ViSAS for Entering Chemical Space: Virtual Screening of Analog Series and Related Advances

José J. Naveja-Romero,[1,2] Fernanda I. Saldívar-González,[3] Diana L. Prado-Romero,[3] Angel J. Ruiz-Moreno, [4] Marco Velasco-Velázquez,[5] Ramón Alain Miranda-Quintana,[6,7] José L. Medina-Franco [3,*]

[1] Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Mainz 55131, Germany. [2] Johannes Gutenberg-Universität Mainz. Institut für Molekulare Biologie gGmbH (IMB), Mainz 55128, Germany. [3] DIFACQUIM Research Group, Department of Pharmacy, National Autonomous University of Mexico, Mexico City 04510, Mexico. [4] University of Groningen, University Medical Center Groningen, The Netherlands. [5] Departamento de Farmacología, Facultad de Medicina, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico. [6] Department of Chemistry, University of Florida, Gainesville, Florida 32611, United States. [7] Quantum Theory Project, University of Florida, Gainesville, Florida 32611, United States.

*Contact author: medinajl@unam.mx

Abstract: The manuscript discusses recent advances on computer-aided drug discovery (CADD) with focus on data-dependent drug discovery. Herein, we do not intend to review the many CADD methodologies comprehensively. Instead, the review discusses progress on selected concepts, methodologies, resources, and applications that are part of multidisciplinary efforts: the manuscript covers advances in artificial intelligence, machine learning, virtual screening, and chemical space including the concept of chemical multiverses, and novel extended similarity methods for chemical space exploration. Throughout the review, we emphasize public resources and open-source code available to the scientific community working in academia and non-profit institutions.

Keywords: chemical space; chemoinformatics; computer-aided drug design; compound databases; de novo design; medicinal chemistry; molecular modeling; open science; similarity; virtual screening.

Abbreviations: AI, Artificial intelligence; CADD, computer-aided drug design; CLN, Chemical Library Networks; COCONUT, Collection of Open Natural Products; CSN, Chemical Space Networks; DL, deep learning; DNMT, DNA methyltransferase; IP, intellectual property; LBDD, ligand-based drug design; ML, machine learning; QSAR, quantitative structure-activity relationships; QSPR, quantitative structure-property relationships; RECAP, retrosynthetic combinatorial analysis procedure; SAR, structure-activity relationships; SBDD, structure-based drug design; SIR, structure-inactivity relationships; TPSA, topological polar surface area; ViSAS, Virtual Screening of Analog Series.

1. Introduction

Chemical compound databases with annotated bioactivity data provide an essential basis for various applications in drug design. Through computer-aided drug design (CADD) and recently with new artificial intelligence (AI) techniques, it has been possible to accelerate the generation of knowledge from big data in biological, chemical and pharmaceutical medicine.[1] The methods developed in CADD, which have been optimized with machine learning (ML) algorithms, can use the vast chemical space combined with its biological information to obtain compounds with safety, efficacy, and low toxicity, a goal in many drug design projects. CADD has led to the identification and development of many drugs used in the clinic and clinical development.[2] Figure 1 shows chemical structures of drugs in clinical use and clinical development where CADD methods have contributed to their identification or development.

