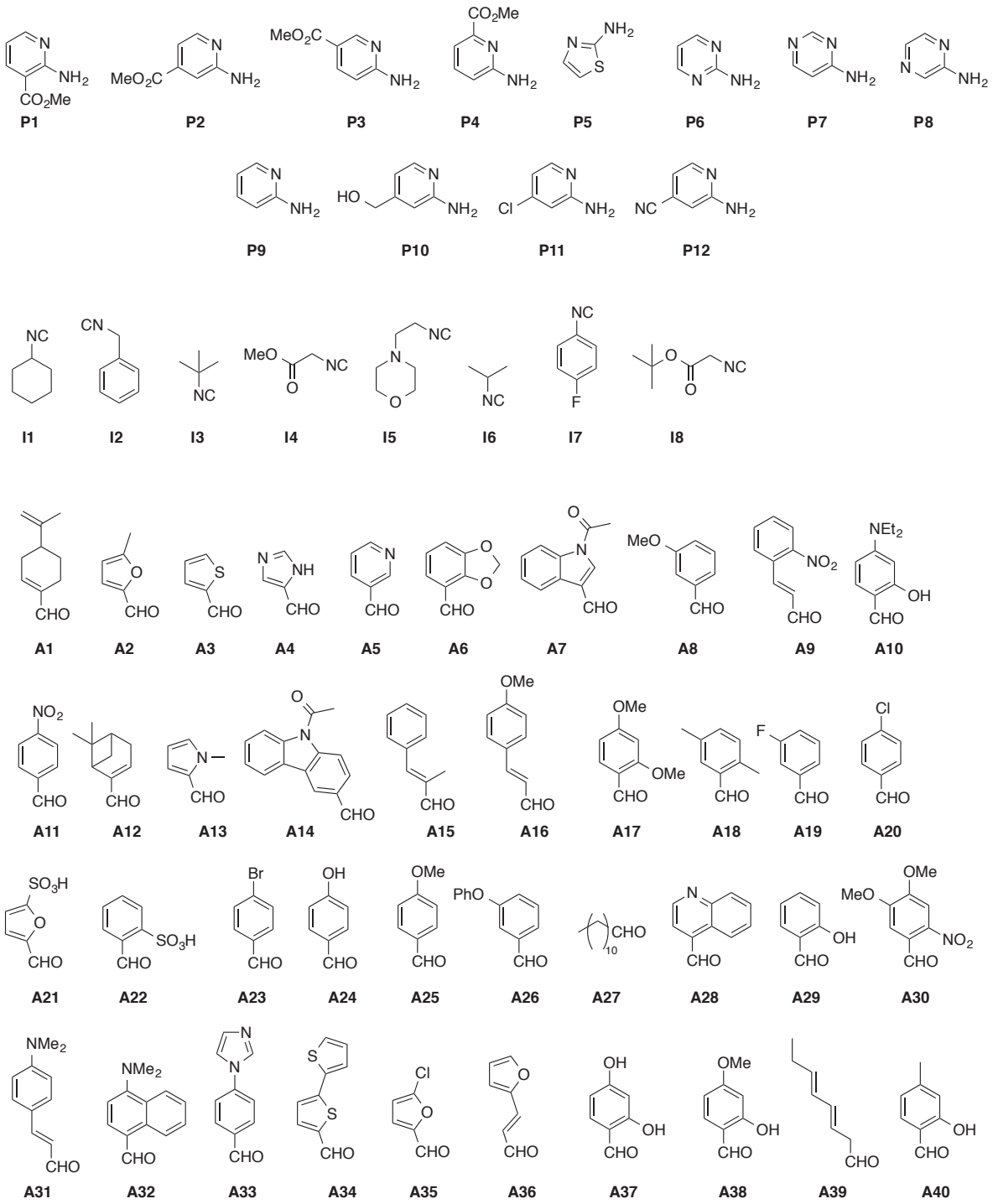
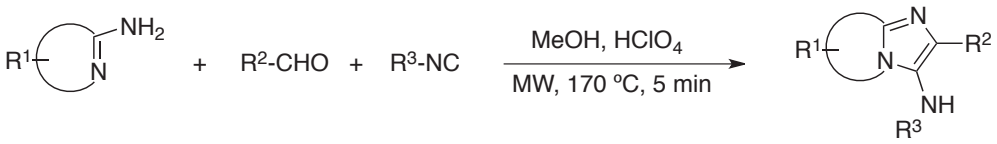


Figure S8. Aminopyridines (P), aldehydes (A) and isocyanides (I) building blocks described by Burchak *et al.*[1] used for enumerating the virtual combinatorial library

[1] Burchak ON, Mugherli L, Ostuni M, Lacapère JJ, Balakirev MY (2011) *J. Am. Chem. Soc.* **133**, 10058-10061.

8.2.3 Synthesis

General procedure for synthesis of imidazopyridine combinatorial library

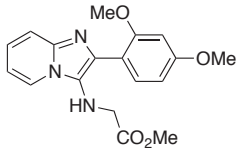


Stock solutions in MeOH were prepared for all building blocks (P, A and I) and catalyst. These solutions had a concentration of 400 mM for P, A and I starting materials, while for perchloric acid a solution of concentration 40 mM was prepared. An automated system dispensed 75 μL of each building block and catalyst, sequentially, into microwave vials. These were sealed and heated for 5 minutes, at 170 ºC under microwaves. The reaction products were analyzed under HPLC-MS using ACN : H_2O (+ 0.1% TFA in each phase) as eluent. A typical run used a gradient of 5-50% ACN run in 12 minutes.

General procedure for synthesis of selected imidazopyridines

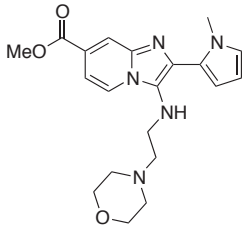
The protocol was adapted from existing literature.[2] 2-Aminopyridine (1.0 molar eq.), aldehyde (1.0 molar eq.) isocyanide (1.0 molar eq.) and perchloric acid (10 mol%) were dissolved in ethanol absolute (2.1 mL/mmol). The solutions were heated at 170 ºC for 5 minutes under microwaves. The resulting crudes were purified via preparative HPLC using ACN : H_2O (+ 0.1% TFA in each phase) as eluent. A typical run used a gradient of 5-50% ACN run over 16 minutes.

Methyl 2-((2-(2,4-dimethoxyphenyl)imidazo[1,2-*a*]pyridin-3-yl)amino)acetate, 2



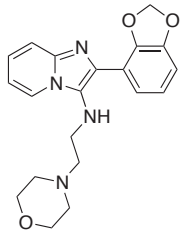
Yellow oil; 81%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 3.44 (3H, s, OCH_3), 3.67 (2H, s, CH_2), 3.75 (3H, s, OCH_3), 3.80 (3H, s, OCH_3), 6.57-6.61 (2H, m, Ar-H), 7.32-7.36 (1H, m, Ar-H), 7.53 (1H, d, *J* = 8.0 Hz, Ar-H), 6.77-7.75 (2H, m, Ar-H), 8.68 (1H, d, *J* = 2.4 Hz, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 48.54, 52.52, 56.18, 56.48, 99.77, 107.08, 108.44, 112.51, 117.43, 123.13, 126.42, 129.04, 132.62, 133.56, 137.46, 159.91, 164.60, 173.19. HRMS-ESI calc. ($C_{18}H_{19}N_3O_4$+H^+): 342.1448, found: 342.1448.

Methyl 2-(1-methyl-1*H*-pyrrol-2-yl)-3-((2-morpholinoethyl)amino)imidazo[1,2-*a*] pyridine-7-carboxylate, 3



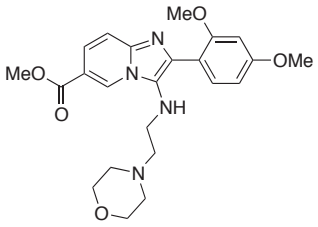
Yellow oil; 74%; (CD_3OD, 400.13 MHz): δ 3.08 (2H, m, CH_2), 3.21 (2H, t, *J* = 6.4 Hz, CH_2), 3.37 (2H, m, CH_2), 3.45 (2H, t, *J* = 6.4 Hz, CH_2), 3.72 (3H, s, CH_3), 3.81 (2H, m, CH_2), 3.97 (2H, m, CH_2), 4.05 (3H, s, CH_3), 6.32 (1H, dd, *J* = 3.8 Hz, Ar-H), 6.63 (1H, dd, *J* = 3.8 Hz Ar-H), 7.06 (1H, m, Ar-H), 7.93 (1H, dd, *J* = 1.6 and 7.2 Hz, Ar-H), 8.40 (1H, m, Ar-H), 8.80 (1H, dd, *J* = 0.8 and 7.2 Hz, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): 35.01, 41.04, 53.33, 53.84, 57.27, 64.81, 110.04, 114.52, 115.63, 116.03, 116.31, 118.78, 126.40, 127.62, 131.28, 134.24, 137.07, 165.19. HRMS-ESI calc. ($C_{20}H_{25}N_5O_3+H^+$): 384.2030, found: 384.2031.

2-(Benzo[*d*][1,3]dioxol-4-yl)-*N*-(2-morpholinoethyl)imidazo[1,2-*a*]pyridin-3-amine, *S*2



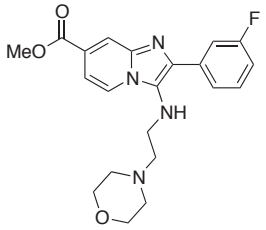
White solid; mp = 131-133 ºC, 75%, ^1H NMR (CD_3OD, 400.13 MHz): δ 3.00 (2H, m, CH_2), 3.21 (3H, t, *J* = 6.2 Hz, CH_2 + CH), 3.34 (3H, t, *J* = 6.4 Hz, CH_2 + CH), 3.56-3.94 (4H, m, CH_2), 6.01 (2H, s, CH_2), 6.87 (1H, dd, *J* = 1.8 and 8.2 Hz, Ar-H), 6.92 (1H, t, *J* = 8.2 Hz, Ar-H), 7.14 (1H, dd, *J* = 1.8 and 8.0 Hz, Ar-H), 7.37 (1H, td, *J* = 1.6 and 7.4 Hz, Ar-H), 7.74 (1H, m, Ar-H), 7.82 (1H, m, Ar-H), 8.65 (1H, m, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 42.50, 53.47, 57.66, 64.84, 103.47, 109.52, 111.50, 112.97, 118.24, 122.26, 122.86, 124.07, 126.52, 128.89, 134.85, 138.74, 146.90, 149.91. HRMS-ESI calc. ($C_{20}H_{22}N_4O_3+H^+$): 367.1765, found: 367.1765.

Methyl 2-(2,4-dimethoxyphenyl)-3-((2-morpholinoethyl)amino)imidazo[1,2-*a*]pyridine-6-carboxylate, S3



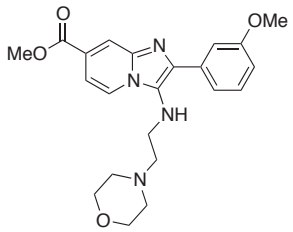
Yellow oil; 92%; ^1H NMR ($CDCl_3$, 400.13 MHz): δ 2.90 (2H, m, CH_2), 3.22 (2H, t, *J* = 6.4 Hz, CH_2), 3.32 (2H, t, *J* = 6.4 Hz, CH_2), 3.39 (2H, d, *J* = 11.6 Hz, CH_2), 3.77 (3H, s, OCH_3), 3.82 (3H, s, OCH_3), 3.84-4.00 (4H, m, CH_2), 4.04 (3H, s, OCH_3), 6.42 (1H, d, *J* = 2.0 Hz, Ar-H), 6.45 (1H, dd, *J* = 2.0 and 8.7 Hz, Ar-H), 7.54 (1H, d, *J* = 8.8 Hz, Ar-H), 7.83 (1H, dd, *J* = 1.0 and 9.2 Hz, Ar-H), 8.20 (1H, dd, *J* = 1.6 and 9.4 Hz, Ar-H), 9.11 (1H, s, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 41.89, 53.36, 53.59, 56.22, 56.52, 57.46, 64.79, 99.86, 107.21, 107.91, 112.67, 116.07, 121.89, 125.05, 129.26, 132.88, 133.31, 138.60, 160.23, 165.11, 165.15. HRMS-ESI calc. ($C_{23}H_{28}N_4O_5+H^+$): 441.2132, found: 441.2127.

Methyl 2-(3-fluorophenyl)-3-((2-morpholinoethyl)amino)imidazo[1,2-*a*]pyridine-7-carboxylate, S4



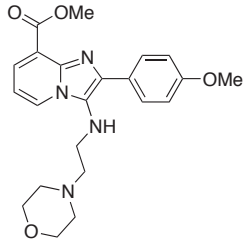
Yellow oil; 70%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.96 (2H, m, CH_2), 3.33 (2H, t, *J* = 6.4 Hz, CH_2), 3.44 (2H, t, *J* = 6.4 Hz, CH_2), 3.44-3.49 (2H, m, CH_2), 3.90-4.01 (4H, m, CH_2), 4.03 (3H, s, OCH_3), 7.11 (1H, td, *J* = 2.2 and 8.4 Hz, Ar-H), 7.45 (1H, m, Ar-H), 7.51 (1H, m, Ar-H), 7.59 (1H, d, *J* = 8.2 Hz, Ar-H), 7.88 (1H, dd, *J* = 1.6 and 7.0 Hz, Ar-H), 8.40 (1H, s, Ar-H), 8.71 (1H, dd, *J* = 1.0 and 7.0 Hz, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 42.56, 53.46, 53.86, 57.64, 64.83, 115.00, 116.12, 116.36, 116.62, 118.40, 118.61, 125.44, 126.73, 129.57, 132.77, 134.75, 138.17, 164.52, 165.21. HRMS-ESI calc. ($C_{21}H_{23}FN_4O_3$+H^+): 399.1827, found: 399.1828.

Methyl 2-(3-methoxyphenyl)-3-((2-morpholinoethyl)amino)imidazo[1,2-*a*]pyridine-7-carboxylate, S5



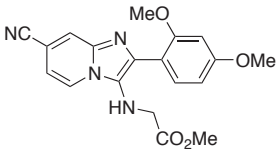
Yellow oil; 93%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.88 (2H, m, CH_2), 3.22 (2H, t, *J* = 6.4 Hz, CH_2), 3.29 (2H, t, *J* = 6.4 Hz, CH_2), 3.37 (2H, m, CH_2), 3.68 (3H, s, CH_3), 3.82-3.91 (4H, m, CH_2), 3.95 (3H, s, OCH_3), 6.66 (1H, dd, *J* = 1.6 and 8.2 Hz, Ar-H), 7.06-7.20 (3H, m, Ar-H), 7.72 (1H, dd, *J* = 1.4 and 7.2 Hz, Ar-H), 8.16 (1H, s Ar-H), 8.55 (1H, d, *J* = 7.2 Hz, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): 41.35, 52.70, 53.52, 55.36, 57.41, 63.53, 112.62, 114.04, 116.17, 116.82, 119.82, 124.39, 126.43, 126.71, 129.48, 130.40, 133.38, 160.18, 161.46, 163.30. HRMS-ESI calc. ($C_{22}H_{26}N_4O_4$+H^+): 411.2027, found: 411.2027.

Methyl 2-(4-methoxyphenyl)-3-((2-morpholinoethyl)amino)imidazo[1,2-*a*]pyridine-8-carboxylate, S6



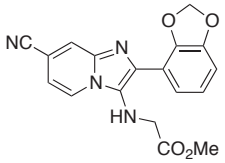
Yellow oil; 72%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.88 (2H, m, CH_2), 3.20 (2H, t, *J* = 6.4 Hz, CH_2), 3.37-3.42 (4H, m, CH_2), 3.81 (3H, s, OCH_3), 3.82-4.03 (4H, m, CH_2), 4.03 (3H, s, OCH_3), 6.98 (2H, d, *J* = 9.0 Hz, Ar-H), 7.46 (1H, dd, *J* = 7.2 Hz Ar-H), 7.60 (2H, d, *J* = 9.0 Hz, Ar-H), 8.38 (1H, dd, *J* = 1.0 and 7.4 Hz, Ar-H), 7.98 (1H, dd, *J* = 1.0 and 7.0 Hz, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 42.46, 53.37, 53.86, 56.07, 57.63, 64.79, 115.73, 116.53, 117.69, 119.14, 128.30, 129.59, 130.70 132.39, 136.29, 137.34, 163.12, 164.19. HRMS-ESI calc. ($C_{22}H_{26}N_4O_4$+H^+): 411.2027, found: 411.2028.

Methyl 2-((7-cyano-2-(2,4-dimethoxyphenyl)imidazo[1,2-*a*]pyridin-3-yl)amino)acetate, S7



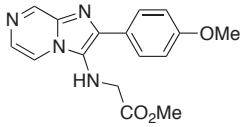
Yellow solid; mp = 60-61 ºC; 87%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 3.55 (3H, s, OCH_3), 3.80 (2H, s, CH_2), 3.90 (3H, s, OCH_3), 3.92 (3H, s, OCH_3), 6.73 (1H, d, *J* = 2.4 Hz, Ar-H), 6.76 (1H, dd, *J* = 4.8 and 2.0 Hz, Ar-H), 7.62-7.68 (2H, m, Ar-H), 8.28 (1H, s, Ar-H), 8.89 (1H, d, *J* = 7.2 Hz, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 48.17, 52.56, 56.22, 56.51, 99.83, 107.29, 108.31, 114.51, 117.25, 117.60, 118.58, 126.04, 127.11, 131.05, 132.79, 135.66, 160.07, 165.05, 173.04. HRMS-ESI calc. ($C_{19}H_{18}N_4O_4$+H^+): 367.1401, found: 367.1401.

Methyl 2-((2-(benzo[*d*][1,3]dioxol-4-yl)-7-cyanoimidazo[1,2-*a*]pyridin-3-yl)amino) acetate, S8



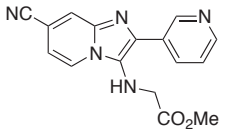
Yellow solid; mp = 160-162 ºC; 56%; ^{1}H NMR (DMSO-d_6, 400.13 MHz): δ 3.49 (3H, s, OCH_3), 3.84 (2H, d, *J* = 6.4 Hz, CH_2), 5.61 (1H, t, *J* = 6.4 Hz, NH), 6.10 (2H, s, CH_2), 6.95-6.97 (2H, m, Ar-H), 7.20 (1H, d, *J* = 7.2 Hz, Ar-H), 7.31 (1H, dd, *J* = 4.4 Hz, Ar-H), 8.23 (1H, s, Ar-H), 8.47 (1H, d, *J* = 6.8 Hz, Ar-H). ^{13}C NMR (DMSO-d_6, 100.61 MHz): δ 47.53, 51.57, 100.83, 103.97, 107.98, 111.44, 115.88, 118.44, 121.73 121.77, 123.34, 124.10, 129.74, 131.54, 138.01, 144.21, 147.36, 171.58. HRMS-ESI calc. ($C_{18}H_{14}N_4O_4$+H^+): 351.1088, found: 351.1087.

Methyl 2-((2-(4-methoxyphenyl)imidazo[1,2-*a*]pyrazin-3-yl)amino)acetate, S9



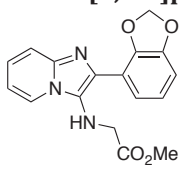
Yellow solid; mp = 150-151 ºC; 67%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 3.50 (3H, s, OCH_3), 3.78 (3H, s, OCH_3), 3.87 (2H, s, CH_2), 7.00 (2H, d, *J* = 6.8 Hz, Ar-H), 7.83-7.87 (3H, m, Ar-H), 8.56 (1H, d, *J* = 4.8 Hz, Ar-H), 8.92 (1H, s, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 48.06, 52.67, 55.96, 115.55, 118.72, 123.80, 124.50, 130.54, 133.56, 134.82, 135.96, 141.33, 162.58, 173.27. HRMS-ESI calc. ($C_{16}H_{16}N_4O_3$+H^+): 313.1295, found: 313.1295.

Methyl 2-((7-cyano-2-(pyridin-3-yl)imidazo[1,2-*a*]pyridin-3-yl)amino)acetate, S10



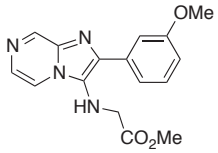
Yellow solid; mp = 183-185 ºC; 39%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 3.54 (3H, s, CH_3), 3.93 (2H, s, CH_2), 7.28 (1H, dd, *J* = 7.2 and 2.0 Hz, Ar-H), 7.84 (1H, ddd, *J* = 8.0, 5.6 and 0.4 Hz, Ar-H), 8.29 (1H, dd, *J* = 1.6 and 1.2 Hz, Ar-H), 8.63 (1H, dd, *J* = 7.2 and 0.8 Hz, Ar-H), 8.71 (1H, dd, *J* = 5.2 and 0.8 Hz, Ar-H), 8.80 (1H, dt, *J* = 8.0 and 2.0 Hz, Ar-H), 9.40 (1H, d, *J* = 2.0 Hz, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 47.93, 51.65, 105.81, 111.63, 118.09, 123.53, 125.27, 125.42, 130.11, 130.94, 131.82, 137.72, 138.61, 143.68, 144.48, 171.97. HRMS-ESI calc. ($C_{16}H_{13}N_5O_2$+H^+): 308.1142, found: 308.1143.

Methyl 2-((2-(benzo[*d*][1,3]dioxol-4-yl)imidazo[1,2-*a*]pyridin-3-yl)amino)acetate, S11



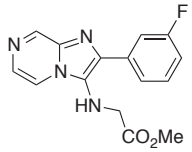
Yellow oil; 80%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 3.36 (3H, s, OCH_3), 3.68 (2H, s, CH_2), 5.93 (2H, s, CH_2), 6.76 (1H, d, *J* = 8.0 Hz, Ar-H), 6.82 (1H, t, *J* = 8.0 Hz, Ar-H), 7.10 (1H, d, *J* = 8.0 Hz, Ar-H), 7.29 (1H, t, *J* = 7.2 Hz, Ar-H), 7.65 (1H, d, *J* = 9.2 Hz, Ar-H), 7.72 (1H, m, Ar-H), 8.63 (1H, d, *J* = 6.8 Hz, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 48.84, 52.57, 103.33, 109.86, 111.18, 112.83, 117.87, 121.63, 123.98, 126.80, 129.62, 134.50, 134.57, 138.48, 146.54, 149.77, 173.17. HRMS-ESI calc. ($C_{17}H_{15}N_3O_4$+H^+): 326.1135, found: 326.1133.