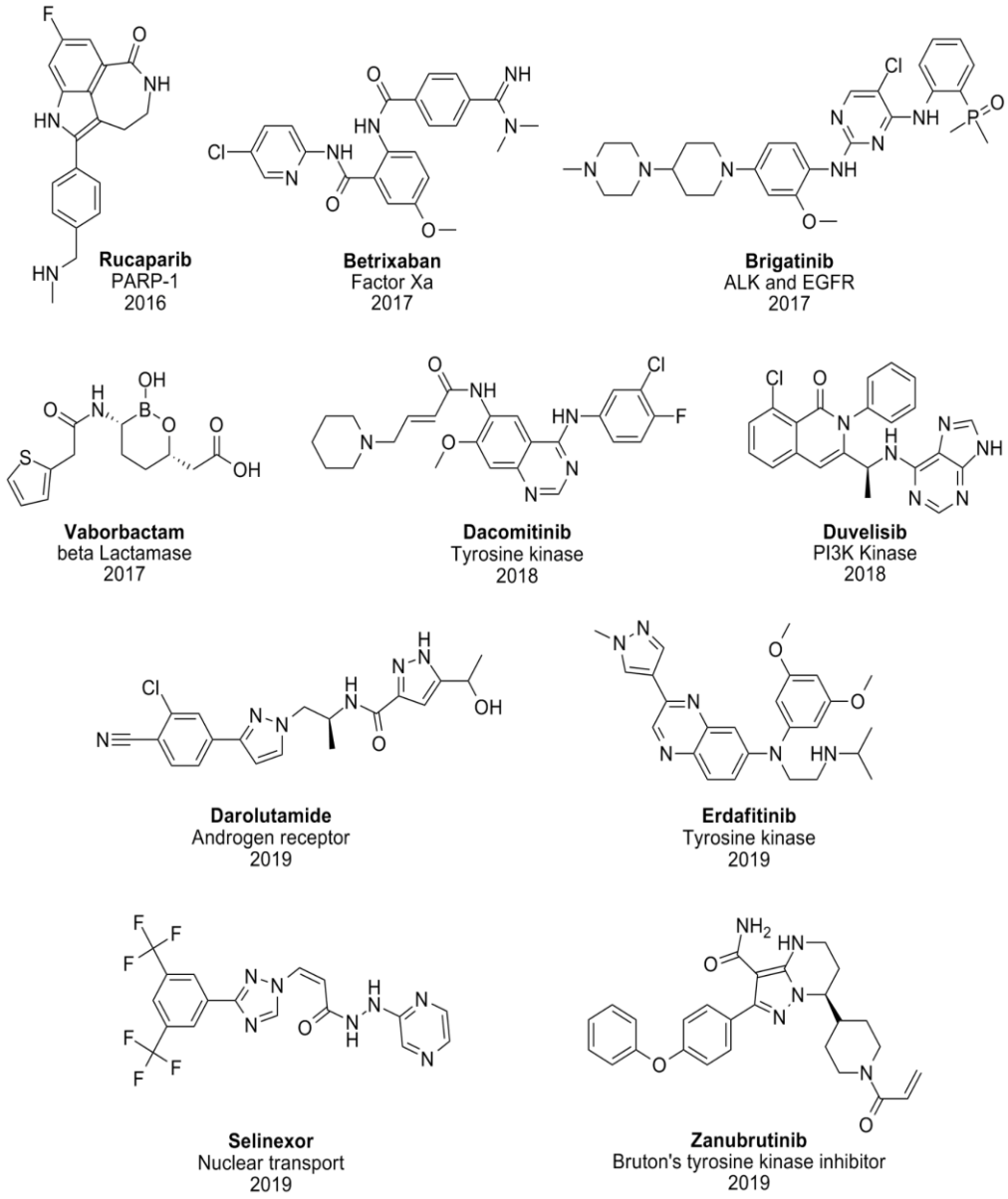


Figure 1. Chemical structures of exemplary drugs recently developed with the aid of computer-aided drug design. The main target and approval year are indicated.

In the last two decades, substantial improvements to structure- and ligand-based drug design methods developed in CADD have been described, many of them driven by AI and its subfields ML and deep learning (DL), as recently discussed in several review papers and special issues.[3–6] For instance, in structure-based drug design (SBDD), the prediction of three-dimensional structures with the AlphaFold2 neural network has generated the most complete and accurate picture of the human proteome,[7] even highlighting its applications in cases where not similar structure is known.[8] Other notable applications of DL are predictions of chemical reactions [9], synthesis automation and *de novo* design.[10]

The goal of the review is to discuss recent progress on selected concepts, resources, methodologies, and applications of CADD. Because of the broad scope of CADD, this manuscript is not meant to be a comprehensive review of the subject. It discusses progress on representative concepts, resources, and applications of CADD that are part of multidisciplinary efforts to advance drug discovery. Throughout the manuscript, we emphasize public resources broadly available to the scientific community. In this regard, we highlight open science. The manuscript is organized into six sections. After this introduction, the next section analyzes the role of bioactivity data in CADD and discusses advances and opportunities in SBDD and ligand-based drug design (LBDD). Section 3 addresses the chemical space and chemical multiverse concept to analyze content and diversity of chemical libraries. Emphasis is placed on constellation plots, which are based on the analog series concept, as a visual representation of the chemical multiverse. Section 4 explores exemplary and recent applications of CADD to identify hit and lead compounds. Therein, we introduce the concept of ViSAS: Virtual Screening of Analog Series, an implementation built upon the analog series formalism by Bajorath et al.[11,12] designed to expand bioactive molecules from the screening of chemically related compounds in ultra-large databases. Section 5 describes recent advances in the development and application of extended similarity methods. Section 6 presents summary conclusions and perspectives.

2. Exploiting bioactivity data in the artificial intelligence era

In the last decade, there has been an important increase in the amount of open bioactivity data in public data repositories. Databases with biological activity annotations have been extensively reviewed.[13–15] As case in point, the latest release of the ChEMBL database (v31) contains data for 14,855 targets, 2,786,911 distinct compounds, and 19,286,751 activities. With the abundance of publicly available bioactivity data, it becomes imperative to extract, curate and explore information of interest for drug discovery, so data-driven drug discovery models have emerged.[16]

Analyzing information derived from large-scale structure-Activity relationships (SAR) data can contribute to understanding the underlying mechanism associated with the structural transformations of compounds that modify their activity, in the prediction of potential target proteins for therapeutics and in network pharmacology,[17,18] linking targets in terms of shared active ligands. Similarly, available bioactivity data plays an important role in computational chemogenomics to develop predictive models [19,20]. Other approaches focus on identifying and describing biologically relevant regions of chemical space to guide the design and synthesis of new compounds.[21]

Recently, the importance of disclosing inactivity data in the public domain has been highlighted. There may be more information in the fact that a compound does not show activity. In this regard, López-López et al. introduced the notion of structure-inactivity relationships (SIR) highlighting the importance of including reliable data of inactive compounds in the development of descriptive and predictive models.[22]

In order to generate information and knowledge, the quantity and quality of data are vital as it improves the development and observed performance of chemoinformatics and AI models. As a scientific community we should prioritize access to complete data, e.g., activity and inactivity data (negative results) that enable researchers to access the “big picture” of the available knowledge. This comprehensive viewpoint could help to cope with data imbalance that we have to deal with on a daily basis in drug design and compound optimization campaigns. Additionally, data curation and the construction of reliable databases are major issues. The poorly curated databases complicate the assessment of the predictive performance of ML and DL models. However, combined efforts could facilitate access to new interesting data. Examples include natural products, metallodrugs, safety, preclinical, and toxicological databases that complement the current data available in the public domain and offer a new perspective on the known data. However, we are aware of the potential conflicts of interest related to the publication of data susceptible to intellectual property, e.g., post-marketing data that has reporting bias related to the time and clarity of shared data.

AI methods enable the parallel study of very large volumes of diverse data for ligand-based drug design. However, SBDD approaches have not yet fully explored the utility of AI, although much research is in progress. One of the reasons is that experimental structural data are still sparse compared to compound activity and physicochemical data. The current protocols' limitations only partially enable the generation of reliable 3D conformational states or binding modes. On the other hand, recent progress in AI-driven *de novo* structure prediction (see Section 4.3) has provided an unprecedented wealth of putatively reliable structural templates, with coverage recently approaching the entire protein universe.