Methyl 2-((2-(3-methoxyphenyl)imidazo[1,2-*a*]pyrazin-3-yl)amino)acetate, S12



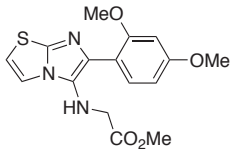
Yellow solid; mp = 137-138 ºC; 61%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 3.60 (3H, s, OCH_3), 3.88 (3H, s, OCH_3), 3.92 (2H, s, CH_2), 7.06 (1H, ddd, *J* = 8.6, 2.6 and 1.2 Hz, Ar-H), 7.45 (1H, t, *J* = 8.0 Hz, Ar-H), 7.53-7.60 (2H, m, Ar-H), 7.95 (1H, d, *J* = 5.4 Hz, Ar-H), 8.66 (1H, dd, *J* = 5.3 and 1.2 Hz, Ar-H), 9.05 (1H, d, *J* = 0.8 Hz, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 48.21, 52.66, 55.90, 114.35, 116.45, 118.91, 121.18, 121.24, 124.15, 131.20, 133.94, 135.13, 137.42, 140.56, 161.60, 173.27. HRMS-ESI calc. ($C_{16}H_{16}N_4O_3$+H^+): 313.1295, found: 313.1294.

Methyl 2-((2-(3-fluorophenyl)imidazo[1,2-*a*]pyrazin-3-yl)amino)acetate, S13



Yellow solid; mp = 170-171 ºC; 46%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 3.50 (3H, s, OCH_3), 3.87 (2H, s, CH_2), 7.12 (1H, tdd, *J* = 8.6, 2.4, 0.8 Hz, Ar-H), 7.46 (1H, m, Ar-H), 7.69 (1H, m, Ar-H), 7.77 (1H, m, Ar-H), 7.83 (1H, d, *J* = 5.2 Hz, Ar-H), 8.57 (1H, dd, *J* = 5.2 and 1.2 Hz, Ar-H), 8.97 (1H, d, *J* = 1.0 Hz, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 48.41, 52.66, 115.41, 115.65, 117.11, 117.32, 119.29, 123.94, 124.72, 131.92, 135.44, 138.33, 163.28, 165.71, 173.31. HRMS-ESI calc. ($C_{15}H_{13}FN_4O_2$+H^+): 301.1095, found: 301.1097.

Methyl 2-((6-(2,4-dimethoxyphenyl)imidazo[2,1-*b*]thiazol-5-yl)amino)acetate, S14

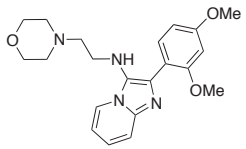


Yellow oil; 52%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 3.39 (3H, s, OCH_3), 3.58 (2H, s, CH_2), 3.68 (3H, s, OCH_3), 3.72 (3H, s, OCH_3), 6.46-6.53 (2H, m, Ar-H), 7.36 (1H, m, Ar-H), 7.46 (1H, m, Ar-H), 7.96 (1H, dd, *J* = 4.4 and 1.2 Hz, Ar-H). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 48.78, 52.53, 56.12, 56.40, 99.75, 106.88, 109.20, 117.89, 121.03, 123.30, 130.74, 132.10, 143.16, 159.66, 164.09, 173.34. HRMS-ESI calc. ($C_{16}H_{17}N_3O_4S$+H^+): 348.1013, found: 348.1013.

8.3 Supporting Information – Combining on-chip synthesis of a focused combinatorial library with *in silico* target prediction reveals imidazopyridine GPCR ligands

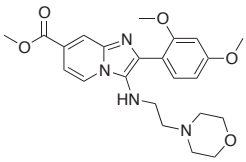
8.3.1 Synthesis

2-(2,4-Dimethoxyphenyl)-*N*-(2-morpholinoethyl)imidazo[1,2-*a*]pyridin-3-amine, formic acid salt (4)



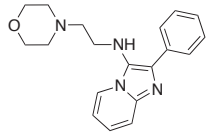
Orange oil; 49%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.15 (4H, m, CH_2), 2.26 (2H, m, CH_2), 2.86 (2H, m, CH_2), 3.51 (4H, m, CH_2), 3.79 (3H, s, OCH_3), 3.84 (3H, s, OCH_3), 6.51 (1H, d, *J* = 2.0 Hz, Ar-H), 6.64 (1H, dd, *J* = 2.4 and 8.4 Hz, Ar-H), 6.88 (1H, td, *J* = 1.6 and 6.8 Hz, Ar-H), 7.22 (1H, m, Ar-H), 7.59 (1H, d, *J* = 8.8 Hz, Ar-H), 7.75 (1H, d, *J* = 8.8 Hz, Ar-H), 8.10 (1H, d, *J* = 6.8 Hz, Ar-H), 8.30 (1H, s, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 43.73, 53.24, 55.53, 56.05, 57.72, 66.70, 99.10, 105.60, 113.04, 113.45, 116.03, 122.44, 125.67, 127.73, 129.22, 132.22, 140.18, 157.31, 161.48, 165.68. HRMS-ESI calc. ($C_{21}H_{26}N_4O_3$+H^+): 383.2078, found: 383.2070.

Methyl 2-(2,4-dimethoxyphenyl)-3-((2-morpholinoethyl)amino) imidazo[1,2-*a*] pyridine-7-carboxylate, formic acid salt (5)



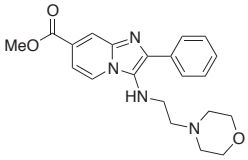
Yellow oil; 35%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.42 (4H, m, CH_2), 2.49 (2H, m, CH_2), 3.05 (2H, m, CH_2), 3.66 (4H, m, CH_2), 3.89 (3H, s, OCH_3), 3.92 (3H, s, OCH_3), 4.00 (3H, s, OCH_3), 6.61 (1H, d, *J* = 2.0 Hz, Ar-H), 6.70 (1H, dd, *J* = 2.0 and 8.8 Hz, Ar-H), 7.53 (1H, dd, *J* = 1.2 and 7.2 Hz, Ar-H), 7.67 (1H, d, *J* = 8.8 Hz, Ar-H), 8.16 (1H, d, *J* = 7.6 Hz, Ar-H), 8.26 (2H, br.s, NH), 8.42 (1H, s, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 44.35, 52.97, 54.49, 56.04, 56.51, 59.04, 67.91, 98.81, 99.80, 106.93, 112.07, 116.53, 117.72, 119.59, 123.49, 125.81, 131.48, 133.19, 159.02, 163.17, 167.29, 189.94. HRMS-ESI calc. ($C_{23}H_{27}N_4O_5$+H^+): 442.3232, found: 441.2133.

***N*-(2-Morpholinoethyl)-2-phenylimidazo[1,2-*a*]pyridin-3-amine, formic acid salt (6)**



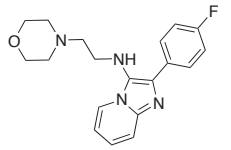
Brown oil; 9%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.42 (4H, m, CH_2), 2.52 (2H, m, CH_2), 3.03 (2H, m, CH_2), 3.66 (4H, m, CH_2), 6.81 (1H, td, *J* = 1.2 and 6.8 Hz, Ar-H), 7.15 (1H, m, Ar-H), 7.27 (1H, t, *J* = 7.2 Hz, Ar-H), 7.38 (2H, m, Ar-H), 6.63 (1H, d, *J* = 8.8 Hz, Ar-H), 8.85 (2H, d, *J* = 8.0 Hz, Ar-H), 8.11 (1H, d, *J* = 6.8 Hz, Ar-H), 8.28 (1H, s, NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 43.77, 53.39, 58.06, 66.47, 112.57, 116.82, 122.59, 125.20, 126.33, 127.41, 127.53, 127.87, 128.75, 132.90, 140.85, 165.39. HRMS-ESI calc. ($C_{19}H_{21}N_4O$+H^+): 323.1866, found: 323.1860.

Methyl 3-((2-morpholinoethyl)amino)-2-phenylimidazo[1,2-*a*]pyridine-7-carboxylate (7)



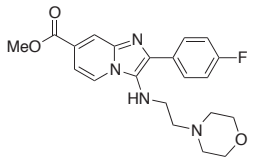
Yellow oil; <5%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 2.38 (4H, m, CH_2), 2.54 (2H, m, CH_2), 3.19 (2H, m, CH_2), 3.54 (4H, m, CH_2), 3.98 (3H, s, OCH_3), 7.41 (1H, m, Ar-H), 7.42-7.53 (3H, m, Ar-H), 8.03 (2H, d, *J* = 8.8 Hz, Ar-H), 8.20 (1H, d, *J* = 1.2 Hz, Ar-H), 8.45 (1H, dd, *J* = 1.0 and 7.2 Hz, Ar-H). HRMS-ESI calc. ($C_{21}H_{23}N_4O_3$+H$^+$): 381.1921, found: 381.1915.

2-(4-Fluorophenyl)-*N*-(2-morpholinoethyl)imidazo[1,2-*a*]pyridin-3-amine (8)



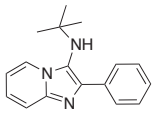
Brown oil; 15%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.41 (4H, m, CH_2), 2.49 (2H, m, CH_2), 3.01 (2H, m, CH_2), 3.67 (4H, m, CH_2), 6.73 (1H, t, *J* = 6.4 Hz, Ar-H), 7.04-7.10 (3H, m, Ar-H), 7.49 (1H, d, *J* = 8.8 Hz, Ar-H), 7.94 (2H, m, Ar-H), 8.05 (1H, d, *J* = 6.8 Hz, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 43.90, 53.66, 58.38, 66.76, 111.93, 115.63, 117.35, 122.42, 124.21, 128.81, 130.31, 134.06, 141.24, 161.02, 163.47. HRMS-ESI calc. ($C_{19}H_{20}FN_4O$+H$^+$): 341.1772, found: 341.1765.

Methyl 2-(4-fluorophenyl)-3-((2-morpholinoethyl)amino) imidazo[1,2-*a*]pyridine-7-carboxylate (9)



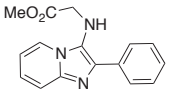
Yellow oil; <5%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 2.42 (4H, m, CH_2), 2.56 (2H, m, CH_2), 3.19 (2H, m, CH_2), 3.55 (4H, m, CH_2), 3.98 (3H, s, OCH_3), 7.25 (2H, m, Ar-H), 7.48 (1H, d, *J* = 1.6 and 7.2 Hz, Ar-H), 8.09 (2H, m, Ar-H), 8.19 (1H, d, *J* = 0.8 and 1.6 Hz, Ar-H), 8.45 (1H, dd, *J* = 0.8 and 7.4 Hz, Ar-H). HRMS-ESI calc. ($C_{21}H_{22}FN_4O_3$+H$^+$): 399.1827, found: 3991824.

***N*-(*Tert*-butyl)-2-phenylimidazo[1,2-*a*]pyridin-3-amine, formic acid salt (10)**



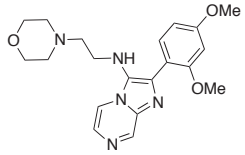
White oil; 8%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 0.91 (9H, s, $(CH_3)_3$), 6.77 (1H, t, *J* = 6.8 Hz, Ar-H), 7.14 (1H, m, Ar-H), 7.26 (1H, t, *J* = 7.2 Hz, Ar-H), 7.37 (2H, t, *J* = 7.2 Hz, Ar-H), 7.60 (1H, d, *J* = 8.8 Hz, Ar-H), 7.78 (2H, d, *J* = 6.8 Hz, Ar-H), 8.20 (2H, d, *J* = 6.8 Hz, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 30.24, 56.50, 112.06, 116.70, 123.61, 125.23, 127.83, 128.32, 128.43, 128.53, 133.98, 138.42, 141.46, 164.88. HRMS-ESI calc. ($C_{17}H_{18}N_3$+H$^+$): 266.1652, found: 266.1646.

Methyl 2-((2-phenylimidazo[1,2-*a*]pyridin-3-yl)amino)acetate (11)



Yellow oil; 8%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 3.65 (3H, s, CH_3), 3.74 (2H, s, CH_2), 6.82 (1H, td, *J* = 1.2 and 6.8 Hz, Ar-H), 7.15 (1H, m, Ar-H), 7.27 (1H, t, *J* = 7.2 Hz, Ar-H), 7.37 (2H, t, *J* = 7.6 Hz, Ar-H), 7.58 (1H, d, *J* = 9.2 Hz, Ar-H), 7.90 (2H, d, *J* = 8.4 Hz, Ar-H), 8.22 (1H, d, *J* = 6.8 Hz, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 49.75, 52.35, 113.48 116.56, 120.02, 121.46, 125.21, 127.09, 128.31, 128.90, 129.73, 134.36, 142.32, 173.91. HRMS-ESI calc. ($C_{16}H_{15}N_3O_2$+H$^+$): 282.8243, found: 282.1234.

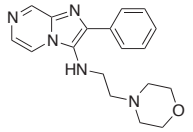
2-(2,4-Dimethoxyphenyl)-*N*-(2-morpholinoethyl)imidazo[1,2-*a*]pyrazin-3-amine (12)



Yellow oil; 28%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.30 (4H, m, CH_2), 2.39 (2H, m, CH_2), 3.01 (2H, m, CH_2), 3.61 (4H, m, CH_2), 3.89 (3H, s, OCH_3), 3.93 (3H, s, OCH_3), 6.62 (1H, d, *J* = 2.0 Hz, Ar-H), 6.68 (1H, dd, *J* = 2.0 and 8.8 Hz, Ar-H), 7.69 (1H, d, *J* = 8.4 Hz, Ar-H), 7.90 (1H, d, *J* = 4.8 Hz, Ar-H), 8.03 (1H, dd, *J* = 1.2 and 8.8 Hz, Ar-H), 8.21 (1H, s, NH), 9.04 (1H, d, *J* = 1.6 Hz, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ

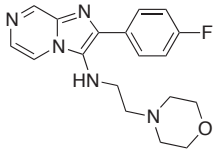
43.21, 53.27, 55.55, 56.08, 57.84, 66.61, 98.90, 105.72, 115.20, 115.60, 128.76, 129.27, 132.37, 135.12, 136.91, 143.13, 157.37, 161.36. HRMS-ESI calc. ($C_{20}H_{24}N_5O_3$+H^+): 384.2030, found: 384.2036.

_N_-(2-morpholinoethyl)-2-phenylimidazo[1,2-_a_]pyrazin-3-amine (13)



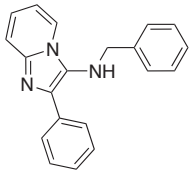
Yellow oil; <5%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 2.50 (4H, m, CH_2), 2.65 (2H, m, CH_2), 3.25 (2H, m, CH_2), 3.57 (4H, m, CH_2), 7.44 (1H, m, Ar-H), 7.54 (2H, m, Ar-H), 7.90 (1H, d, *J* = 4.4 Hz, Ar-H), 8.06 (2H, m, Ar-H), 8.41 (1H, dd, *J* = 2.0 and 4.8 Hz, Ar-H), 8.90 (1H, d, *J* = 1.4 Hz, Ar-H). HRMS-ESI calc. ($C_{18}H_{20}N_5O$+H^+): 324.1819, found: 324.1815.

2-(4-Fluorophenyl)-_N_-(2-morpholinoethyl)imidazo[1,2-_a_]pyrazin-3-amine (14)



Yellow oil; <5%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.40-2.60 (6H, m, CH_2), 3.09 (2H, m, CH_2), 3.72 (4H, m, CH_2), 7.10 (2H, m, Ar-H), 7.81 (1H, d, *J* = 4.4 Hz, Ar-H), 7.92-8.00 (3H, m, Ar-H), 8.92 (1H, d, *J* = 1.4 Hz, Ar-H). HRMS-ESI calc. ($C_{18}H_{20}FN_5O$+H^+): 342.1725, found: 342.1732.

_N_-Benzyl-2-phenylimidazo[1,2-_a_]pyridin-3-amine, formic acid salt (15)



Yellow solid; mp = 119-121ºC; 35%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 4.16 (2H, s, CH_2), 7.06 (1H, t, *J* = 6.8 Hz, Ar-H), 7.16-7.18 (5H, m, Ar-H), 7.41 (1H, m, Ar-H), 7.47-7.52 (3H, m, Ar-H), 7.59 (1H, d, *J* = 9.2 Hz, Ar-H), 7.92 (2H, d, *J* = 7.2 Hz, Ar-H), 8.25 (1H, s, NH), 8.31 (1H, d, *J* = 6.8 Hz, Ar-H). ^{3}C NMR (CD_3OD, 100.61 MHz): δ 52.56, 115.01, 115.31, 125.04, 128.27, 128.48, 128.60, 129.45, 129.53, 129.61, 129.68, 129.89, 132.29, 132.83, 140.45, 140.84, 166.22. HRMS-ESI calc. ($C_{20}H_{17}N_3$+H^+): 300.1495, found: 300.1499.

8.3.2 Functional assay

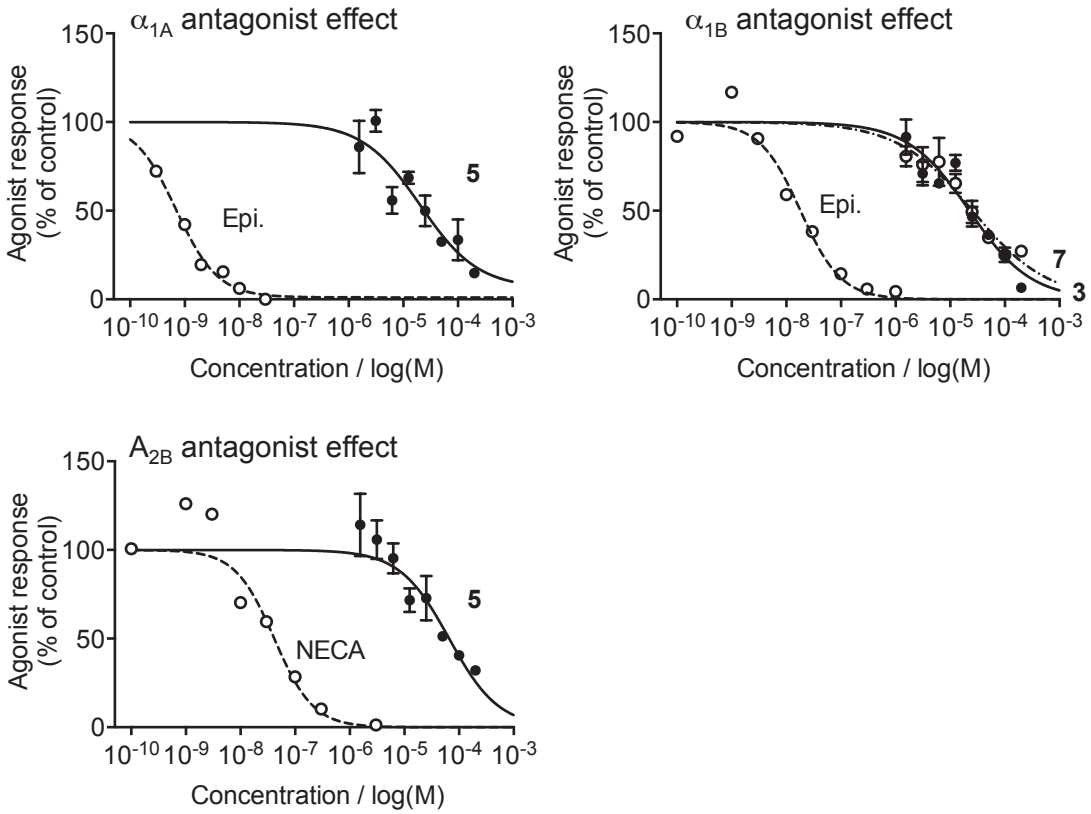


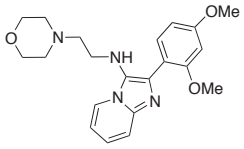
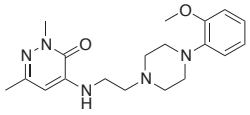
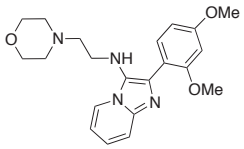
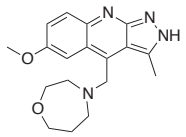
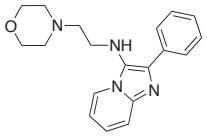
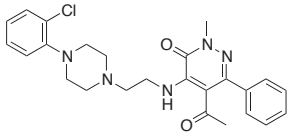
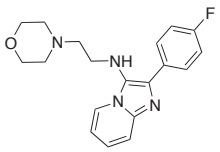
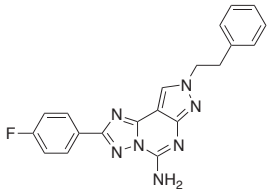
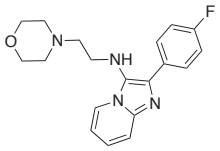
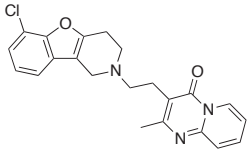
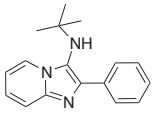
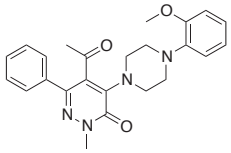
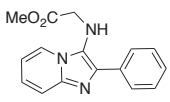
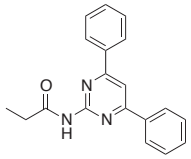
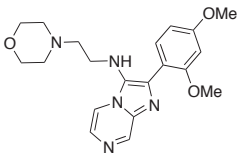
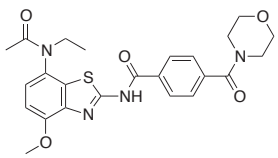
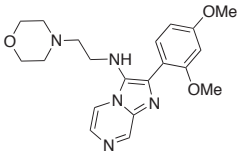
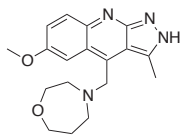
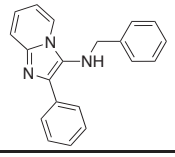
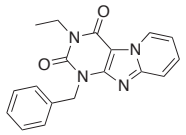
Figure S9. Functional assay results for active compounds **6**, **8**, **10**. Results for hA_1 are not shown (inactive compounds).

Table S1. Parameters of the functional GPCR assays.