There are several recent reviews discussing examples of AI applications in drug design and development.[4,23–26] Some challenges and opportunities that face AI discussed by experts in the industry, academia, and other institutions were recently discussed on a public online event.[27] It has been emphasized that sufficient knowledge and correct application (beyond the hype) are necessary. For this reason, it has been proposed to integrate “augmented intelligence” models into drug design, which shows a trend towards almost total automation (“*Human-assisted*”). This model of partnership between human intelligence and AI aims to improve cognitive performance, including learning, decision-making and the generation of new experiences by leveraging the capabilities offered by AI models and the medicinal chemist's own expertise.[28]

2.1 Databases annotated with biological activity

Given the usefulness of chemical databases, many companies and researchers have taken it upon themselves to compile and put on web servers databases with diverse information. In an effort to classify these databases, Masoudi-Sobhanzadeh et al. distinguish five classes useful for drug repositioning based on

data content, including: raw data (e.g. data from literature, *in-house* and clinical), target-based (include genes, proteins, pathways and side effects information), specific data (traditional medicine, disease-specific or geographical databases), drug design (containing the 3D structure of molecules, and molecular replacement information which are based on the resolved protein structures), and tool-based DB (tools and web servers).[15]

Other databases that are having a major current trend in drug discovery are virtual compound libraries and *de novo* designed libraries.[29] Also, it is ongoing an effort to build and curate a compound database with metal-containing molecules in preclinical and clinical development, and approved for therapeutic use.[30]

Natural products have been sources of bioactive compounds that later have been used in the clinic or that have been used as starting points of drug candidates.[31] The application of computational methods including chemoinformatic approaches and AI to further advance natural product research is a current trend.[32–34] As early as 2012 there have been efforts to put together natural product databases in the public domain.[35] The most recent large compound collection of natural products in the public domain is the Collection of Open Natural Products, COCONUT.[36] In Latin America, there is an ongoing effort to put together and curate compound databases composed of natural products from the vast diversity contained in Latin American countries.[37,38]

2.2 Opportunities of AI in ligand-based drug design

In addition to *in vitro* and *in vivo* methods, we can use *in silico* methods to mitigate serendipity and rationalize those phenomena that experimental methods cannot explain. During rational drug design, serendipity might occur, leading to unexpected but potentially positive results such as the
discovery of Lyrica (pregabalin).[39] However, the dramatic increase of data in chemical databases with biological annotations limits the chance of serendipitous positive results and calls for enhanced methods for the identification of molecules with clinical application.[6] Such tasks can be addressed by using AI.[6, 40]

Increasing numbers of AI algorithms are being developed for predicting the relationship between chemical structure and biological activity. For example, DeepChem is an open-source platform with tools for applying AI algorithms that allow the prediction not only of biological activity but also of multiple drug properties, including physicochemical features, and toxicity.[41] DeepChem has supported the development and benchmarking of new AI models with diverse goals. [42–44]

Efficacy prediction using diverse inputs (chemical and physicochemical properties, biological -*in vitro* and *in vivo*- information, -omics, preclinical, clinical, and post-marketing data) is one of the main objectives of applying AI in drug design. For example, Wang et al. used a model based on Support Vector Machines (SVMs) to discover nine new compounds and their interactions with four key targets. The model was trained on 15,000 protein-ligand interactions and was developed based on primary protein sequences and structural characteristics of small molecules. [45] Yu et al. used two random forests (RF) models with high sensitivity and specificity to predict possible drug-protein interactions by combining pharmacological and chemical data. [46]

KinomeX, an AI-based online platform trained with over ~14 000 bioactivity data points derived from over ~300 kinases, can efficiently investigate the overall selectivity of compounds for the kinase family and specific subfamilies of kinases, which can aid in developing new chemical modifiers.[47] A last example is PyRMD, an AI algorithm based in Random Matrix Discriminant (MD) -a subtype of ML- that can be trained to recognize the distinctive pharmacophoric features from the target bioactivity data available at the ChEMBL. Selected negative data are incorporated in the learning to identify structural features that are irrelevant or detrimental for the intended bioactivity.[48]

Identification of toxic effects in early stages of drug design allows to remove undesirable characteristics of bioactive compounds. At present, multiple AI-based are employed to assess toxicity by predicting the off-target ligand binding. For example, Ligand Express, Cyclica's cloud-based AI platform, uses proteome-screening data to find receptors that can interact with a specific small molecule, predicting on- and off-target interactions and suggesting the drug's potential side effects.[49] Other AI web-based tools that help predict toxicity include LimTox, pkCSM, admetSAR, and Toxtree.[50] A particularly remarkable case is DeepTox, an ML-based algorithm that using features within chemical descriptors accurately predicted the toxicity of 12,707 environmental compounds and drugs during the Tox21 Data Challenge.[51] The DeepTox algorithm uses static descriptors, such as molecular weight, Van der Waals volume, or the presence/ absence of a predefined substructure or a toxicophore descriptor, as well as calculated dynamic descriptors.[51] Despite a potentially infinite number of different dynamic features, the method keeps the dataset within manageable limits and shows good accuracy in predicting the toxicology of compounds.

After a molecule has been virtually screened for potential bioactivity and toxicology, a chemical synthesis pathway is required for their evaluation in relevant models of disease. Despite knowledge of hundreds of thousands of transformation steps, novel molecules cannot be efficiently synthesized due to novel structural features or conflicting reactivities.[52] AI can help to identify possible and less complicated synthesis routes for compounds simultaneously or sequentially with prediction of bioactivity. [53] Computer-aided synthesis planning can also suggest millions of structures that can be synthesized and predict multiple synthesis routes for each of them.[54]

New AI methods can support multiple applications such as analog series identification (fragmentation), *de novo* drug design signatures study, SAR visualization, reactivity predictions, similarity searching, and visualization of chemical space. Two examples of such methods are Extended Similarity Indices developed by the research group of Miranda-Quintana,[55, 56] and the Structure-Activity Relationships Matrix approach and its deep learning extension by Bajorath et al.[57]

A strategy still to be consolidated is data expansion or augmentation[58] using multiple layers of inputs. This approximation could allow the generation of the most representative similarity searching to identify chemical mimetics capable of reverting disease signatures (instead of altering one molecule activity). For example, drug design approaches might be developed for reverting (or preventing) molecular pathway alterations or for predicting toxicity or safety issues for marketed drugs.

2.3 Opportunities in structure-based drug design

SBDD has reached notable maturity over the past decades, especially structure-based virtual screening, despite its intrinsic limitations.[10] In recent years, DL has been used in attempts to improve the performance of SBDD methods further. Perhaps the most well-known example of this is the usage of DL for protein structure prediction. *De novo* structure prediction with Alphafold [8] and RoseTTAfold [59] or other programs has yielded many protein models of near-experimental accuracy which has further expanded the opportunities and applicability domain of homology modeling. Other uses of AI in SBDD include but are not limited to potentials similar to quantum-chemical descriptions (ANAKIN-ME) [60] force field development;[61] Boltzmann generators trained to identify transition states f;[62] protein-ligand interaction fingerprints [63] such as SPLIF [64] or extended connectivity interaction features (ECIF),[65] and scoring functions like GNINA.[66]

Recently, the geometric DL approach was used to learn distance distribution and ligand-target interactions or predict the binding conformation of bioactive compounds. This potential performs similarly or better than well-established scoring functions.[67] Geometry DL uses a mesh on the protein surface as a molecular representation.