Receptor / Assay	Source	Stimulus	Incubation	Measured component	Detection method
hA2B (antagonist effect)	human recombinant (HEK-293 cells)	NECA (1000 nM)	10 min, 37ºC	cAMP	HTRF
ha_{1A} (antagonist effect)	human recombinant (CHO cells)	Epinephrine (3 nM)	RT	intacellular $[Ca^{2+}]$	Fluorimetry
ha_{1B} (antagonist effect)	human recombinant (CHO cells)	Epinephrine (3000 nM)	30 min, 37ºC	cAMP	HTRF
hA1 (antagonist effect)	human recombinant (CHO cells)	CPA (1 nM)	28ºC	impedance	Cellular dielectric spectroscopy

8.3.3 Computational

Table S2. Imidazopyridines synthesized and tested in this study, their nearest neighbours from the training data for the respective target and their activities, and their pairwise structural similarity expressed as structure-based Tanimoto similarity index (Morgan fingerprints with radius = 3). ChEMBL IDs given without "CHEMBL" prefix.

Imidazopyridines			Nearest neighbours			
Cmp. No.	**Chemical structure**	**Target**	**Chemical structure**	**ChEMBL ID**	**pK_i**	**Tc**
4		α_{1A}		80132	7.7	0.22
4		PDE10A		1940054	6.8	0.23
6		α_{1B}		310599	9.1	0.24
8		A_{2B}		1093432	5.0	0.18
8		α_{1A}		164612	6.6	0.22
10		α_{1B}		78584	6.6	0.16
11		A_1		222718	7.4	0.28
12		A_{2B}		1170134	7.9	0.18
12		PDE10A		1940054	6.8	0.22
15		A_1		113512	6.5	0.30

8.4 Supporting Information – Multi-objective molecular *de novo* design by adaptive fragment prioritization

8.4.1 Experimental section

General considerations

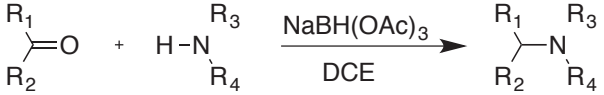
Starting materials and solvents were purchased from Sigma-Aldrich, Fluka, Chembridge, Maybridge or ABCR and were used without further purification. Syntheses were performed on a Radleys Tech Carousel with 12 reaction stations. Melting points (mp) were recorded on a Büchi M560 apparatus and are uncorrected. Proton and carbon nuclear magnetic resonance (^{1}H and ^{13}C NMR) spectra were recorded on a Bruker Avance 400 (400 and 100 MHz, respectively). All chemical shifts are quoted on the δ scale in ppm using residual solvent peaks as the internal standard. Coupling constants (*J*) are reported in Hz with the following splitting abbreviations: s = singlet, br.s. = broad singlet, d = doublet, dd = doublet of doublets, t = triplet, td = triplet of doublets, q = quartet, m = multiplet.

Analytical HPLC-MS was carried out in a Shimadzu LC-MS2020 system, equipped with a Nucleodur C_{18} HTec column, under an appropriate gradient of acetonitrile: H_2O (+ 0.1% formic acid in each solvent), and a total flow rate of 0.5 mL/min. High resolution mass spectrometry (HRMS) analyses were performed on a Bruker Daltonics maXis ESI-QTOF device. Mass spectrometry analyses were operated in positive-ion mode with ESI. Nominal and exact *m/z* values are reported in Daltons.

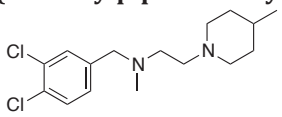
Flash chromatography was performed on a Biotage Isolera One device equipped with SNAP cartridges KP-C_{18}-HS 12, 30 or 60 g. All compounds present purity ≥ 95% based on LC-MS analysis.

Synthesis

General synthesis scheme for compounds **17-32**.

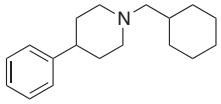


_N_-(3,4-Dichlorobenzyl)-_N_-methyl-2-(4-methylpiperidin-1-yl)ethanamine, formic acid salt (17)



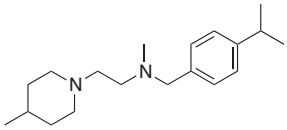
Educt 1: 1-(3,4-dichlorophenyl)-_N_-methylmethanamine; Educt 2: 2-(4-methylpiperidin-1-yl)acetaldehyde. White oil; 40%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 1.02 (3H, d, J = 6.0 Hz, CH_3), 1.62-1.70 (3H, m, CH_2+CH), 1.80 (2H, m, CH_2), 2.23 (3H, s, CH_3), 2.63 (2H, m, CH_2), 2.83 (2H, t, J = 6.4 Hz, CH_2), 3.06 (2H, t, J = 6.4 Hz, CH_2), 3.51-3.54 (4H, m, CH_2), 7.16 (1H, d, J = 2.0 and 8.4 Hz, Ar-H), 7.40-7.42 (2H, m, Ar-H), 8.45 (1H, br.s., NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 21.05, 29.37, 31.23, 42.07, 52.22, 52.72, 53.98, 61.33, 128.16, 130.32, 130.64, 131.15, 132.42, 138.86, 167.76. HRMS-ESI calc. ($C_{16}H_{23}Cl_2N_2$+H^+): 315.1389, found: 315.1384.

1-(Cyclohexylmethyl)-4-phenylpiperidine, formic acid salt (18)



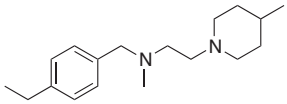
Educt 1: 4-phenylpiperidine; Educt 2: cyclohexanecarbaldehyde. White powder; mp = 61-62ºC 71%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 1.13-1.32 (5H, m, CH_2+CH), 1.65-1.86 (6H, m, CH_2), 1.99 (2H, d, J = 14 Hz, CH_2), 2.29 (2H, m, CH_2), 2.70-2.86 (5H, m, CH_2+CH), 3.69 (2H, m, CH_2), 7.21-7.25 (3H, m, Ar-H), 7.31-7.34 (2H, m, Ar-H), 8.47 (1H, br.s, NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 25.55, 25.71, 29.73, 31.35, 33.23, 40.50, 53.19, 62.97, 126.70, 127.02, 128.77, 143.28, 166.76. HRMS-ESI calc. ($C_{18}H_{26}$+H^+): 258.2216, found: 258.2214.

_N_-(4-Isopropylbenzyl)-_N_-methyl-2-(4-methylpiperidin-1-yl)ethanamine, formic acid salt (19)



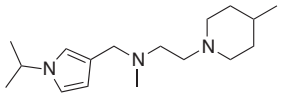
Educt 1: _N_-methyl-2-(4-methylpiperidin-1-yl)ethanamine; Educt 2: 4-isopropylbenzaldehyde. White oil; 96%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 0.98 (3H, d, J = 6.0 Hz, CH_3), 1.25 (6H, d, J = 7.2 Hz, CH_3), 1.50-1.71 (3H, m, CH_2+CH), 1.72 (2H, m, CH_2), 2.19 (3H, s, CH_3), 2.52-2.52 (2H, m, CH_2), 2.83 (2H, t, J = 7.2 Hz, CH_2), 2.93 (1H, m, CH), 3.00 (2H, t, J = 7.2 Hz, CH_2), 3.38 (2H, m, CH_2), 3.58 (2H, s, CH_2), 7.20-7.25 (4H, m, Ar-H), 8.50 (1H, br.s., NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 21.18, 24.00, 29.50, 31.72, 33.77, 42.02, 51.97, 52.59, 53.87, 62.00, 126.41, 129.28, 134.86, 148.16, 168.18. HRMS-ESI calc. ($C_{19}H_{31}N_2$+H^+): 289.2638, found: 289.2636.

_N_-(4-Ethylbenzyl)-_N_-methyl-2-(4-methylpiperidin-1-yl)ethanamine, formic acid salt (20)



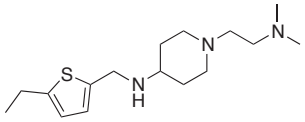
Educt 1: 1-(4-ethylphenyl)-_N_-methylmethanamine; Educt 2: 2-(4-methylpiperidin-1-yl)acetaldehyde. White oil; 74%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 0.95 (3H, d, J = 5.2 Hz, CH_3), 1.21 (3H, t, J = 7.6 Hz, CH_3), 1.52-1.57 (3H, m, CH_2+CH), 1.71 (2H, m, CH_2), 2.82 (3H, s, CH_3), 2.54-2.65 (4H, m, CH_2), 2.82 (2H, t, J = 6.4 Hz, CH_2), 3.01 (2H, t, J = 6.4 Hz, CH_2), 3.37 (2H, d, J = 12 Hz, CH_2), 3.57 (2H, s, CH_2), 7.15 (2H, d, J = 8.0 Hz, Ar-H), 7.20 (2H, d, J = 8.0 Hz, Ar-H), 8.44 (1H, br.s., NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 15.58, 20.96, 28.49, 29.07, 31.05, 41.67, 51.05, 52.43, 53.12, 61.82, 127.97, 129.48, 133.81, 143.86, 167.14. HRMS-ESI calc. ($C_{18}H_{29}N_2$+H^+): 275.2482, found: 275.2476.

_N_-((1-Isopropyl-1_H_-pyrrol-3-yl)methyl)-_N_-methyl-2-(4-methylpiperidin-1-yl)ethanamine, formic acid salt (21)



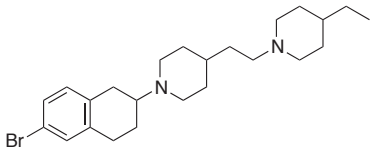
Educt 1: 1-(1-isopropyl-1*H*-pyrrol-3-yl)-*N*-methylmethanamine; Educt 2: 2-(4-methylpiperidin-1-yl)acetaldehyde. Yellow oil; 34%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 0.97 (3H, d, *J* = 6.0 Hz, CH_3), 1.39-1.47 (9H, m), 1.71 (2H, m, CH_2), 2.39 (2H, m, CH_2), 2.57 (3H, s, CH_3), 2.98-3.19 (4H, m, CH_2), 3.17 (2H, m, CH_2), 3.91 (2H, s, CH_2), 4.22 (1H, m, CH), 6.13 (1H, m, Ar-H), 6.71 (1H, m, Ar-H), 6.81 (1H, m, Ar-H), 8.49 (2H, br.s., NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 21.35, 23.85, 29.82, 32.60, 39.86, 49.97, 50.98, 52.78, 53.15, 53.26, 109.86, 112.60, 119.00, 119.96, 167.70. HRMS-ESI calc. ($C_{17}H_{30}N_3$+H$^+$): 278.2591, found: 278.2588.

1-(2-(Dimethylamino)ethyl)-_N_-((5-ethylthiophen-2-yl)methyl)piperidin-4-amine, formic acid salt (22)



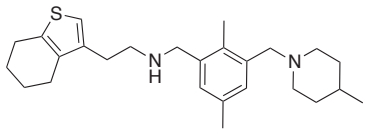
Educt 1: 1-(2-(dimethylamino)ethyl)piperidin-4-amine; Educt 2: 5-ethylthiophene-2-carbaldehyde. Yellow oil; 40%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 1.30 (3H, t, *J* = 4.0 Hz, CH_3), 1.72 (2H, m, CH_2), 2.04 (2H, m, CH_2), 2.29 (2H, m, CH_2), 2.66 (6H, s, CH_3), 2.79-2.87 (5H, m), 2.98-3.09 (4H, m, CH_2), 4.07 (2H, s, CH_2), 6.66 (1H, d, *J* = 3.2 Hz, Ar-H), 6.88 (1H, d, *J* = 3.2 Hz, Ar-H), 8.23 (2H, br.s., NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 15.77, 23.45, 29.02, 30.92, 42.95, 43.53, 51.53, 52.19, 52.85, 54.00, 123.41, 128.33, 133.45, 148.83. HRMS-ESI calc. ($C_{16}H_{28}N_3S$+H$^+$): 296.2155, found: 296.2151.

1-(6-Bromo-1,2,3,4-tetrahydronaphthalen-2-yl)-4-(2-(4-ethylpiperidin-1-yl)ethyl)piperidine, formic acid salt (23)



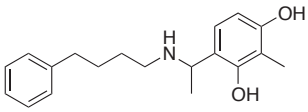
Educt 1: 4-ethyl-1-(2-(piperidin-4-yl)ethyl)piperidine; Educt 2: 6-bromo-3,4-dihydronaphthalen-2(1*H*)-one. Brown oil; 51%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 0.93 (3H, t, *J* = 6.8 Hz, CH_3), 1.33-1.38 (3H, m, CH_2+CH), 1.59-1.92 (12H, m), 2.35 (1H, m, CH), 2.95 (2H, m, CH_2), 2.79-3.15 (8H, m), 3.48-3.55 (5H, m), 6.99 (1H, d, *J* = 8.0 Hz, Ar-H), 7.26-7.28 (2H, m, Ar-H), 8.43 (2H, br.s, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 11.06, 23.93, 28.18, 28.51, 29.09, 29.17, 29.43, 30.90, 32.16, 35.79, 47.96, 48.48, 52.30, 54.12, 61.03, 120.10, 129.31, 130.93, 131.25, 132.09, 137.25, 167.37. HRMS-ESI calc. ($C_{24}H_{26}BrN_2$+H$^+$): 433.2213, found: 433.2210.

_N_-(2,5-Dimethyl-3-((4-methylpiperidin-1-yl)methyl)benzyl)-2-(4,5,6,7-tetrahydrobenzo[_b_]thiophen-3-yl)ethanamine, formic acid salt (24)



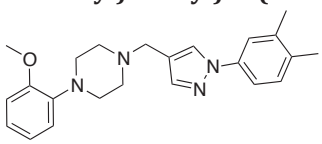
Educt 1: (2,5-dimethyl-3-((4-methylpiperidin-1-yl)methyl)phenyl)methanamine; Educt 2: 2-(4,5,6,7-tetrahydrobenzo[*b*]thiophen-3-yl)acetaldehyde. White oil; 23%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 0.95 (3H, d, *J* = 4.7 Hz, CH_3), 1.39-1.55 (3H, m, CH_2+CH), 1.68 (2H, d, *J* = 9.4 Hz, CH_2), 1.74-1.86 (4H, m, CH_2), 2.24-2.46 (10H, m), 2.69-2.87 (4H, m, CH_2), 2.98-3.08 (2H, m, CH_2), 3.17 (2H, d, *J* = 10.6 Hz, CH_2), 3.79 (2H, s, CH_2), 3.99 (2H, s, CH_2), 6.73 (1H, s, Ar-H), 7.16 (1H, s, Ar-H), 7.19 (1H, s, Ar-H), 8.45 (1H, br.s, NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 14.94, 20.75, 21.27, 22.58, 23.27, 24.26, 25.29, 26.14, 32.08, 47.09, 49.24, 52.74, 58.80, 118.11, 131.67, 132.30, 132.57, 132.93, 133.99, 134.17, 135.76, 136.01, 136.82, 167.53. HRMS-ESI calc. ($C_{26}H_{37}N_2S$+H$^+$): 411.2828, found: 411.2819.

2-Methyl-4-(1-((4-phenylbutyl)amino)ethyl)benzene-1,3-diol, formic acid salt (25)



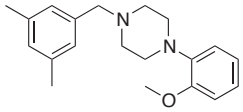
Educt 1: 4-phenylbutan-1-amine; Educt 2: 1-(2,4-dihydroxy-3-methylphenyl)ethanone. White oil; 30%; ^{1}H NMR (CD_3OD, 400.13 MHz): δ 1.63-1.68 (7H, m, CH_2 and CH_3), 2.22 (3H, s, CH_3), 2.61 (2H, m, CH_2), 2.81 (2H, m, CH_2), 4.47 (1H, q, *J* = 6.8 Hz, CH), 6.45 (2H, d, *J* = 8.4 Hz, Ar-H), 6.91 (2H, d, *J* = 8.4 Hz, Ar-H), 7.14-7.18 (3H, m, Ar-H), 7.25 (2H, m, Ar-H), 8.34 (2H, br. s, NH). ^{13}C NMR (CD_3OD, 100.61 MHz): δ 8.90, 18.57, 26.67, 29.32, 36.10, 46.50, 56.08, 108.60, 113.45, 115.55, 126.77, 127.02, 129.43, 129.45, 142.74, 155.21, 158.52, 167.37. HRMS-ESI calc. ($C_{19}H_{25}NO_2$+H$^+$): 300.1958, found: 300.1962.

1-((1-(3,4-Dimethylphenyl)-1*H*-pyrazol-4-yl)methyl)-4-(2-methoxyphenyl) piperazine (26)



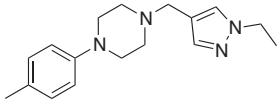
Educt 1: 1-(2-methoxyphenyl)piperazine; Educt 2: 1-(3,4-dimethylphenyl)-1*H*-pyrazole-4-carbaldehyde. Brown powder; mp = 122-124ºC; 61%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.31 (3H, s, CH_3), 2.34 (3H, s, CH_3), 2.74 (4H, br.s, CH_2), 3.14 (4H, br.s, CH_2), 3.60 (2H, s, CH_2), 3.88 (3H, s, CH_3), 6.88 (1H, dd, *J* = 1.2 and 8.0 Hz, Ar-H), 6.92-7.04 (3H, m, Ar-H), 7.11 (1H, d, *J* = 8.0 Hz, Ar-H), 7.39 (1H, dd, *J* = 2.4 and 8.4 Hz, Ar-H), 7.52 (1H, d, *J* = 2.4 Hz, Ar-H), 7.67 (1H, s, Ar-H), 7.88 (1H, s, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 19.27, 19.93, 50.56, 52.60, 53.03, 55.34, 111.14, 116.23, 118.23, 118.92, 120.33, 120.98, 122.92, 126.61, 130.34, 134.79, 137.85, 138.16, 141.29, 141.54, 152.27. HRMS-ESI calc. ($C_{23}H_{27}N_4O$+H$^+$): 377.2336, found: 377.2339.

1-(3,5-Dimethylbenzyl)-4-(2-methoxyphenyl)piperazine, formic acid salt (27)



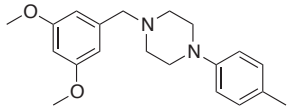
Educt 1: 1-(2-methoxyphenyl)piperazine; Educt 2: 2-(3,5-dimethylphenyl)acetaldehyde. Orange oil; 78%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.33 (6H, s, CH_3), 3.07 (4H, br.s, CH_2), 3.25 (4H, br.s, CH_2), 3.85 (3H, s, CH_3), 3.93 (2H, s, CH_2), 6.86 (1H, d, *J* = 7.6 Hz, Ar-H), 6.91-6.93 (2H, m, Ar-H), 7.01-7.04 (3H, m, Ar-H) 8.48 (1H, s, NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 21.23, 48.56, 51.70, 55.37, 61.09, 111.20, 118.53, 121.10, 123.73, 128.51, 130.51, 131.54, 138.41, 139.97, 152.09, 166.94. HRMS-ESI calc. ($C_{20}H_{25}N_2O$+H$^+$): 311.2118, found: 311.2113.

1-((1-Ethyl-1*H*-pyrazol-4-yl)methyl)-4-(*p*-tolyl)piperazine, formic acid salt (28)



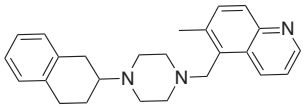
Educt 1: 1-(*p*-tolyl)piperazine; Educt 2: 1-ethyl-1*H*-pyrazole-4-carbaldehyde. Brown oil; 80%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 1.48 (3H, t, *J* = 8.0 Hz, CH_3), 2.26 (3H, s, CH_3), 3.00 (4H, m, CH_2), 3.28 (4H, m, CH_2), 3.86 (2H, s, CH_2), 4.17, (2H, q, *J* = 8.0 Hz, CH_2), 6.81 (2H, d, *J* = 8.0 Hz, Ar-H), 7.07 (2H, d, *J* = 8.0 Hz, Ar-H), 7.47 (1H, s, Ar-H), 7.54 (1H, s, Ar-H), 8.41 (1H, br.s, NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 15.41, 20.44, 47.18, 48.11, 50.97, 51.02, 111.52, 117.01, 129.79, 129.91, 130.41, 140.35, 148.14, 166.86. HRMS-ESI calc. ($C_{17}H_{23}N_4$+H$^+$): 285.2074, found: 285.2071.

1-(3,5-Dimethoxybenzyl)-4-(*p*-tolyl)piperazine (29)



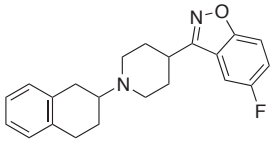
Educt 1: 1-(*p*-tolyl)piperazine; Educt 2: 3,5-dimethoxybenzaldehyde. White powder; mp = 113-115ºC; 50%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 2.29 (3H, s, CH_3), 2.64 (4H, br.s., CH_2), 3.18 (4H, br.s., CH_2), 3.53 (2H, s, CH_2), 3.82 (6H, s, OCH_3), 6.40 (1H, s, Ar-H), 6.57 (2H, s, Ar-H), 6.87 (2H, d, *J* = 7.6 Hz, Ar-H), 7.09 (2H, d, *J* = 7.6 Hz, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 20.42, 49.74, 53.18, 55.34, 63.14, 99.05, 106.93, 116.38, 129.10, 129.60, 140.70, 149.34, 160.75. HRMS-ESI calc. ($C_{20}H_{25}N_2O_2$+H$^+$): 327.2067, found: 327.2062.

7-Methyl-8-((4-(1,2,3,4-tetrahydronaphthalen-2-yl)piperazin-1-yl)methyl) quinolone, formic acid salt (30)



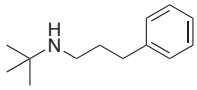
Educt 1: 6-methyl-5-(piperazin-1-ylmethyl)quinoline; Educt 2: 3,4-dihydronaphthalen-2(1*H*)-one. Brown oil; 77%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 1.83 (1H, m, CH), 2.36 (1H, m, CH), 2.61 (3H, s, CH_3), 2.89-3.16 (12H, m), 3.48 (1H, m, CH), 4.05 (2H, s, CH_2), 7.08-7.16 (4H, m, Ar-H), 7.48 (1H, dd, *J* = 4.0 and 8.8 Hz, Ar-H), 7.50 (1H, d, *J* = 8.4 Hz, Ar-H), 8.03 (1H, d, *J* = 8.4 Hz, Ar-H), 8.41 (2H, br.s., NH), 8.62 (1H, d, *J* = 8.4 Hz, Ar-H), 8.91 (1H, dd, *J* = 2.0 and 4.4 Hz, Ar-H). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 20.40, 24.42, 28.69, 29.82, 48.27, 50.24, 54.28, 61.11, 120.90, 126.20, 126.50, 128.28, 128.51, 128.71, 129.32, 130.17, 132.87, 133.07, 133.52, 135.08, 136.51, 146.95, 148.85, 166.75. HRMS-ESI calc. ($C_{25}H_{28}N_3$+H^+): 372.2434, found: 372.2430.