In all, AI should be used and practiced for the right reason and not because of just hype or a trendy fashion in current drug discovery.[68]

3. Chemical space and chemical multiverse

Chemical space, sometimes referred to in the literature as the “chemical universe” [69] and recently extended to “chemical multiverse”,[70] is a concept that is central and distinctive of chemoinformatics as an independent theoretical discipline.[71] Chemical space refers to all possible molecules as well as multi-dimensional conceptual spaces representing their structural and functional properties, depending on the type of representation.[72] Indeed, in contrast to cosmic space, the chemical space is relative to the structural and functional properties used to construct or define a given chemical space. Since there is not a molecular representation that captures all structural and functional properties, the concept of “chemical universe” has been introduced to account for the alternative chemical spaces of a compound data set that can be generated by different sets of descriptors.[70] Structural representation is the most relevant feature in basically any chemoinformatics application, and computational study [73] and it is an area under constant research.[74] In virtual screening, defining the chemical space to be explored is crucial, as it defines the applicability domain that will be searched. In practice, it is common to conduct virtual screening campaigns focused on regions of the medicinally relevant chemical space.[75, 76] Nonetheless, it is becoming a regular practice to explore novel regions of the chemical space, given by the emerging large- and ultra-large chemical libraries.[77]

The chemical space concept has practical applications in many areas of chemistry including drug discovery, organic synthesis, food chemistry, and material sciences to name a few examples. A key distinction between the different ways of mapping compound data sets into the chemical space lies in the type of descriptors that are used to represent the compounds of interest. For instance, the nature of the descriptors employed to represent small organic molecules will be different to describe chemicals with applications in, for

example, material sciences. In some cases, the chemical space concept is used to guide drug discovery projects, but a generalized or unique manner to represent visually the chemical space remains elusive. A typical example of this challenge is the visual representation of the chemical space of metal-containing compounds.[30]

In drug discovery, chemical space was used as a spatial navigation framework, helpful for understanding and generating knowledge of pharmacokinetic properties and molecular diversity of biologically relevant compounds [78]. As the number of compounds and their information in chemical databases increased, more sophisticated molecular descriptors and visualization techniques were developed to expand their applications. For example, explorations of chemical space have considerably improved our understanding of biology and have led to the development of many tools for the exploration of SAR and structure-property relationships (SPR).[79, 80] The availability of software libraries and the rise of AI [9] have led to the emergence of several tools that integrate ML methods as versatile tools to design, generate, and visualize the chemical space of small molecules.[81, 82]

3.1 Recent progress on chemical space

The chemical space concept has been of interest in several areas of chemistry for a number of years. However, the rapid and continued increase of large- and ultra-large libraries has renewed the interest of the scientific community to generate/implement methods to handle and use for practical applications the large- and ultra-large chemical space associated with the newly generated compound libraries.[29] Hence, the chemical space concept continues to be of significant interest to study the very large chemical libraries. There are several reviews addressing the concept of chemical space, covering different aspects such as enumeration of chemical compounds using *de novo* design, calculation of molecular descriptors, progress on visualization methods with emphasis on publicly available tools, web servers to explore the chemical space of chemical libraries, applications to study structure-property relationships.[81–83] A recent development in this area is the chemical library networks.[84] This development is further elaborated in Section 5.

3.2 Chemical multiverse and constellation plots

Another recent progress in the research on chemical space is the introduction of the term *chemical multiverse* as an expanded view of the chemical space.[70] This novel term is based on the chemical space concept that implies that a set of *m* molecules described with different descriptors would lead to distinct chemical spaces. Varnek and Baskin points out that “unlike real physical space, a chemical space is not unique: each ensemble of graphs and descriptors defines its own chemical space”.[71] It follows that molecules with different chemical natures, e.g., macromolecules, metal-containing compounds, biologics, etc., yield distinct chemical spaces because of the nature of the descriptors required to represent the compounds.