5-Fluoro-3-(1-(1,2,3,4-tetrahydronaphthalen-2-yl)piperidin-4-yl)benzo[*d*] isoxazole, formic acid salt (31)



Educt 1: 5-fluoro-3-(piperidin-4-yl)benzo[*d*]isoxazole; Educt 2: 3,4-dihydronaphthalen-2(1*H*)-one. Brown powder; mp = 120-122ºC; 31%; ^{1}H NMR ($CDCl_3$, 400.13 MHz): δ 1.88 (1H, m, CH), 2.40-2.48 (5H, m, CH_2+CH), 2.98-3.19 (6H, m, CH_2), 3.55-3.60 (4H, m, CH_2), 7.12-7.18 (4H, m, Ar-H), 7.35 (1H, td, *J* = 2.4 and 9.2 Hz, Ar-H), 7.47 (1H, d, *J* = 6.4 Hz, Ar-H), 7.56 (1H, dd, *J* = 3.7 and 8.8 Hz, Ar-H), 8.49 (1H, br.s, NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 24.30, 26.95, 28.71, 29.44, 46.96, 47.27, 61.76, 106.29, 111.13, 118.86, 119.13, 126.36, 126.68, 128.59, 129.38, 132.70, 134.96, 157.89, 159.78, 160.30, 166.98. HRMS-ESI calc. ($C_{22}H_{22}FN_2O$+H^+): 351.1867, found: 351.1864.

***N*-(*Tert*-butyl)-3-phenylpropan-1-amine, formic acid salt (32)**



Educt 1: 2-methylpropan-2-amine; Educt 2: 3-phenylpropanal. White amorphous solid; 90%; ^{1}H NMR ($CDCl_3$, 400.13 MHz) δ 1.32 (9H, s, CH_3), 2.05 (2H, m, CH_2), 2.61 (2H, m, CH_2), 2.82 (2H, m, CH_2), 7.16-7.20 (3H, m, Ar-H), 7.26 (2H, m, Ar-H), 8.52 (1H, s, NH). ^{13}C NMR ($CDCl_3$, 100.61 MHz): δ 25.75, 27.73, 32.87, 40.84, 55.85, 126.18, 128.24, 128.49, 140.23, 167.46. HRMS-ESI calc. ($C_{13}H_{21}N$+H^+): 192.1747, found: 192.1753.

8.4.2 Supplementary data

Building block SMARTS filter rules

Table S3: Modified ZINC's[2] basic SMARTS expression for building block triage.

Maximal occurrence	SMARTS pattern	Description
20	[a,A]	Non-hydrogens atoms
10	[#7,#8,#16]	N, O, S
6	[Cl,Br]	Cl, Br
3	F	Fluorines
2	[C+,Cl+,S+]	Quaternary C,Cl,S
2	C(=O)C[N+,n+]	Beta cabonyl quaternary
2	[N;R0][N;R0]C(=O)	Acylhydrazides
2	C1[O,S,N]C1	(Thio)epoxides, aziridines
2	cC[N+]	Benzylic Quaternary
2	C[O,S;R0][C;R0](=S)	Thioesters
2	[#7]O[#6,#16]=O	Aminooxy(oxo)
2	N(~[OD1])~[OD1]	Nitros
2	C=[N;R0]*	Imines
2	N#CC=C	Acrylonitriles
2	C=CC(=O)[!#7;!#8]	Propenals
1	S(=O)(=O)[Cl,Br]	Sulfonyl Halides
1	[S,C](=[O,S])[F,Br,Cl,I]	Acid Halides
1	[Br,Cl][CX4;CH,CH2]	Alkyl Halides
1	SC#N	Thiocyanates
1	COS(=O)O[C,c]	Sulfate esters
1	COS(=O)(=O)[C,c]	Sulfonates
1	[ND4+]	Quaternary N
0	[CD1][CD2][CD2][CD2][CD2][CD2][CD2]	Heptanes
0	OCl(O)(O)(O)	Perchlorates
0	O=CN=[N+]=[N-]	Carbazides
0	S[Cl,Br,F]	S-Halides
0	N=C=N	Carbodiimides
0	N#CC[OH]	Cyanohydrines
0	C(=O)Oc1c(F)c(F)c(F)c(F)c1(F)	Pentafluorophenyl esters

Table continues on next page …

[2] Irwin JJ, Shoichet BK (2005) ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177-182.

… continued from previous page

Maximal occurrence	SMARTS pattern	Description
0	C(=O)Oc1ccc(N(=O)=O)cc1	Paranitrophenyl esters
0	C(=O)Onnn	HOBt esters
0	OS(=O)(=O)C(F)(F)F	Triflates
0	[Cl]C([C&R0])=N	Chloramidines
0	C(=O)N(C(=O))OC(=O)	Triacyloximes
0	N#CC(=O)	Acyl Cyanides
0	S(=O)(=O)C#N	Sulfonyl Cyanides
0	[N;R0]=[N;R0]C#N	Azocyanamides
0	[N;R0]=[N;R0]CC=O	Azoalkanals
0	B([O;H0])([O;H0])	Boronic acid
0	B(F)(F)(F)	Trifluoroborane
0	S(O)(O)(O)	Sulfanetriol
0	C(=O)OC(=O)	Acid anhydrides
0	OO	Peroxides
0	[*]-C=CC(=O)	α,β-unsaturated carbonyl
0	[N+]#[C-]	Isonitriles
0	N=N=N	Azides
0	C(=O)[C;H2]C(=O)	1,3-dicarbonyl
0	[#6][CX3](=O)[#7]	Amides
0	C(=O)C(=O)	Oxalaldehyde
0	C(=O)[O;H1]	Carboxylic acids
0	C(=O)[Cl,Br,F]	Acyl Halides
0	N=C=[S,O]	Iso(thio)cyanates
0	N(=O)O	Nitro
0	[#6][NX3;H1]N	Hydrazine

Building blocks were restricted to a maximal molecular weight of 250 containing only elements H, C, N, O, F, S, Cl, F and Br. Further the following SMARTS pattern were applied and building blocks were filtered according the maximal allowed occurrence.

Retrospective evaluation of the machine-learning models

Table S4. Comparison of the retrospective performance of GP, GP_{LCB}, BKD and RF for the focused target panel. Values are given as *mean* ± *sd*.

Target	#actives / #inactives	Method	BEDROC	Recall 3%	Q^2	MAE[a]
Sigma opioid	1636 / 800	GP	0.71 ± 0.03	0.84 ± 0.03	0.84 ± 0.02	0.53 ± 0.03
		GP_{LCB}	0.77 ± 0.02	0.89 ± 0.02	0.84 ± 0.02	0.75 ± 0.05
		BKD	0.59 ± 0.03	0.73 ± 0.03	0.61 ± 0.04	-
		RF	0.74 ± 0.02	0.85 ± 0.03	0.84 ± 0.03	0.59 ± 0.04
Delta opioid	2745 / 1572	GP	0.90 ± 0.01	0.93 ± 0.01	0.88 ± 0.01	0.51 ± 0.02
		GP_{LCB}	0.95 ± 0.01	0.97 ± 0.01	0.88 ± 0.02	0.76 ± 0.04
		BKD	0.84 ± 0.02	0.88 ± 0.02	0.54 ± 0.03	-
		RF	0.89 ± 0.01	0.93 ± 0.02	0.87 ± 0.01	0.57 ± 0.03
Kappa opioid	3390 / 1327	GP	0.81 ± 0.02	0.85 ± 0.02	0.84 ± 0.01	0.58 ± 0.02
		GP_{LCB}	0.90 ± 0.01	0.93 ± 0.01	0.83 ± 0.01	0.88 ± 0.05
		BKD	0.82 ± 0.02	0.85 ± 0.02	0.51 ± 0.02	-
		RF	0.85 ± 0.02	0.87 ± 0.02	0.83 ± 0.01	0.63 ± 0.02
Mu opioid	3266 / 1260	GP	0.83 ± 0.02	0.87 ± 0.03	0.84 ± 0.02	0.57 ± 0.03
		GP_{LCB}	0.92 ± 0.01	0.95 ± 0.01	0.84 ± 0.02	0.85 ± 0.03
		BKD	0.82 ± 0.02	0.86 ± 0.02	0.56 ± 0.02	-
		RF	0.83 ± 0.01	0.86 ± 0.02	0.83 ± 0.02	0.62 ± 0.03
Histamine H3	2836 / 58	GP	0.73 ± 0.02	0.77 ± 0.02	0.73 ± 0.03	0.44 ± 0.01
		GP_{LCB}	0.84 ± 0.01	0.87 ± 0.01	0.73 ± 0.03	0.54 ± 0.02
		BKD	0.69 ± 0.03	0.73 ± 0.03	0.26 ± 0.03	-
		RF	0.70 ± 0.02	0.70 ± 0.02	0.70 ± 0.03	0.48 ± 0.02
Dopamine D_4	1658 / 953	GP	0.79 ± 0.02	0.92 ± 0.04	0.83 ± 0.03	0.60 ± 0.04
		GP_{LCB}	0.85 ± 0.02	0.95 ± 0.02	0.83 ± 0.03	0.99 ± 0.05
		BKD	0.67 ± 0.03	0.79 ± 0.02	0.59 ± 0.03	-
		RF	0.78 ± 0.02	0.90 ± 0.03	0.83 ± 0.03	0.64 ± 0.04
Dopamine D_1	581 / 922	GP	0.86 ± 0.03	0.90 ± 0.03	0.86 ± 0.02	0.49 ± 0.04
		GP_{LCB}	0.89 ± 0.03	0.92 ± 0.03	0.86 ± 0.03	0.72 ± 0.05
		BKD	0.76 ± 0.04	0.81 ± 0.05	0.63 ± 0.04	-
		RF	0.89 ± 0.02	0.92 ± 0.02	0.86 ± 0.03	0.53 ± 0.03
Dopamine D_2	4183 / 1389	GP	0.77 ± 0.02	0.84 ± 0.02	0.80 ± 0.02	0.50 ± 0.02
		GP_{LCB}	0.87 ± 0.01	0.92 ± 0.01	0.80 ± 0.02	0.69 ± 0.02
		BKD	0.64 ± 0.02	0.72 ± 0.02	0.35 ± 0.03	-
		RF	0.79 ± 0.02	0.82 ± 0.03	0.78 ± 0.02	0.55 ± 0.02
Dopamine D_3	2708 / 932	GP	0.75 ± 0.02	0.79 ± 0.03	0.83 ± 0.02	0.53 ± 0.03
		GP_{LCB}	0.83 ± 0.02	0.87 ± 0.02	0.83 ± 0.02	0.74 ± 0.03
		BKD	0.67 ± 0.04	0.76 ± 0.03	0.46 ± 0.04	-
		RF	0.78 ± 0.02	0.80 ± 0.02	0.81 ± 0.03	0.58 ± 0.04
Dopamine D_5	148 / 87	GP	0.68 ± 0.12	0.71 ± 0.13	0.72 ± 0.15	0.55 ± 0.13
		GP_{LCB}	0.84 ± 0.08	0.87 ± 0.07	0.71 ± 0.15	0.85 ± 0.12
		BKD	0.79 ± 0.12	0.86 ± 0.12	0.58 ± 0.12	-
		RF	0.68 ± 0.11	0.72 ± 0.12	0.71 ± 0.12	0.61 ± 0.11
5-HT_{1a}	2640 / 173	GP	0.63 ± 0.03	0.67 ± 0.03	0.69 ± 0.04	0.54 ± 0.04
		GP_{LCB}	0.76 ± 0.02	0.80 ± 0.03	0.68 ± 0.04	0.73 ± 0.03
		BKD	0.66 ± 0.02	0.74 ± 0.03	0.28 ± 0.03	-
		RF	0.66 ± 0.02	0.66 ± 0.02	0.68 ± 0.04	0.57 ± 0.04

[a] MAE is not applicable to BKD due to different value ranges

Table S5: Y-Scrambling performance evaluation. Values are given as *mean* ± *sd*.

Target	#actives / #inactives	Method	BEDROC	Recall 3%	Q^2	MAE[a]
Sigma-1	1636 / 800	GP	0.09 ± 0.03	0.12 ± 0.03	0.00 ± 0.00	1.65 ± 0.04
		GP_{LCB}	0.19 ± 0.02	0.23 ± 0.03	0.00 ± 0.00	3.46 ± 0.13
		BKD	0.16 ± 0.07	0.19 ± 0.08	0.00 ± 0.00	-
		RF	0.19 ± 0.04	0.23 ± 0.04	0.01 ± 0.01	1.69 ± 0.04
Delta opioid	2745 / 1572	GP	0.09 ± 0.04	0.11 ± 0.05	0.00 ± 0.00	1.74 ± 0.04
		GP_{LCB}	0.20 ± 0.08	0.22 ± 0.08	0.00 ± 0.00	4.23 ± 0.12
		BKD	0.18 ± 0.08	0.20 ± 0.08	0.00 ± 0.00	-
		RF	0.23 ± 0.05	0.26 ± 0.05	0.00 ± 0.00	1.81 ± 0.03
Kappa opioid	3390 / 1327	GP	0.11 ± 0.03	0.13 ± 0.04	0.00 ± 0.00	1.66 ± 0.05
		GP_{LCB}	0.23 ± 0.04	0.26 ± 0.05	0.00 ± 0.00	4.03 ± 0.08
		BKD	0.19 ± 0.05	0.20 ± 0.06	0.00 ± 0.01	-
		RF	0.26 ± 0.02	0.27 ± 0.02	0.00 ± 0.00	1.73 ± 0.06
Mu opioid	3266 / 1260	GP	0.09 ± 0.04	0.11 ± 0.05	0.00 ± 0.00	1.64 ± 0.04
		GP_{LCB}	0.21 ± 0.06	0.25 ± 0.06	0.00 ± 0.00	3.96 ± 0.08
		BKD	0.22 ± 0.07	0.23 ± 0.07	0.00 ± 0.00	-
		RF	0.25 ± 0.03	0.28 ± 0.02	0.00 ± 0.00	1.70 ± 0.05
Histamine H3	2836 / 58	GP	0.20 ± 0.08	0.24 ± 0.09	0.01 ± 0.00	0.96 ± 0.05
		GP_{LCB}	0.43 ± 0.14	0.49 ± 0.16	0.01 ± 0.00	1.51 ± 0.04
		BKD	0.30 ± 0.10	0.33 ± 0.10	0.00 ± 0.01	-
		RF	0.32 ± 0.05	0.34 ± 0.04	0.01 ± 0.01	0.99 ± 0.05
Dopamine D_4	1658 / 953	GP	0.09 ± 0.04	0.13 ± 0.05	0.00 ± 0.00	1.82 ± 0.06
		GP_{LCB}	0.19 ± 0.05	0.24 ± 0.06	0.00 ± 0.00	4.42 ± 0.17
		BKD	0.14 ± 0.06	0.16 ± 0.07	0.00 ± 0.00	-
		RF	0.20 ± 0.04	0.24 ± 0.05	0.00 ± 0.00	1.86 ± 0.05
Dopamine D_1	581 / 922	GP	0.07 ± 0.04	0.11 ± 0.06	0.01 ± 0.01	1.64 ± 0.04
		GP_{LCB}	0.16 ± 0.06	0.21 ± 0.07	0.01 ± 0.01	3.60 ± 0.12
		BKD	0.17 ± 0.06	0.19 ± 0.07	0.01 ± 0.02	-
		RF	0.08 ± 0.03	0.09 ± 0.04	0.01 ± 0.02	1.65 ± 0.07
Dopamine D_2	4183 / 1389	GP	0.09 ± 0.04	0.11 ± 0.04	0.00 ± 0.01	1.24 ± 0.04
		GP_{LCB}	0.17 ± 0.07	0.18 ± 0.07	0.00 ± 0.01	2.75 ± 0.08
		BKD	0.15 ± 0.06	0.16 ± 0.06	0.00 ± 0.00	-
		RF	0.27 ± 0.03	0.29 ± 0.03	0.00 ± 0.01	1.29 ± 0.04
Dopamine D_3	2708 / 932	GP	0.10 ± 0.04	0.12 ± 0.05	0.00 ± 0.00	1.49 ± 0.05
		GP_{LCB}	0.21 ± 0.05	0.26 ± 0.06	0.00 ± 0.00	3.15 ± 0.11
		BKD	0.21 ± 0.12	0.22 ± 0.12	0.00 ± 0.00	-
		RF	0.22 ± 0.04	0.24 ± 0.03	0.00 ± 0.01	1.53 ± 0.07
Dopamine D_5	148 / 87	GP	0.25 ± 0.15	0.29 ± 0.18	0.02 ± 0.03	1.29 ± 0.13
		GP_{LCB}	0.59 ± 0.19	0.65 ± 0.21	0.02 ± 0.04	2.26 ± 0.36
		BKD	0.30 ± 0.18	0.35 ± 0.20	0.05 ± 0.08	-
		RF	0.16 ± 0.14	0.17 ± 0.14	0.02 ± 0.03	1.30 ± 0.12
5-HT_{1a}	2640 / 173	GP	0.16 ± 0.06	0.19 ± 0.07	0.00 ± 0.01	1.08 ± 0.03
		GP_{LCB}	0.34 ± 0.12	0.39 ± 0.13	0.00 ± 0.01	1.88 ± 0.07
		BKD	0.22 ± 0.07	0.24 ± 0.07	0.01 ± 0.01	-
		RF	0.29 ± 0.07	0.30 ± 0.07	0.00 ± 0.00	1.11 ± 0.03

[a] MAE is not applicable to BKD due to different value ranges

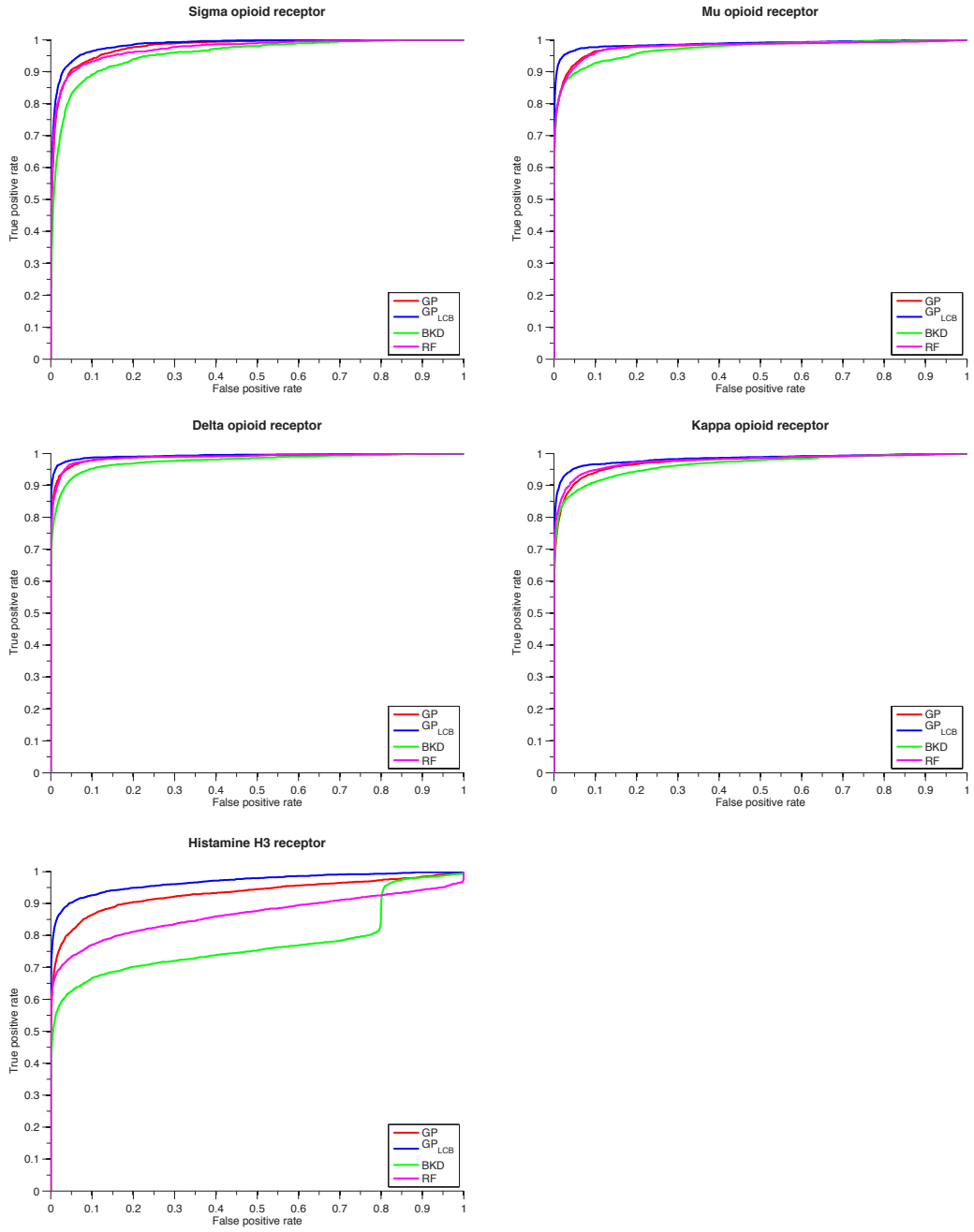


Figure S10. Receiver operating characteristic (ROC) curves for the sigma-1 panel.