In physics, Everett’s multiverse [85] is “a hypothetical collection of potentially diverse observable universes, each of which would comprise everything that is experimentally accessible by a connected community of observers.” Thus, the multiverse “is a hypothetical group of multiple universes.” In analogy with

the cosmic multiverse, the chemical multiverse was defined as “the group of numerical vectors that describe it differently from the same set of molecules”.[70] A chemical multiverse can also be seen as a “group of multiple chemical spaces, each one defined by a given set of descriptors.” As reviewed recently [70] different chemical space representations can lead to alternative spaces, and the relationships between chemical compounds could change. It has been shown that the concept of chemical multiverse is applicable to different types of molecules such as small organic molecules and peptides for drug discovery applications, food chemicals, and natural products. Eventually, the chemical multiverse can be expanded to any type of compounds, including inorganic compounds.

One approach to analyze chemical multiverses is through constellation plots. A common limitation of most visualization methods of chemical space is that they capture a single type of molecular representation, emphasizing the dependence of the chemical space on the structure representation. To address this issue, constellation plots, generally depicted in Figure 2, combine, in a single graph, multiple structural representations providing a broader perspective of the contents, diversity, and, if desired, a property of interest (e.g., biological activity, either experimental or predicted). Specifically, constellation plots combine a coordinate-based chemical space representation of analog series. Constellation plots facilitate the identification of entire zones in chemical space enriched with active compounds (‘bright’ SAR) or with predominantly or all inactive molecules (‘dark’ regions or “black holes”). In analogy with the cosmic space, the name ‘constellations’ is associated with clusters of analog series with similar chemical structures (given by similar coordinates in the two-dimensional plot). Combining multiple structural representations or, more generally, complementary approaches, is founded on the general notion that multiple and well integrated approaches perform overall better than individual methods.[86–89] Since constellation plots combine various structural representations in a single plot, it can be proposed that these plots are a manner to represent visually chemical multiverses.