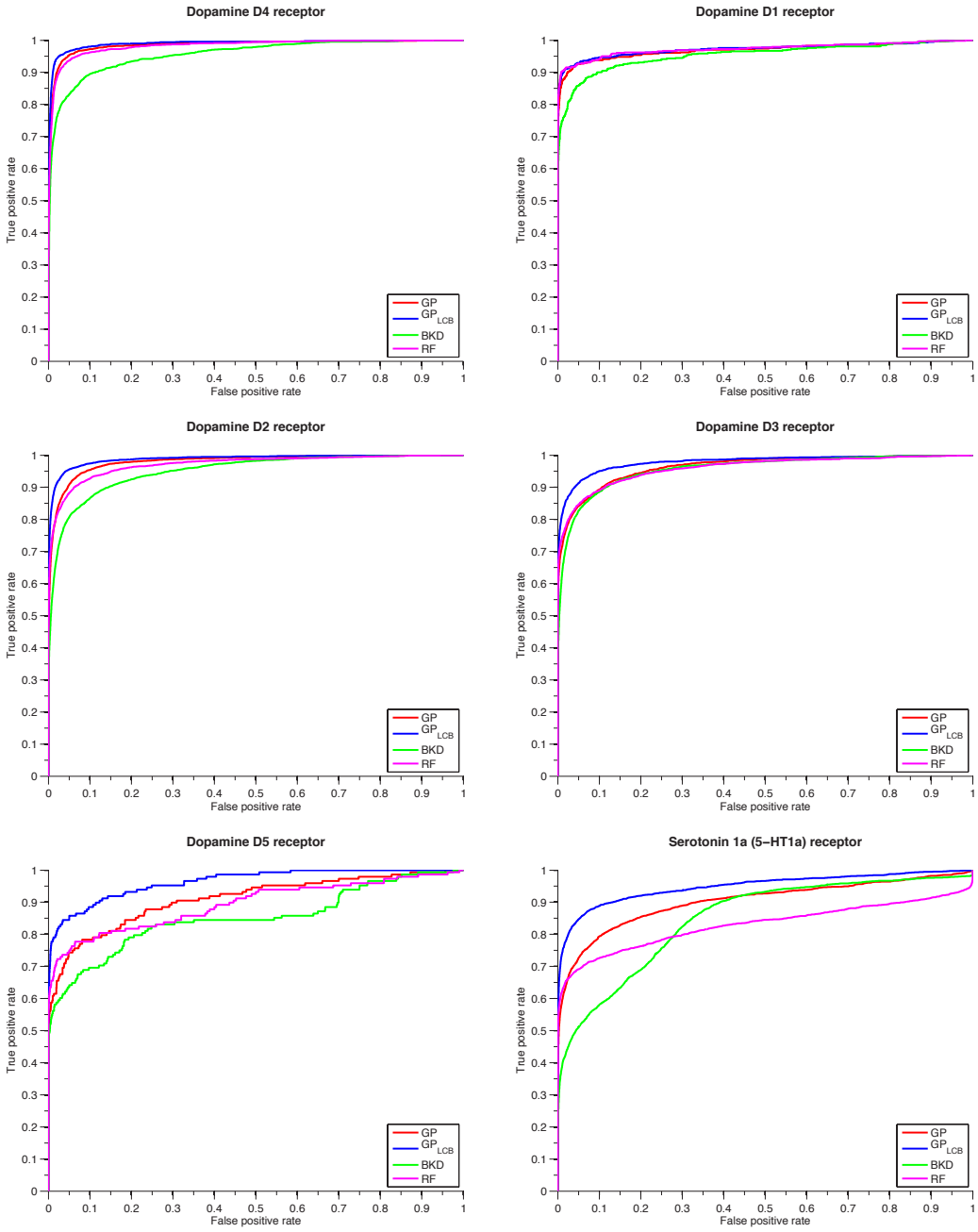


Figure S11. Receiver operating characteristic (ROC) curves for the dopamine D_4 panel.

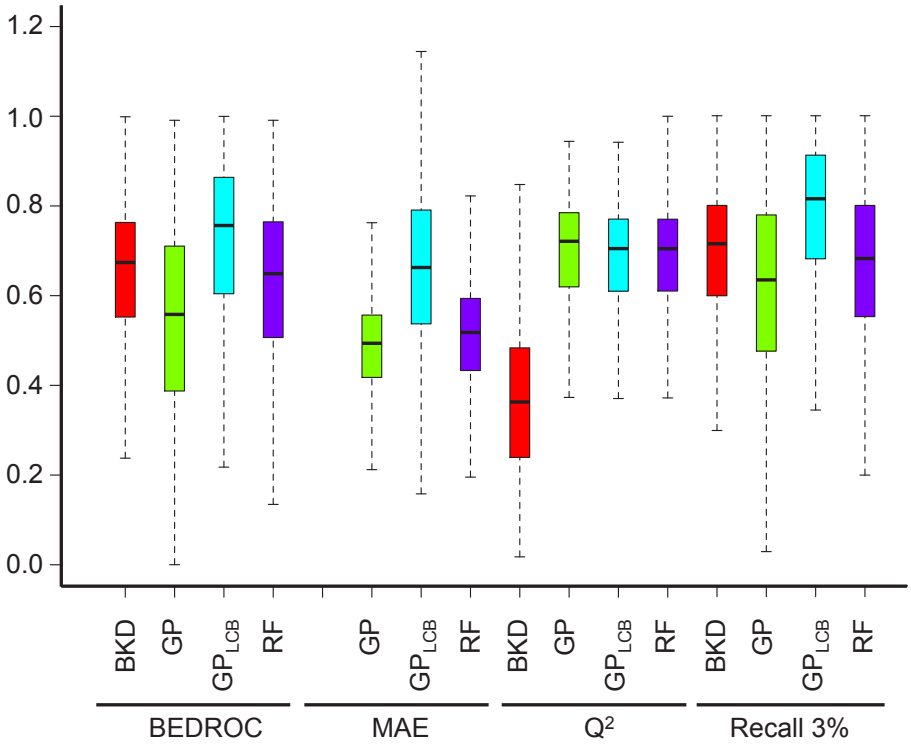


Figure S12. Retrospective, 10-fold cross-validated performance analysis of the machine learning models on 640 human drug targets. Performance is expressed by four different metrics, measuring early enrichment (BEDORC / Recall 3%) and regression performance (Q^2 / MAE).

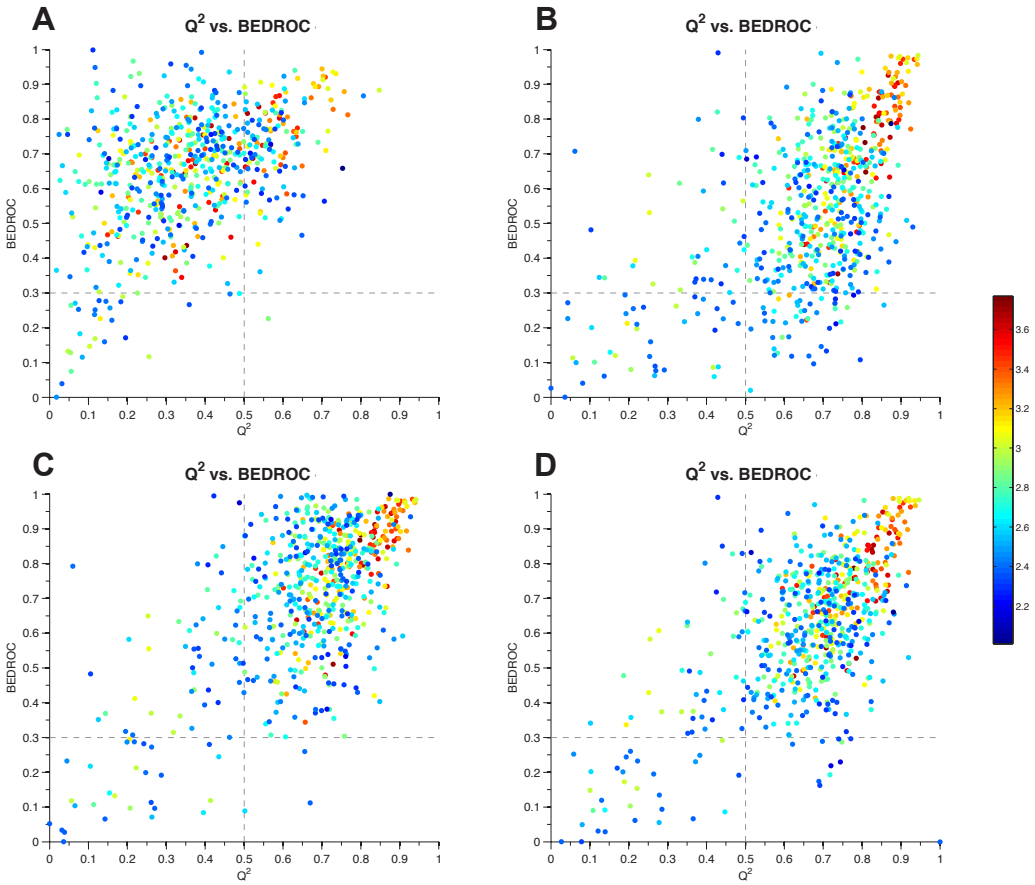


Figure S13. Pairwise comparison of Q^2 and BEDROC performance. BKD (**A**), GP (**B**), GP_{LCB} (**C**) and RF (**D**). Points are colored according to log scaled training size.

Dose-response curves

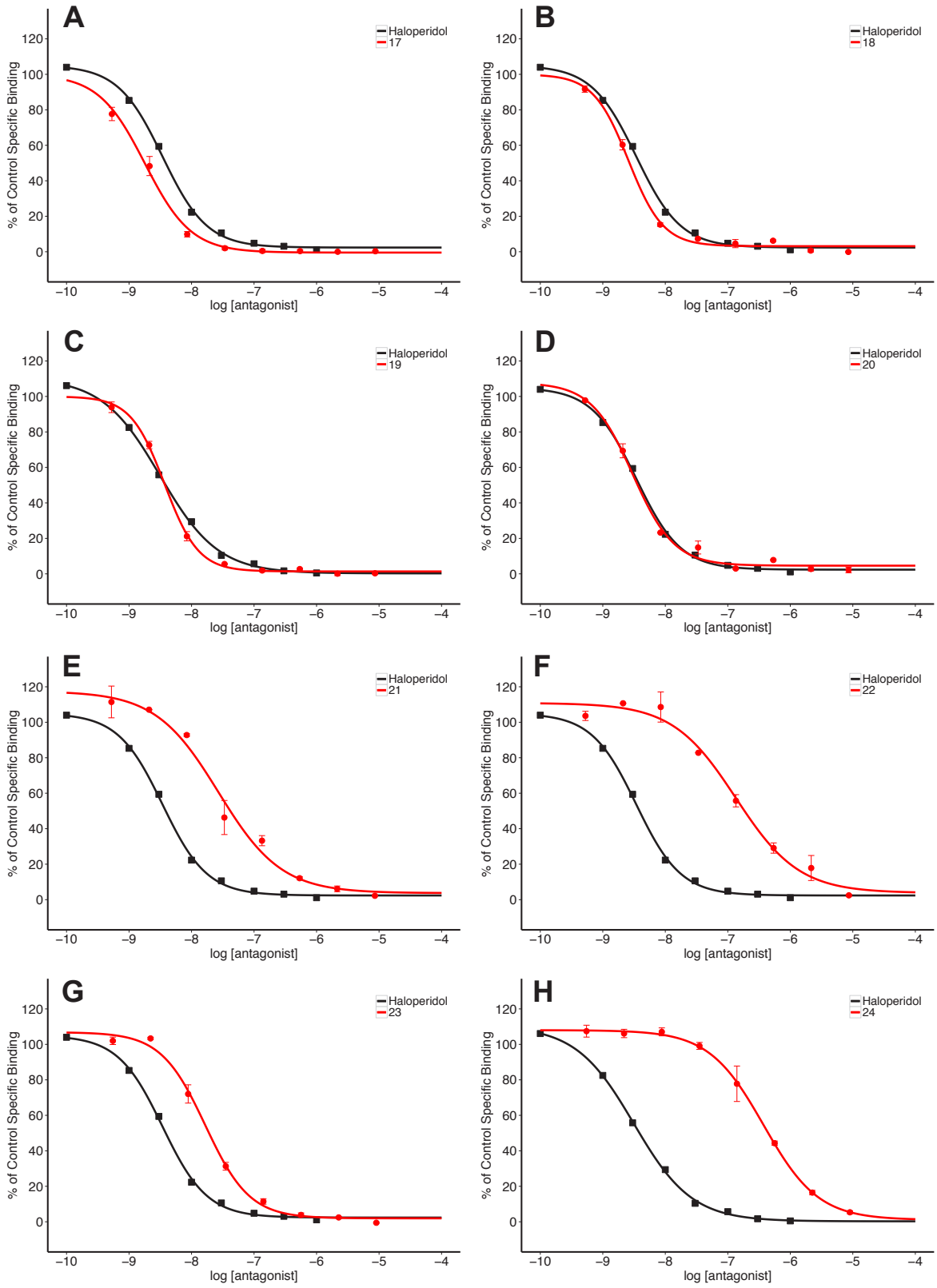


Figure S14. Sigma-1 (agonist radioligand) dose-response curves for **17** (**A**), **18** (**B**), **19** (**C**), **20** (**D**), **21** (**E**), **22** (**F**), **23** (**G**), **24** (**H**). Reference haloperidol as black line (K_i = 1.6 nM).

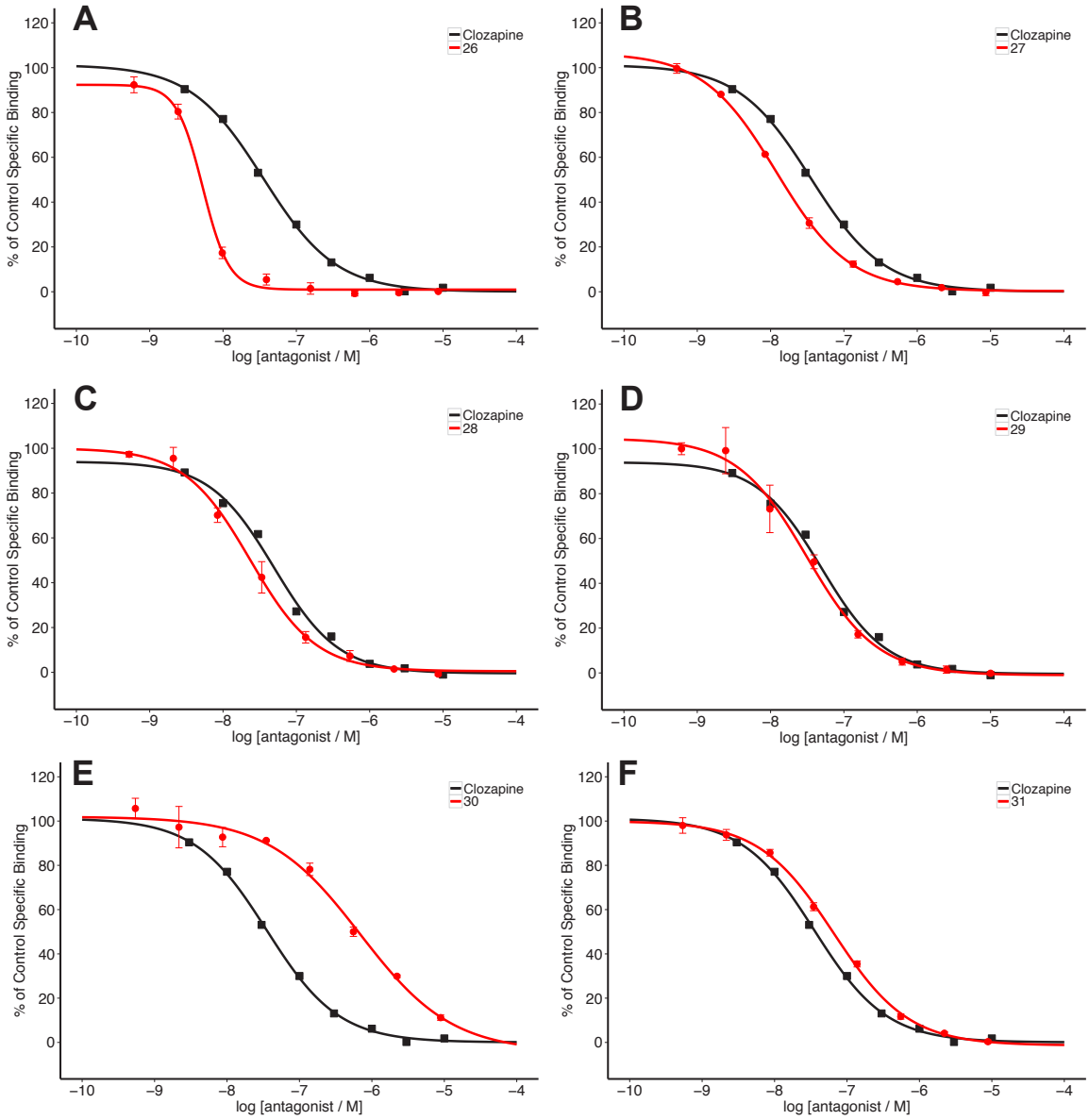
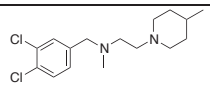
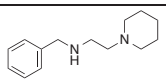
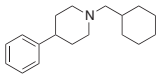
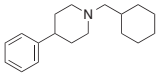
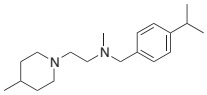
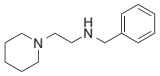
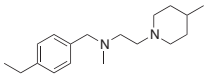
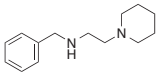
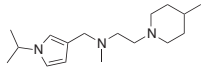
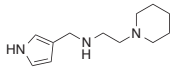
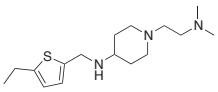
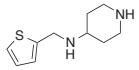
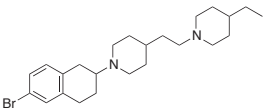
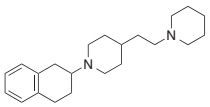
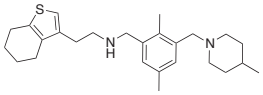
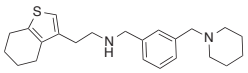
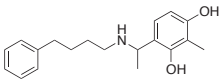
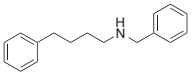
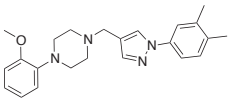
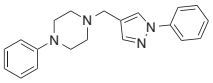
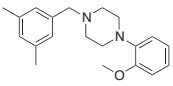
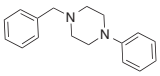
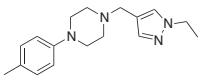
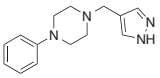
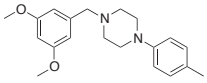
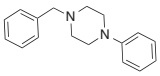
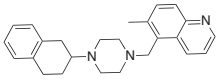
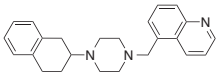
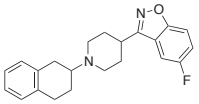
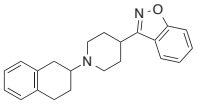
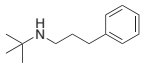


Figure S15. Dopamine D4.4 (antagonist radioligand) dose-response curves for **26** (**A**), **27** (**B**), **28** (**C**), **29** (**D**), **30** (**E**), **31** (**F**). Reference clozapine as black line (K_i = 19 nM).

Table S6: Scaffold frequency analysis of designed MAntA compounds in the respective target training data. For **17-25** Sigma-1 (ChEMBL TID 11272) and for **26-32** Dopamine D_4 (ChEMBL TID 90). Murcko scaffolds were calculated using RDKit.

	MAntA Design		Training data	
ID	**Chemical structure**	**Murcko scaffold**	**Scaffold frequency**	**ChEMBL IDs**
17			0	-
18			0	-
19			0	-
20			0	-
21			0	-
22			0	-
23			0	-
24			0	-
25			1	CHEMBL19628
26			7	CHEMBL208018 CHEMBL210405 CHEMBL210717 CHEMBL210955 + 3 more
27			4	CHEMBL100952 CHEMBL101032 CHEMBL103900 CHEMBL88365
28			0	-
29			4	CHEMBL100952 CHEMBL101032 CHEMBL103900 CHEMBL88365
30			0	-
31			0	-
32			95	CHEMBL1009 CHEMBL108545 CHEMBL1089 CHEMBL112 + 91 more

Chemical space coverage and individual target activity landscapes

Table S7. Projection rank error calculations for sigma-1 and dopamine D_4 selectivity panels. All indices were calculated for K = 10 nearest neighbors.

	Sigma-1	Dopamine D_4
Number of data points	10,658	8,346
Trustworthiness W_T	0.90	0.88
Continuity W_C	0.98	0.98
MRRE W_n	0.09	0.11
MRRE W_v	0.01	0.01
LCMC	0.37	0.35

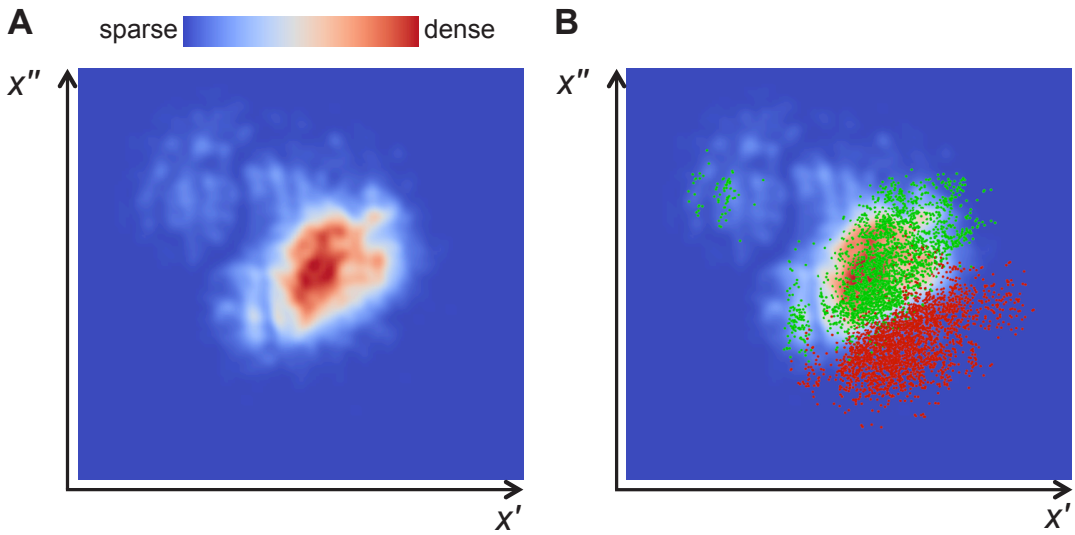


Figure S16. Drug-like chemical space coverage 2D-landscapes. Distribution of 10,000 random ChEMBL molecules (**A**) and with additionally highlighted 5,000 compounds randomly generated using two combinatorial reactions (Red = Ugi-3 component reaction, Green = reductive amination) (**B**).

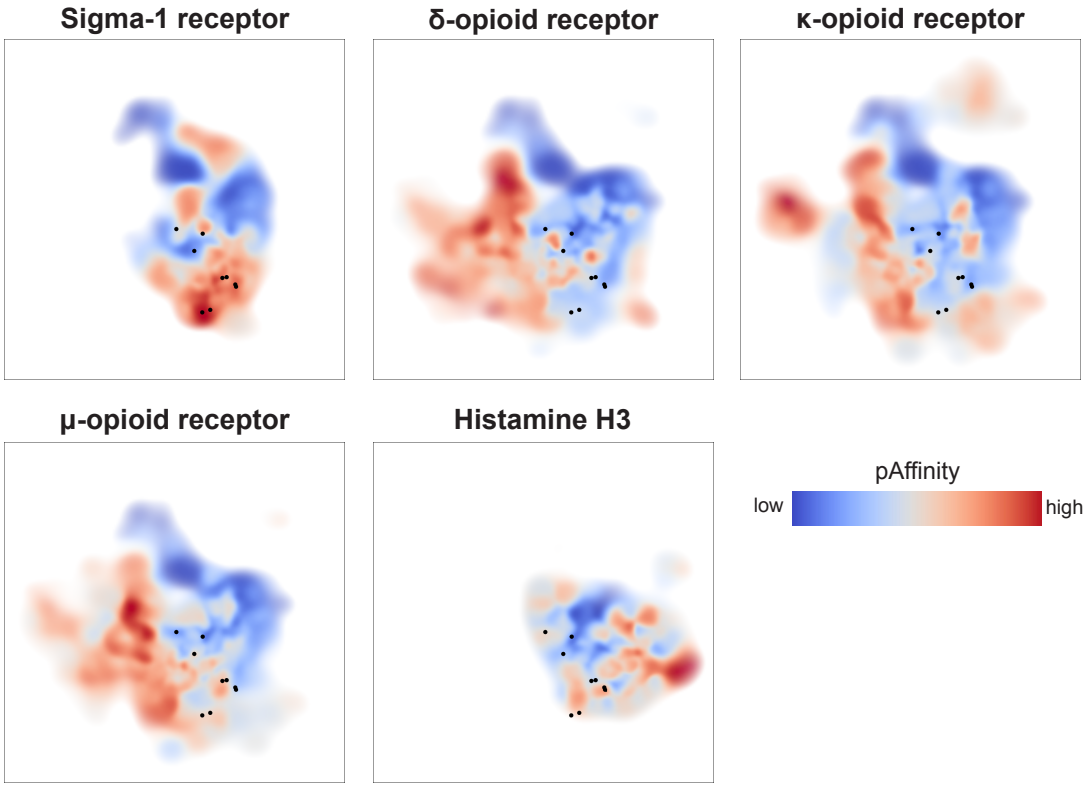


Figure S17. Activity landscapes of the sigma-1 selectivity panel. Coloring of the landscape according to fitted p*Affinity* surface values. Transparency encodes local data density.

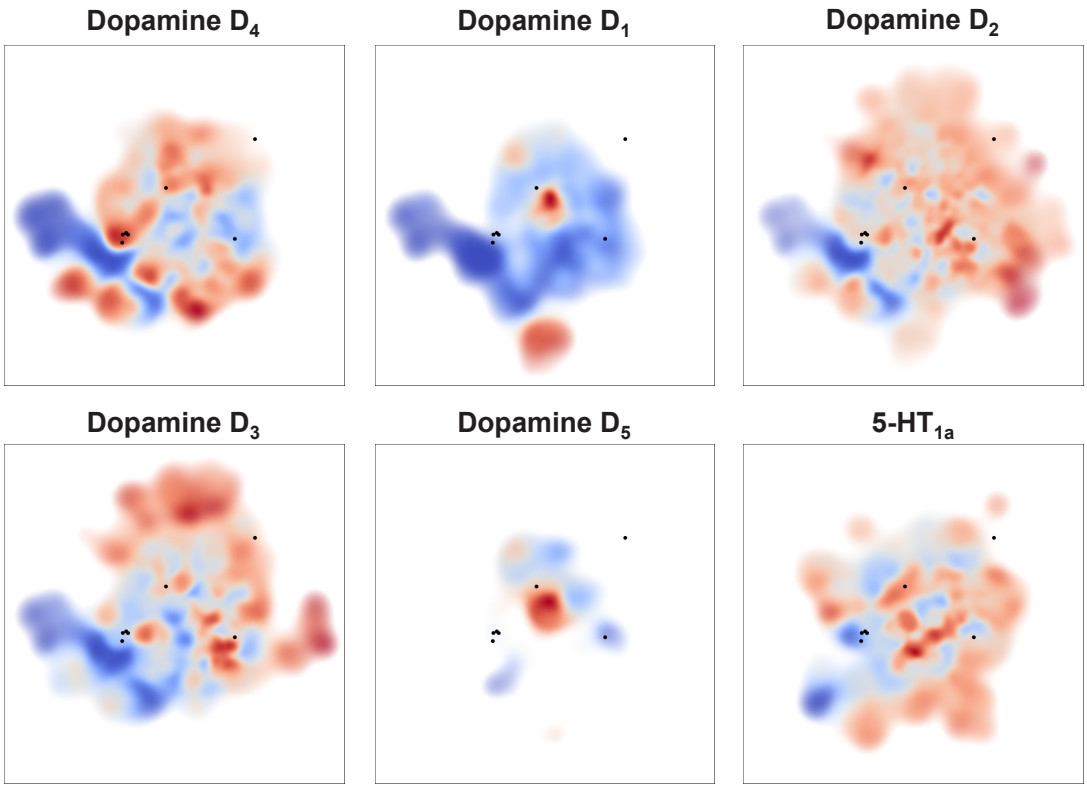


Figure S18. Activity landscapes of the dopamine D_4 selectivity panel. Coloring according to Fig. S17.

Spectral data

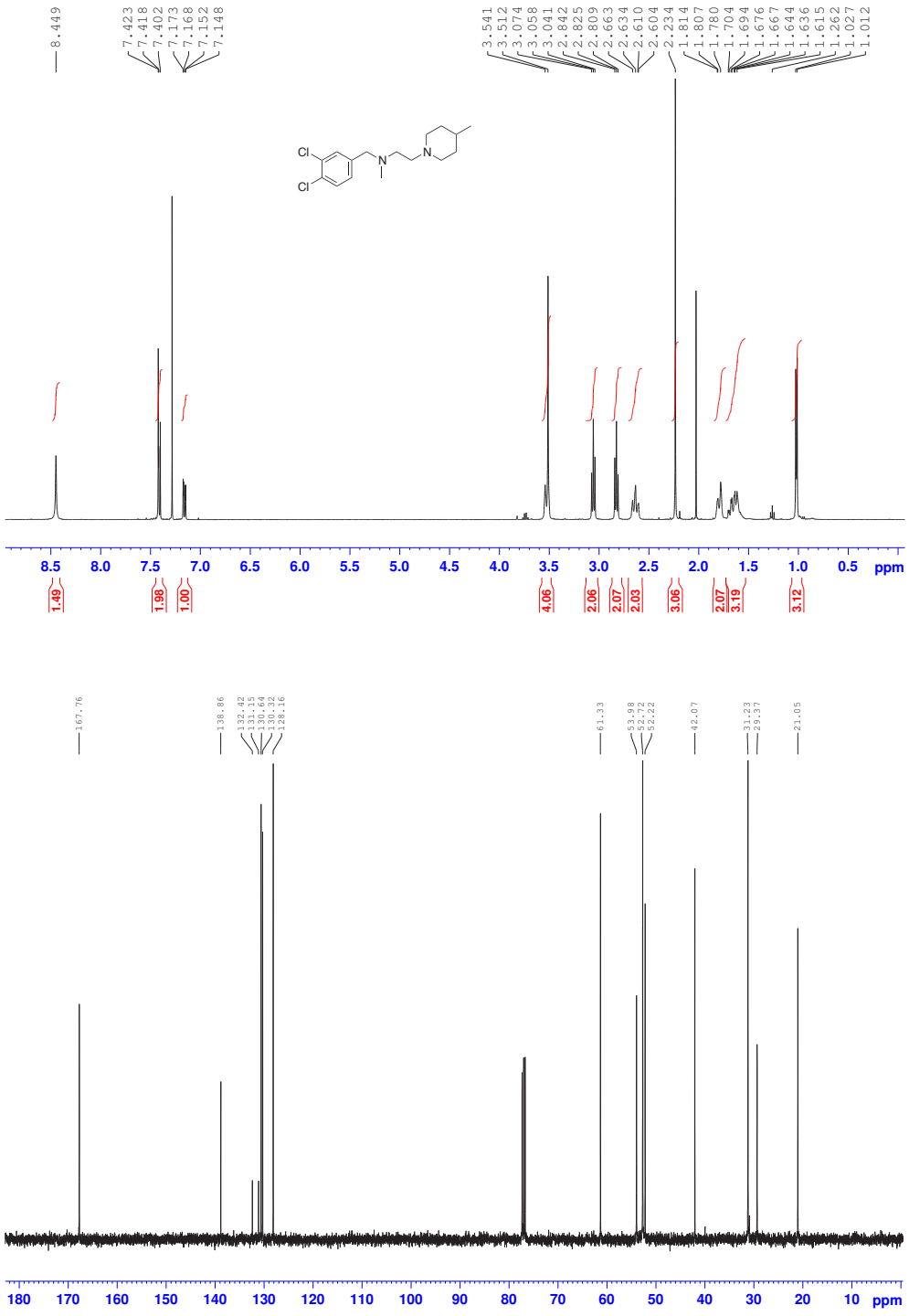


Figure S19. ^{1}H and ^{13}C NMR spectra of **17**.

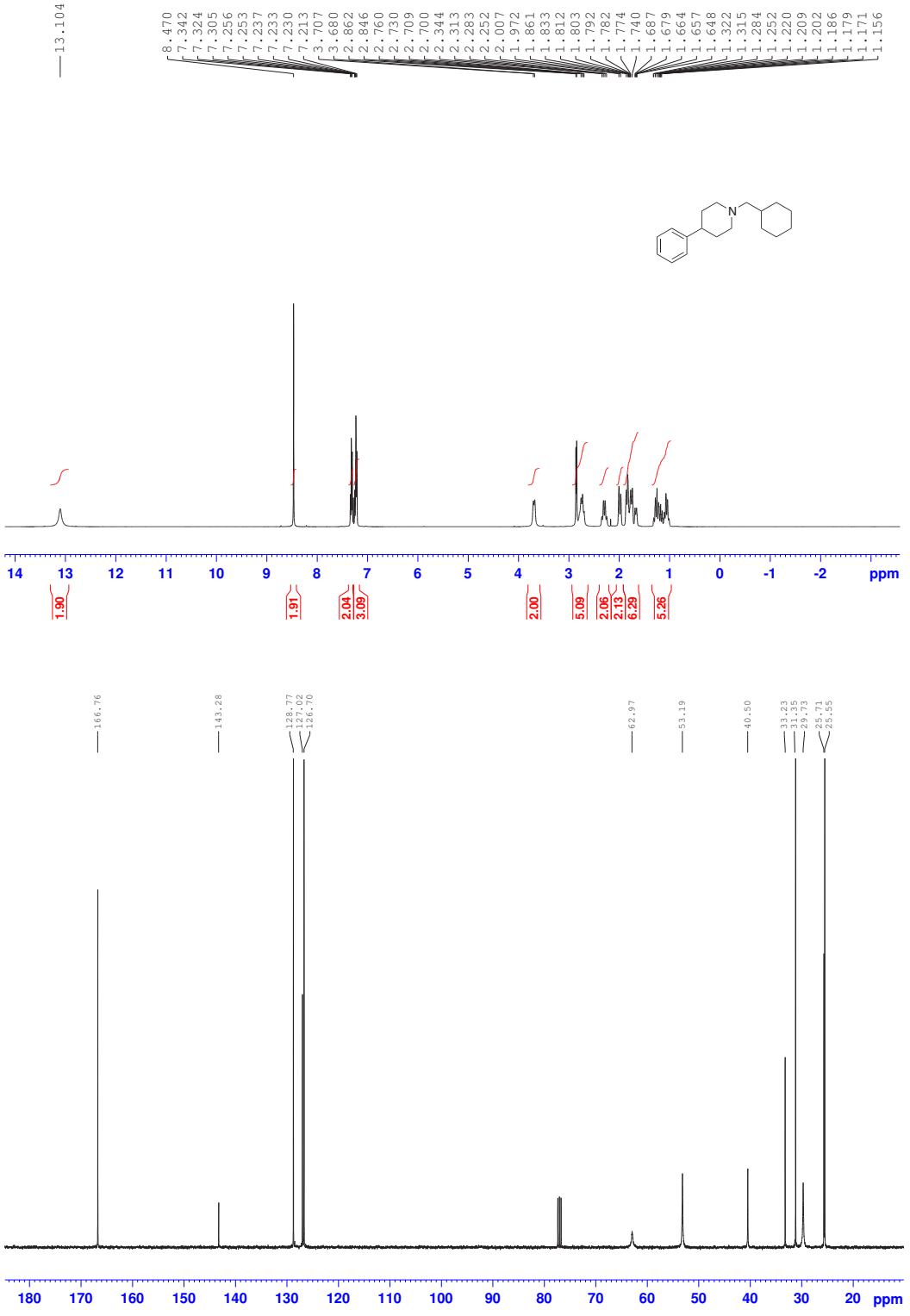


Figure S20. ^{1}H and ^{13}C NMR spectra of **18**.

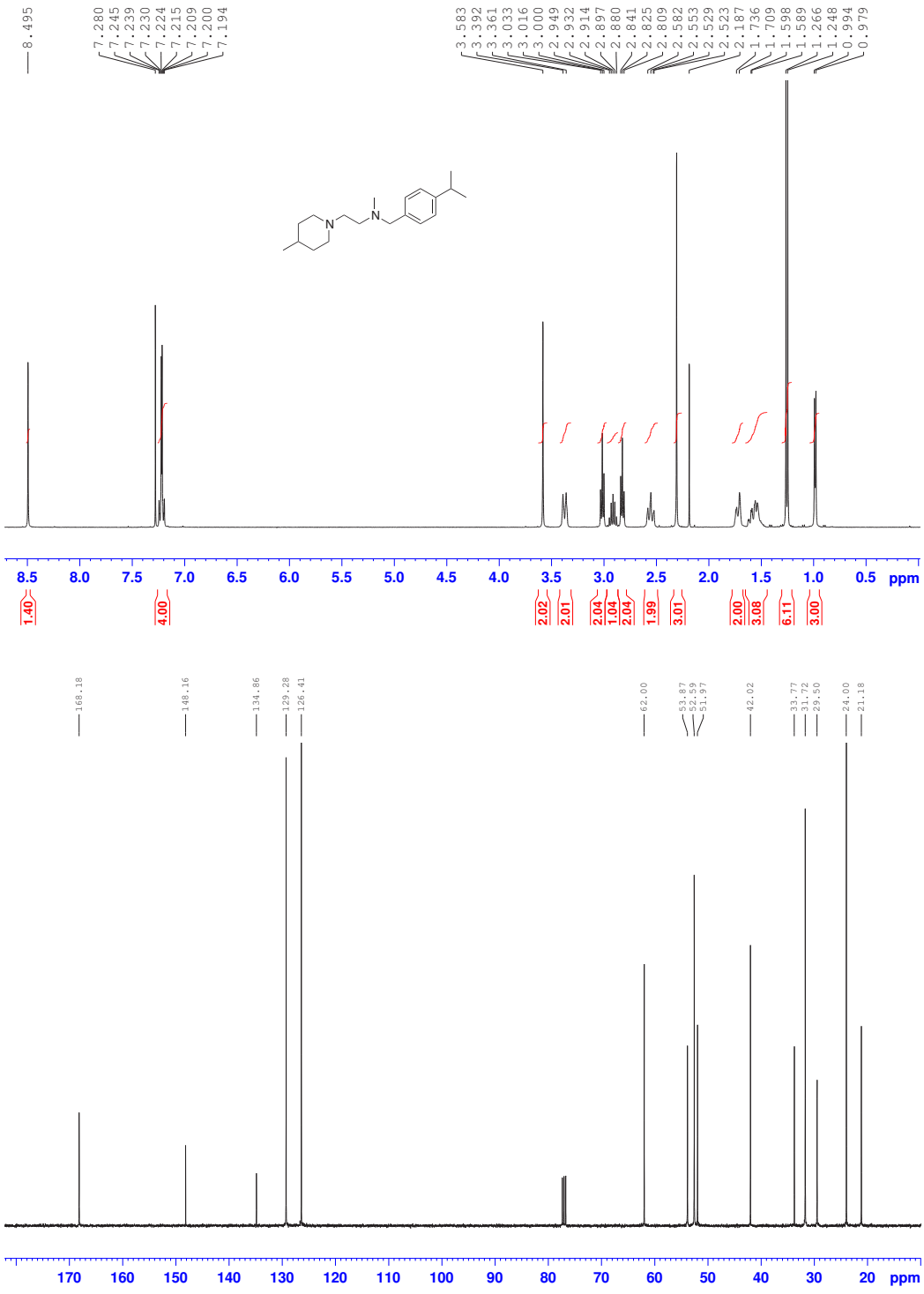


Figure S21. ^{1}H and ^{13}C NMR spectra of **19**.

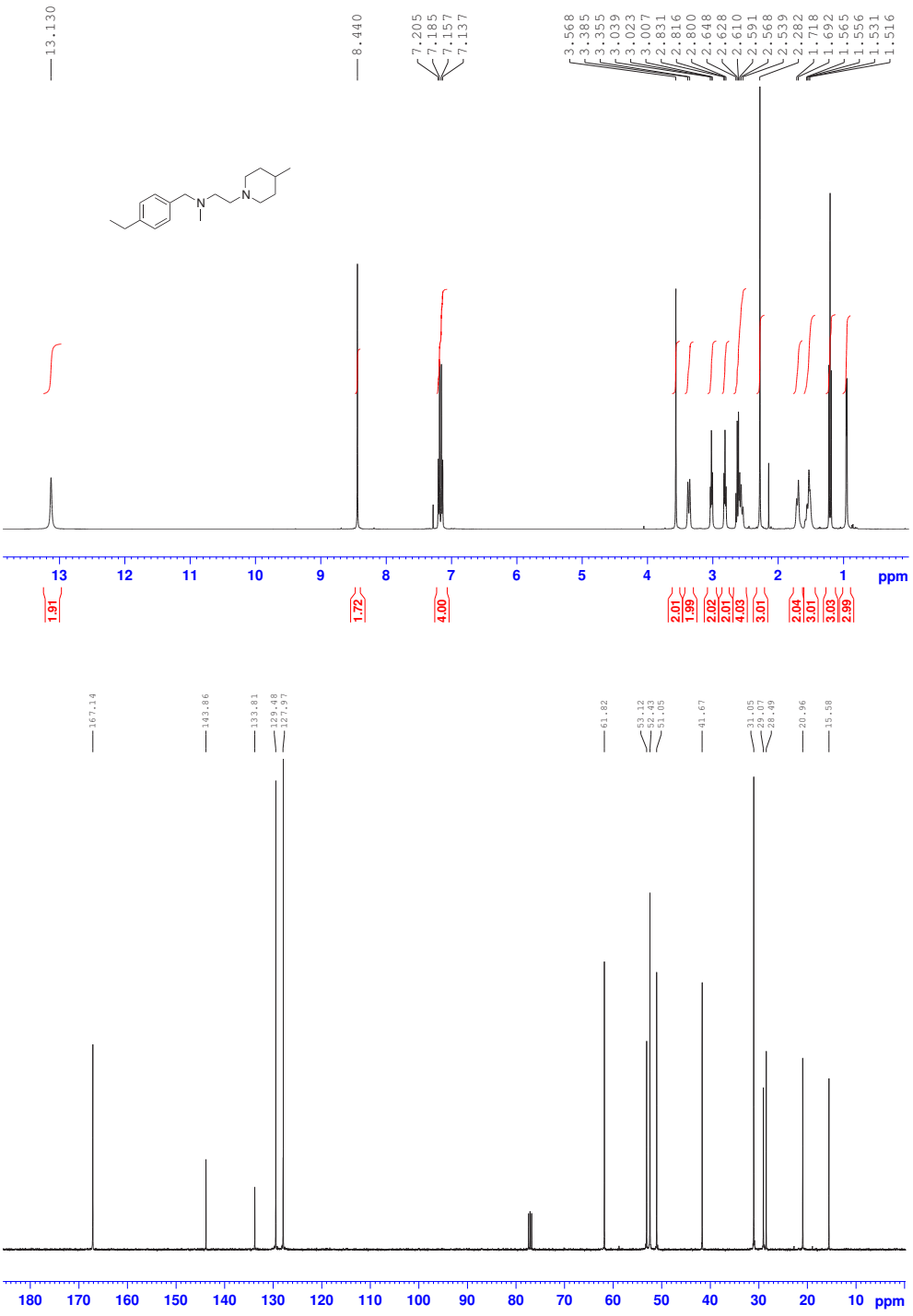


Figure S22. ^{1}H and ^{13}C NMR spectra of **20**.

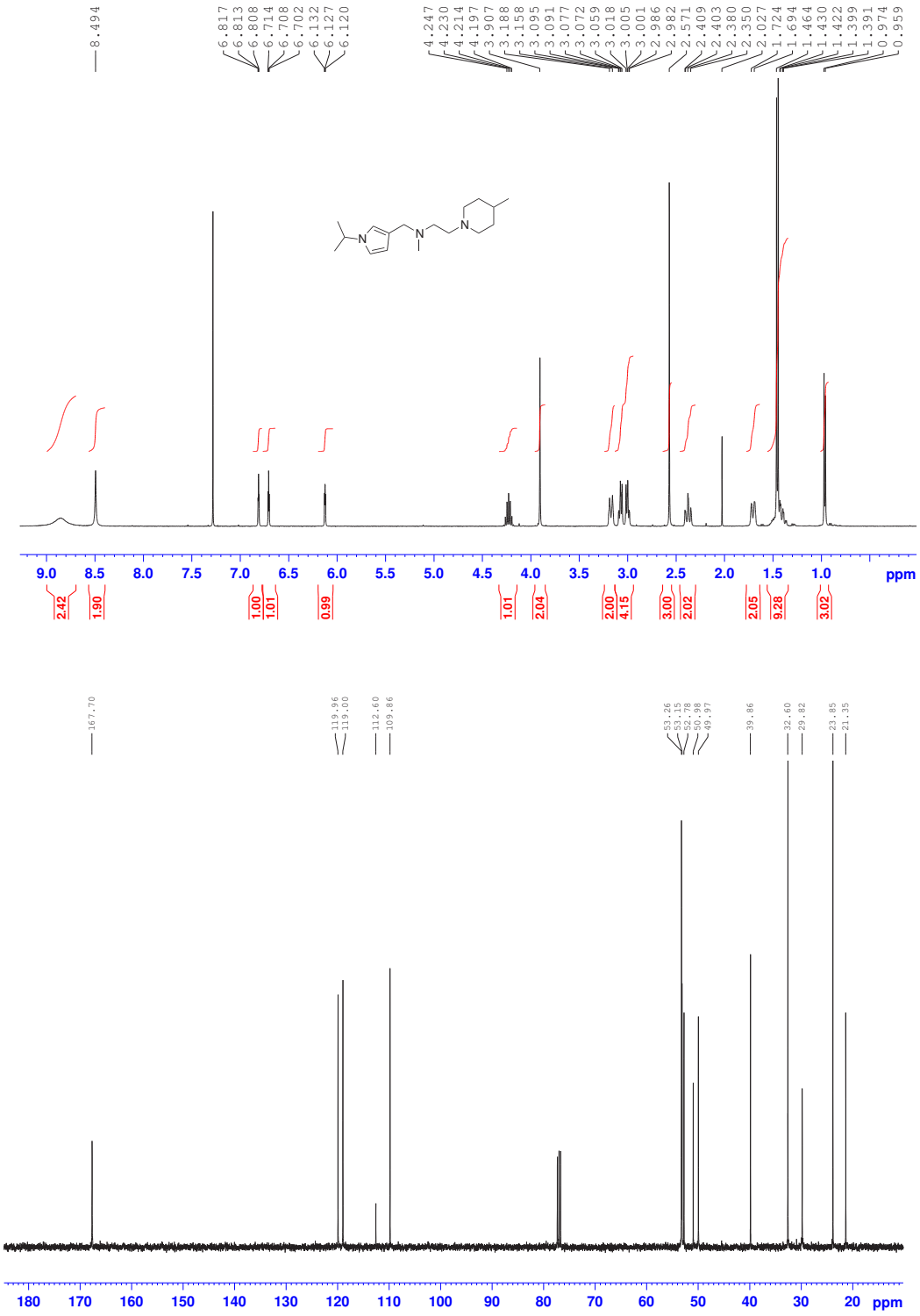


Figure S23. ^{1}H and ^{13}C NMR spectra of **21**.

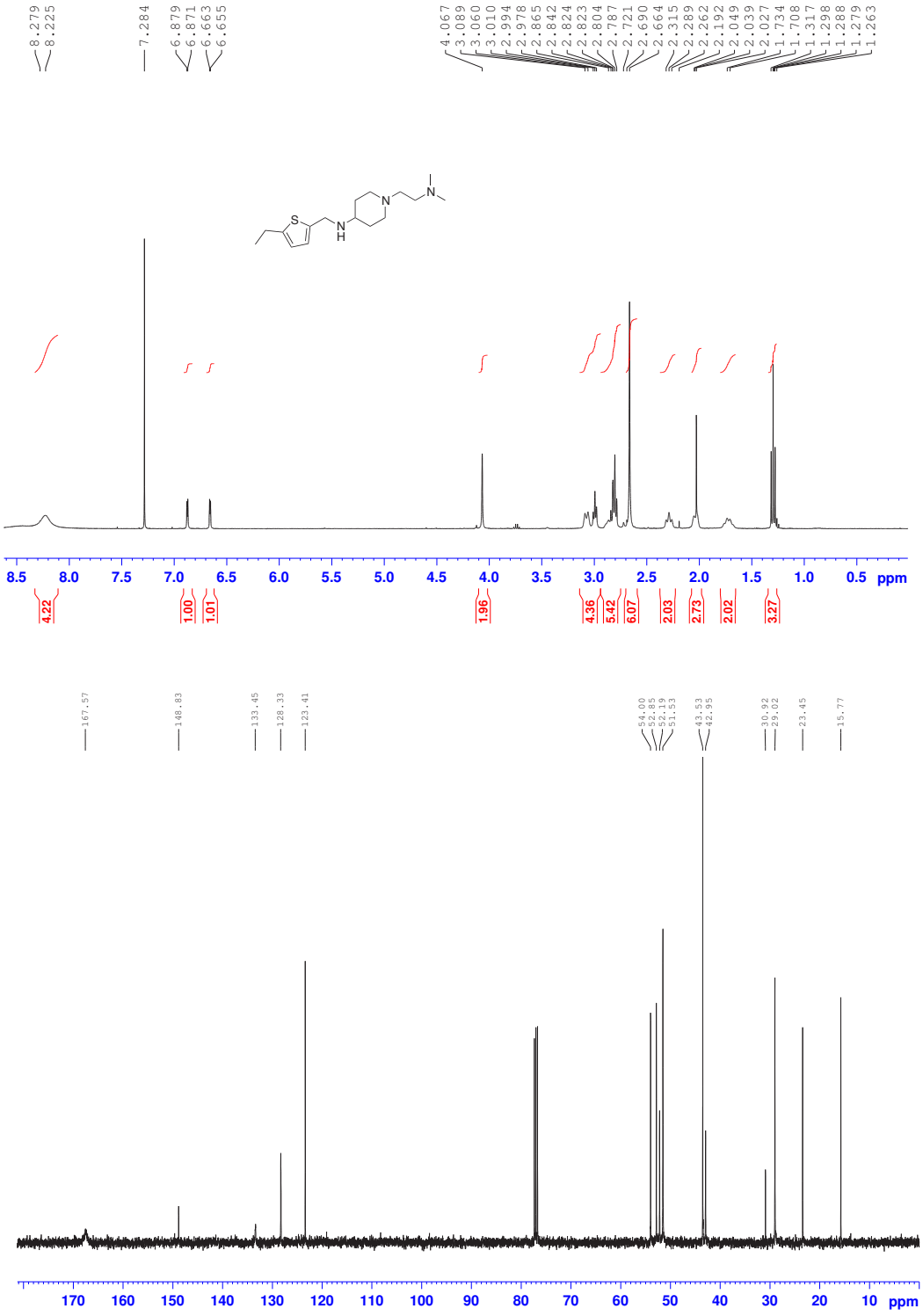


Figure S24. ^{1}H and ^{13}C NMR spectra of **22**.

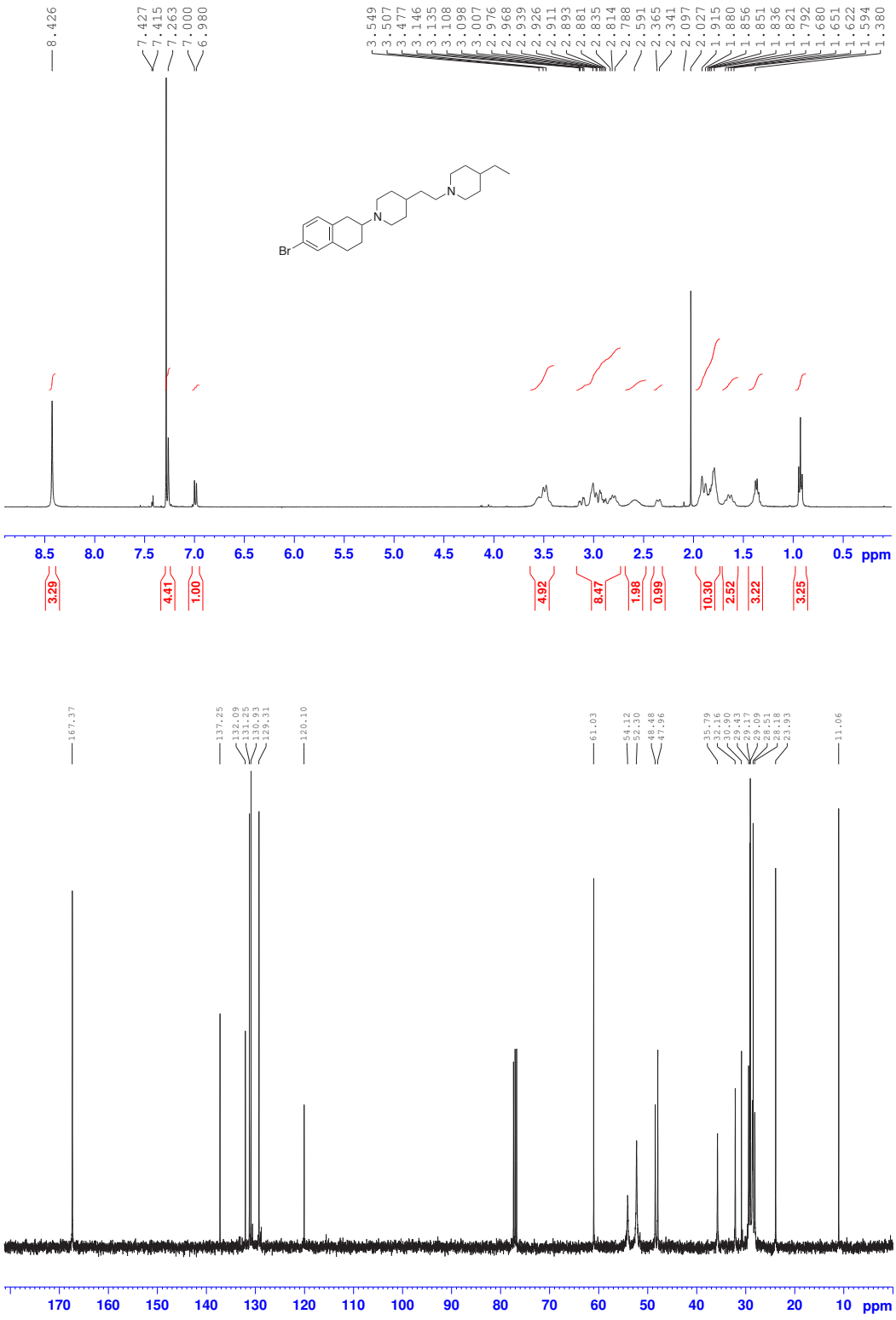


Figure S25. ^{1}H and ^{13}C NMR spectra of **23**.

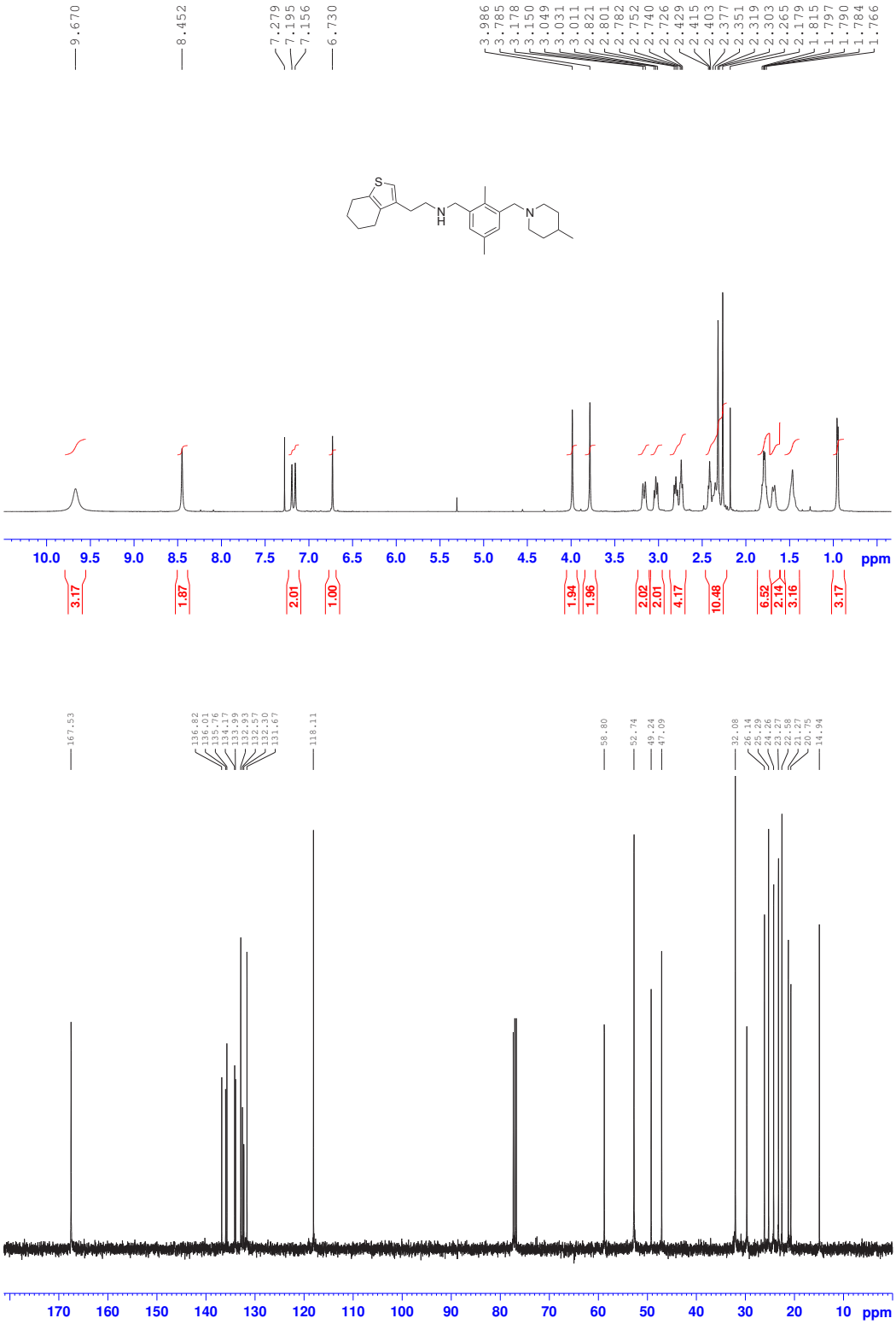


Figure S26. ^{1}H and ^{13}C NMR spectra of **24**.

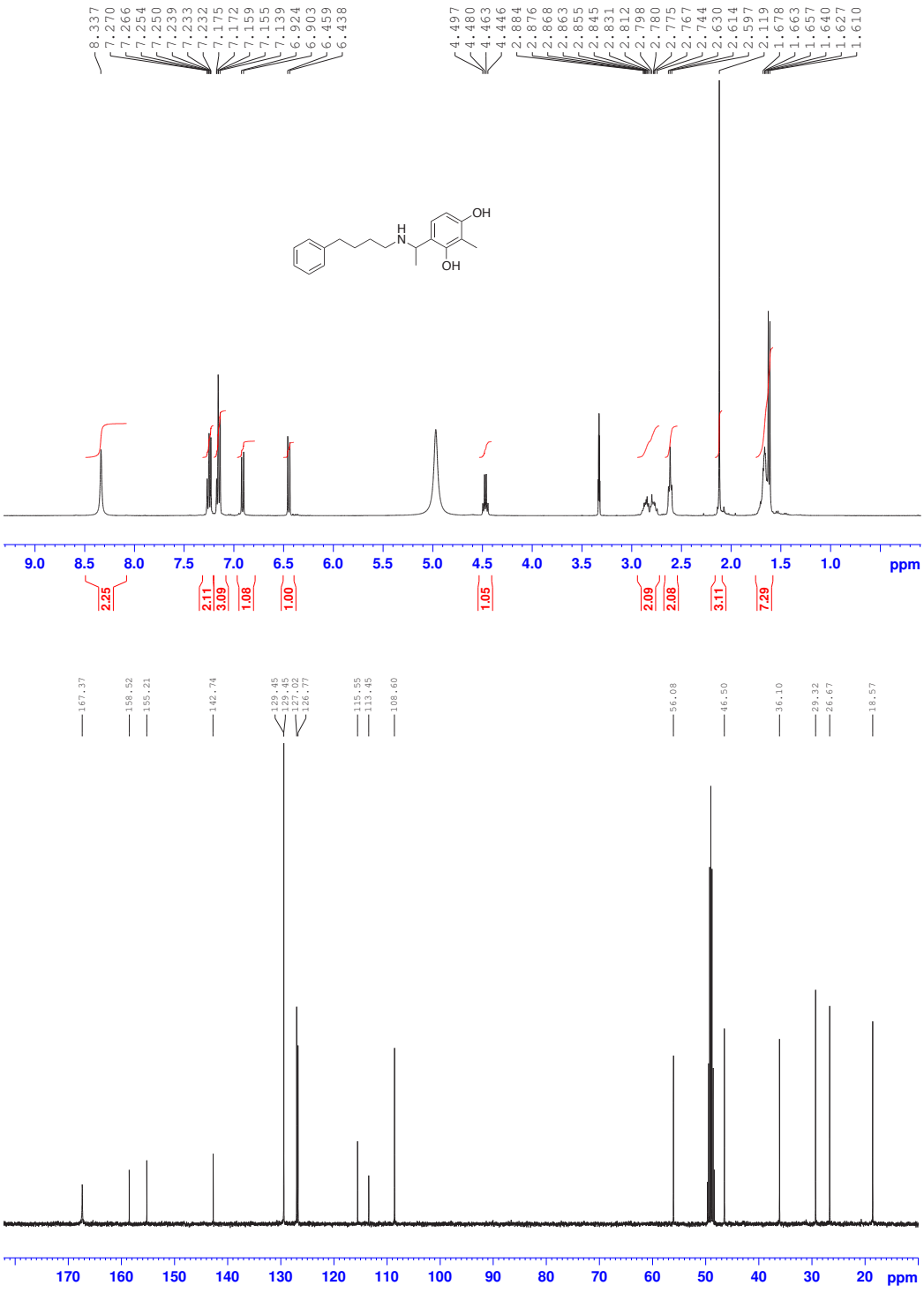


Figure S27. ^{1}H and ^{13}C NMR spectra of **25**.

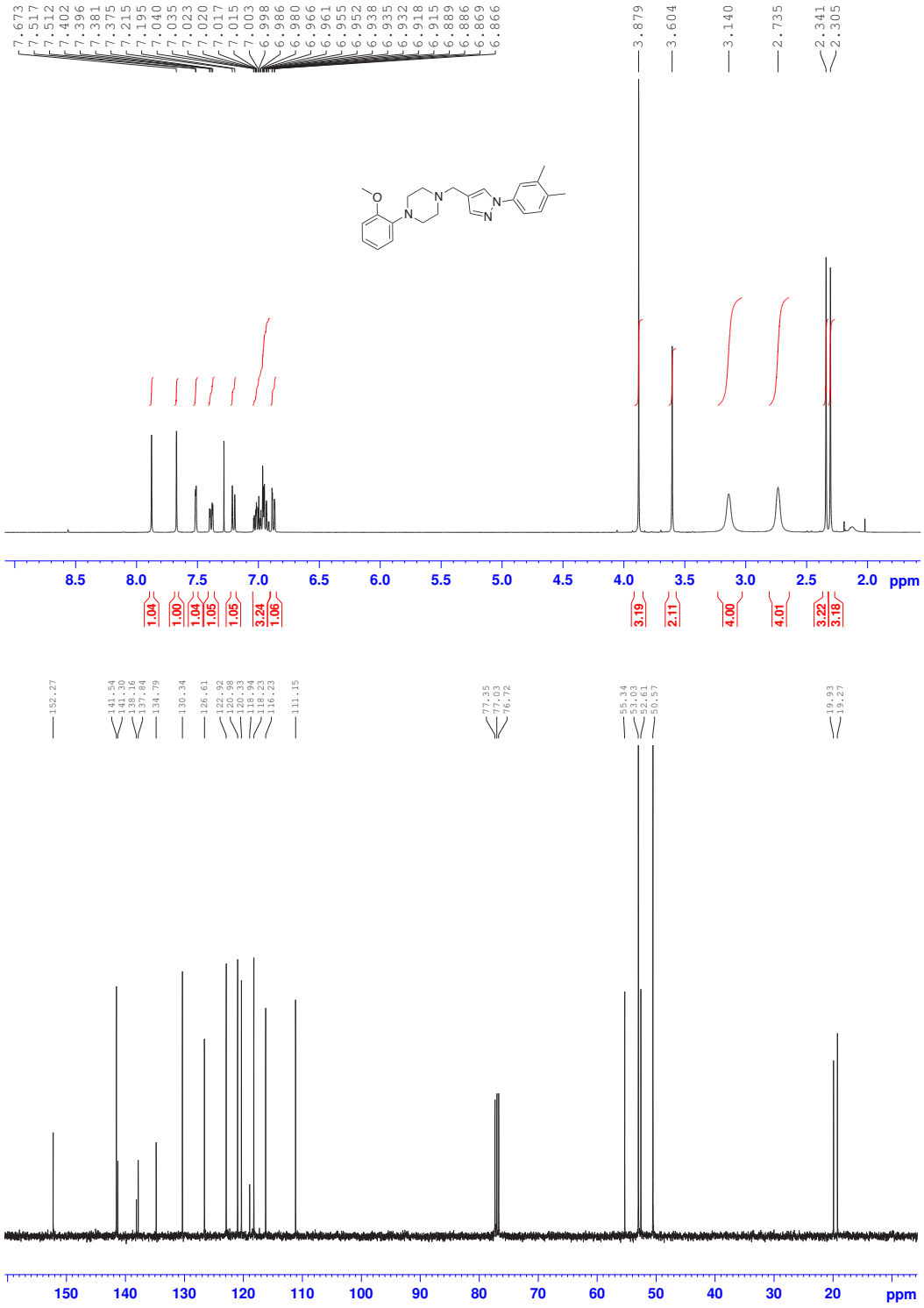


Figure S28. ^{1}H and ^{13}C NMR spectra of **26**.

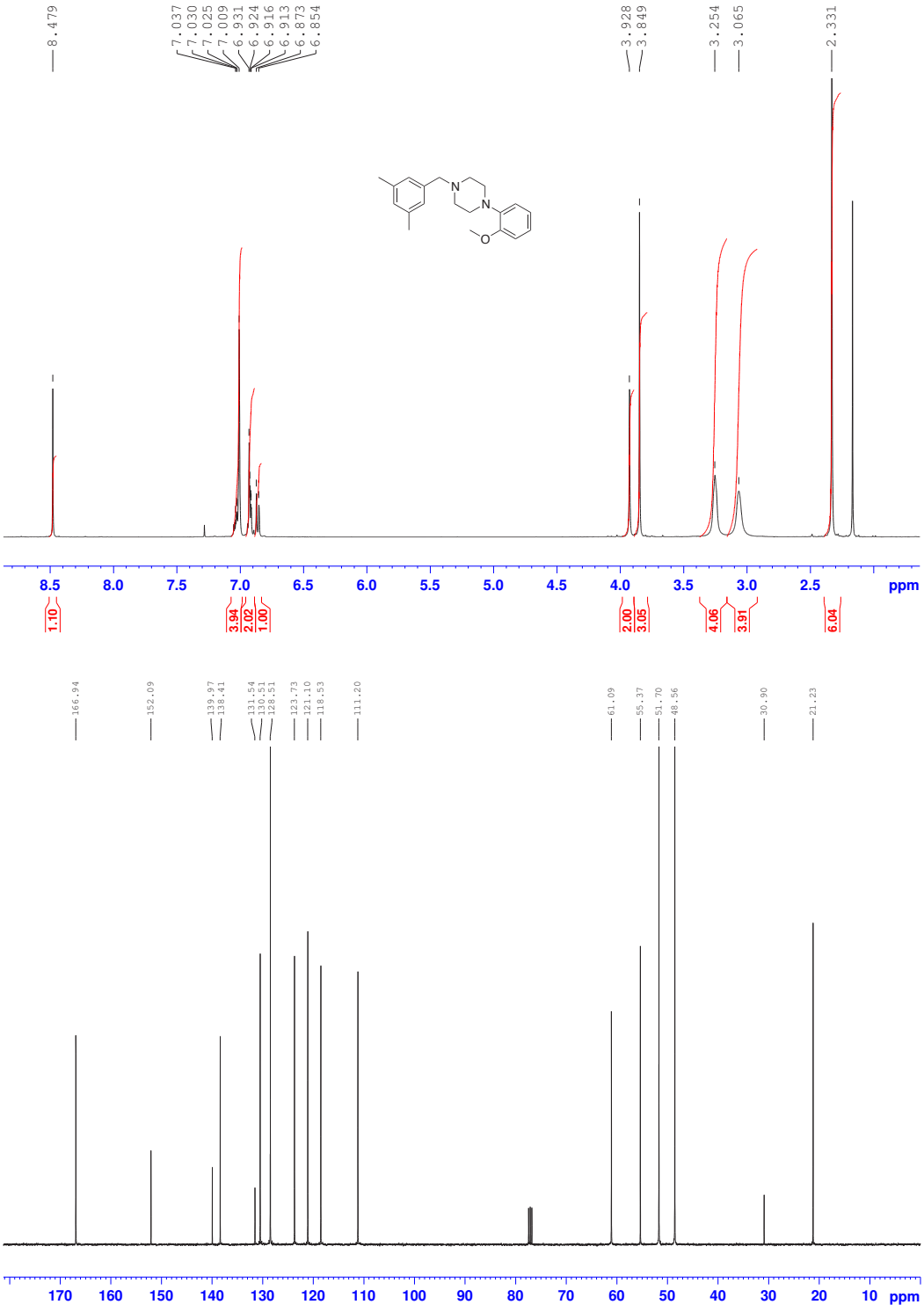


Figure S29. ^{1}H and ^{13}C NMR spectra of **27**.

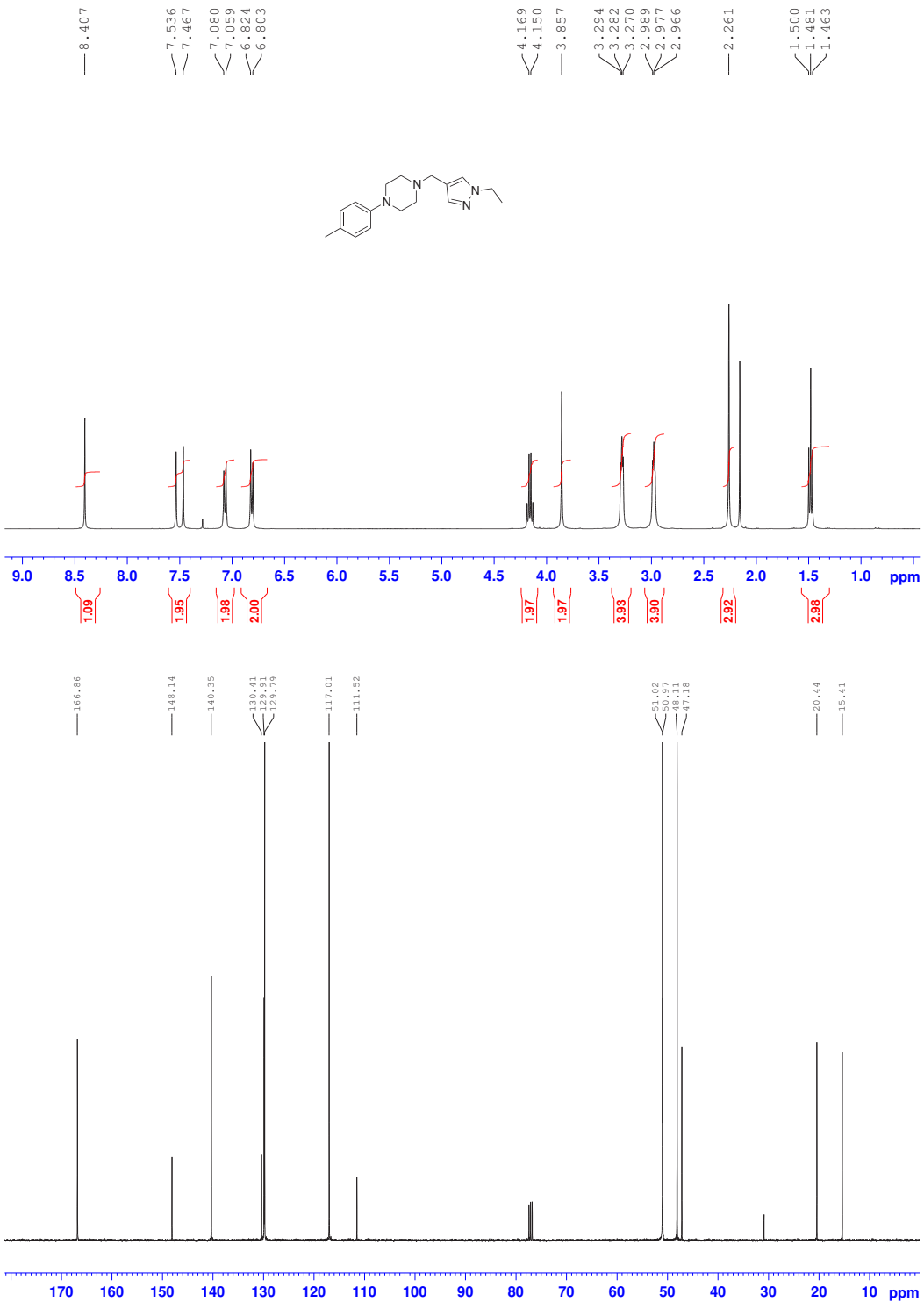


Figure S30. ^{1}H and ^{13}C NMR spectra of **28**.

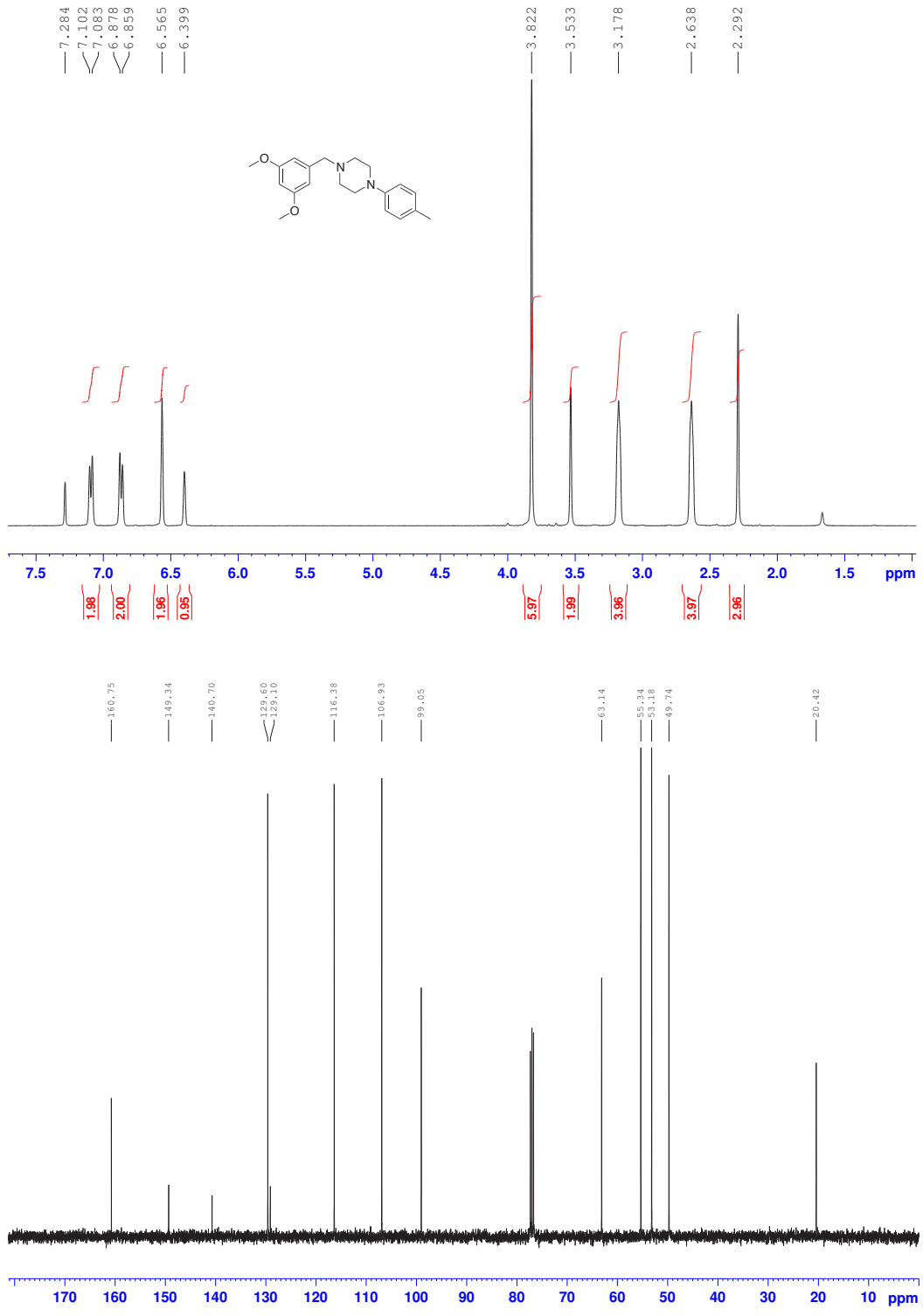


Figure S31. ^{1}H and ^{13}C NMR spectra of **29**.

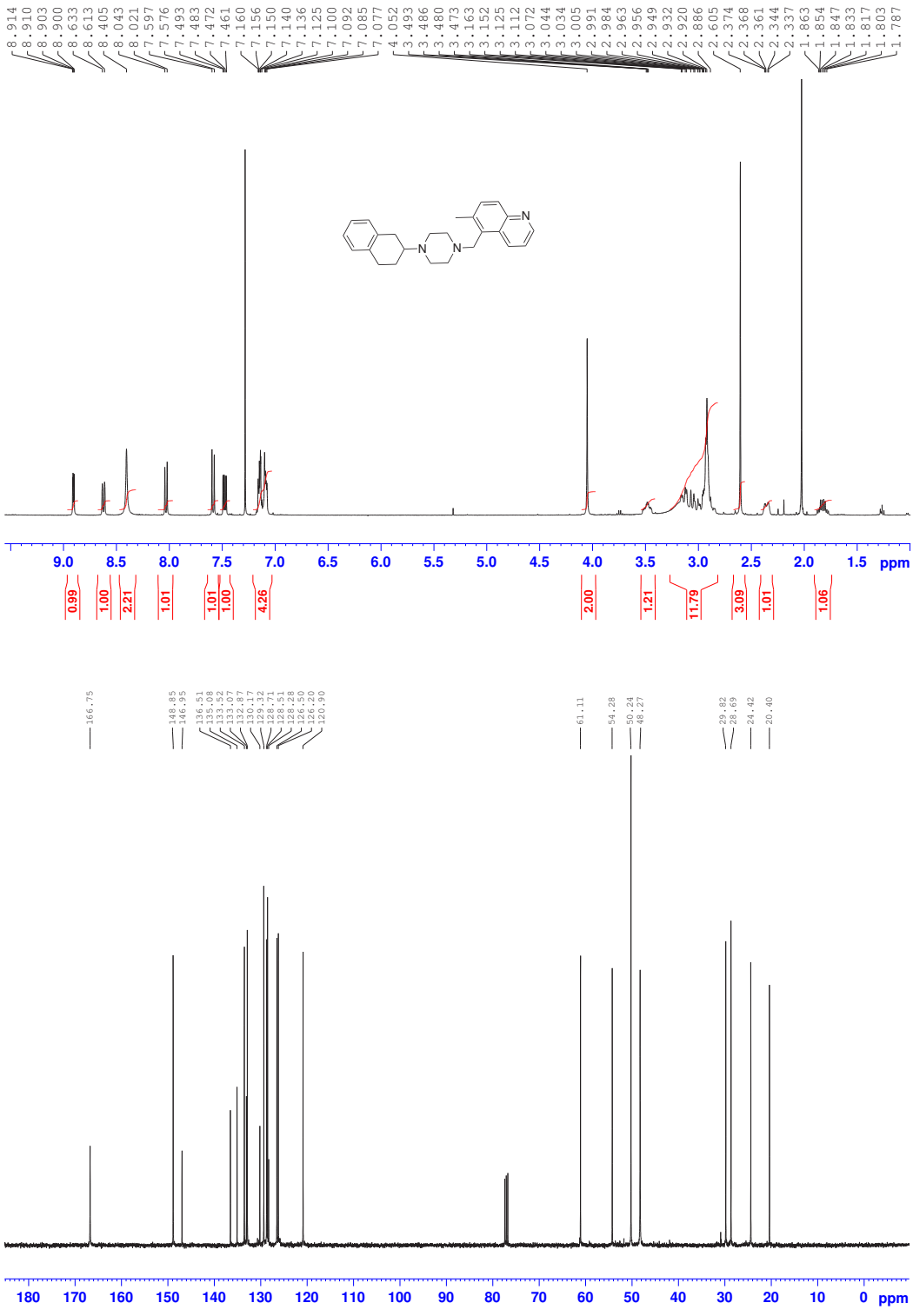


Figure S32. ^{1}H and ^{13}C NMR spectra of **30**.

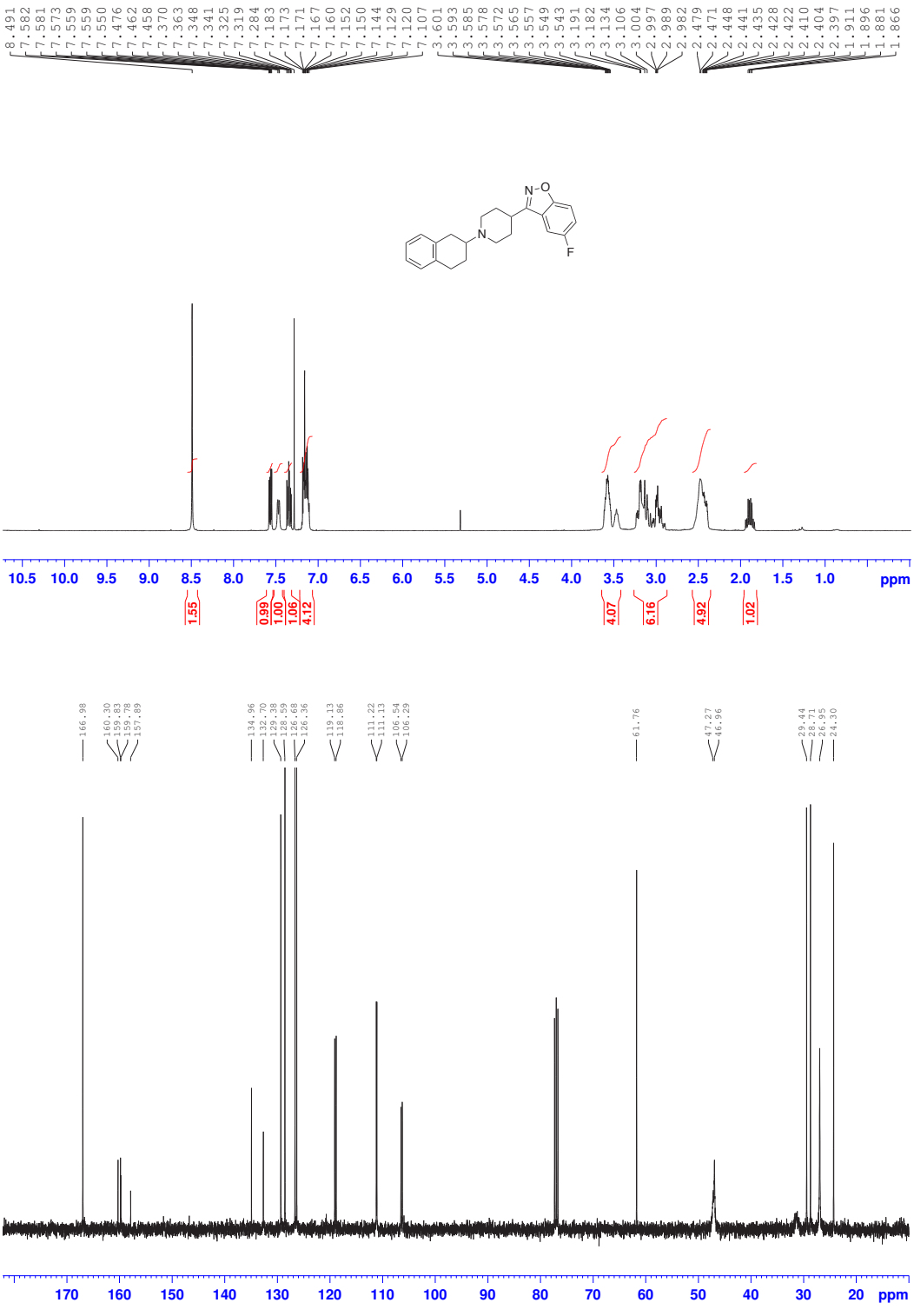


Figure S33. ^{1}H and ^{13}C NMR spectra of **31**.

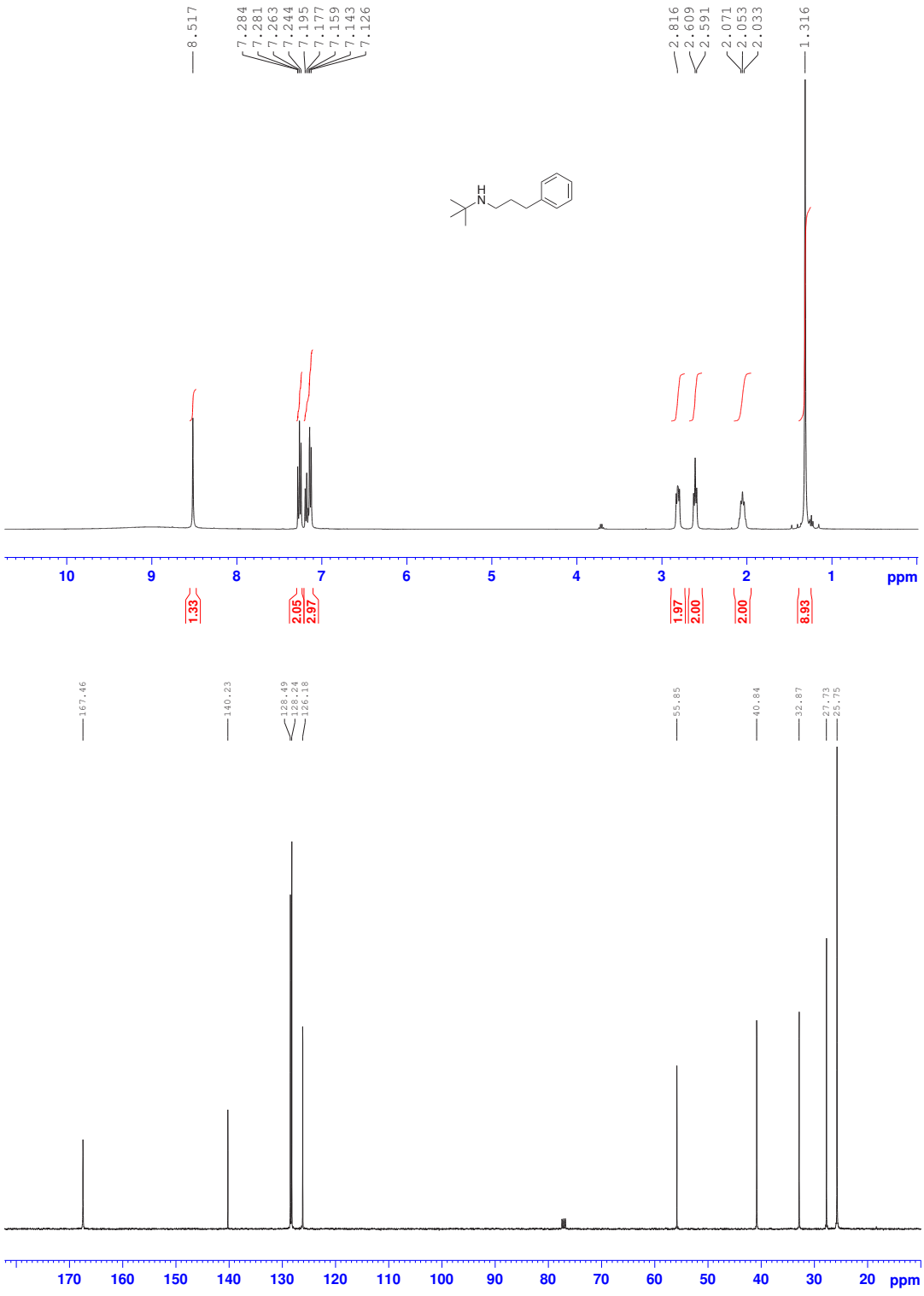


Figure S34 ^{1}H and ^{13}C NMR spectra of **32**.