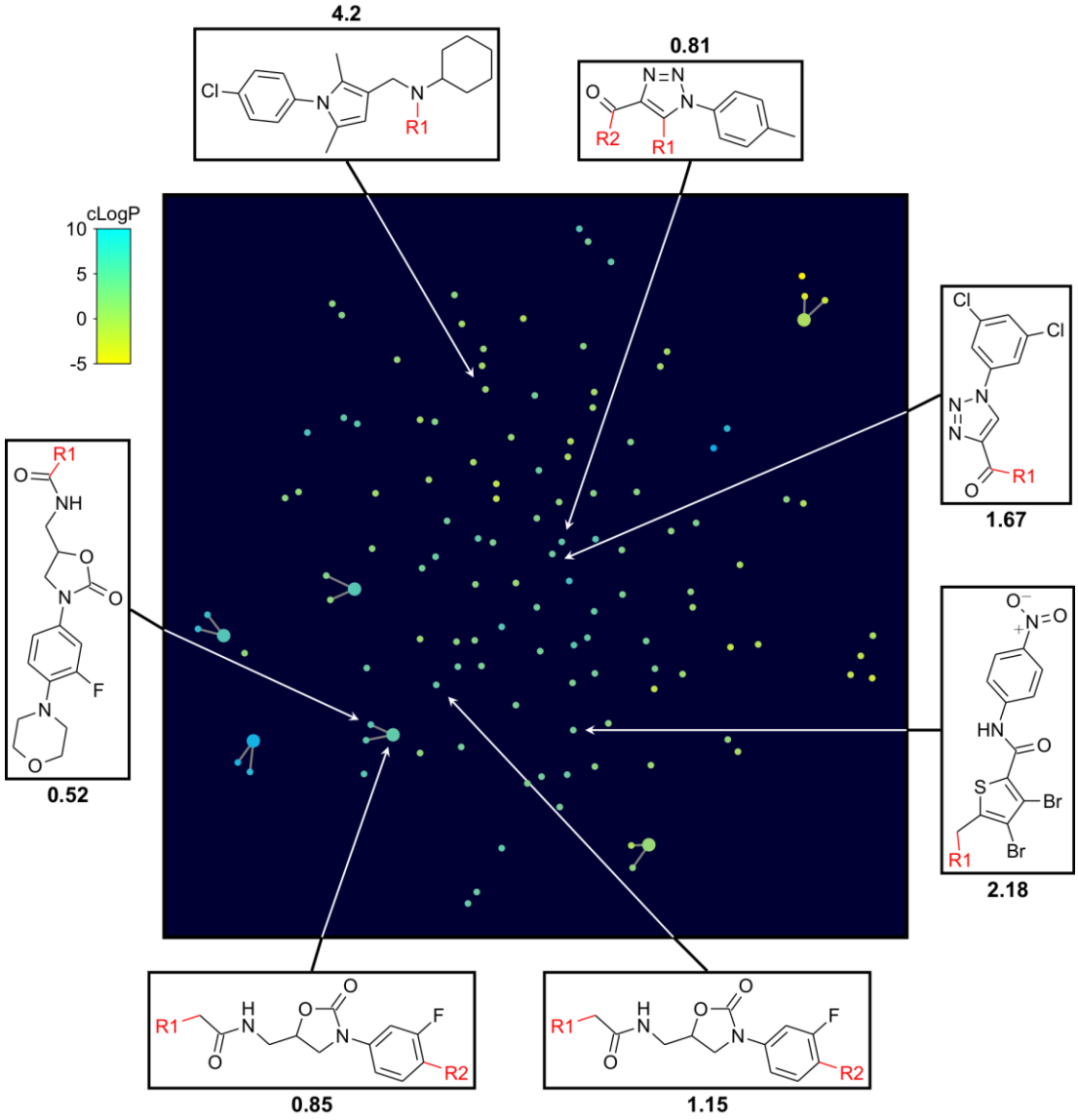


Figure 2. The general form of a constellation plot is illustrated in this image. Every core is represented by a dot, the size of which is proportional to the number of compounds mapping to it. Edges represent cores connected by at least one shared molecule in the dataset. The color coding can represent any feature, such as the average scores of the molecules represented by the corresponding core in virtual screening. In this example, the color indicates the average of the cLogP values of the compounds sharing the core structure.

Virtually any property of interest can be depicted in a constellation plot, such as experimental activity data or results from virtual screening, for instance, docking scores, predicted binding affinities, similarity values, or any other estimated value. This can be useful to identify, for instance, promising analog series for prioritization in experimental screening or additional computational studies before final selection for experimental evaluation.

Constellation plots have already been used to aid the visualization of chemical space for different practical applications. For example, the authors analyzed the results of a docking-based virtual screening of 2789 molecules from a commercial virtual library focused on inhibitors of DNA methyltransferase (DNMT). The

docking scores were visually represented on the plot, enabling the rapid identification and grouping of analogs (e.g, “constellations”) of compounds to be prioritized for further screening.[79] Constellation plots have also been used to explore the SAR of 827 inhibitors of AKT1 obtained from a public database, and the structure-multiple-activity relationships -SmART- of 286 molecules experimentally tested as inhibitors of three DNMTs and assembled from public sources [79, 90–92] consistent cell-selective analog series of chemical compounds. This analysis was done through a systematic analysis of high-throughput screening data of 41,821 compounds consistently assayed against the same panel of 73 human cancer cell lines used by the National Cancer Institute of the United States. In that study, the most relevant analog series were identified as measured by a therein developed combined selectivity and consensus score. Also, all the 3,750 cores of the entire data set were used as queries or reference structures to virtually screen the entire ZINC 15 database identifying 82,409 purchasable analogs for 1,980 of the 3,750 cores.[91]

One more recent application of the constellation plots was to contribute to a comprehensive SAR analysis of 851 compounds tested as tubulin inhibitors and bioactivity data in different cancer cell lines. A total of 147 analog series were identified and analyzed in a constellation plot. Visual analysis of the plot (an interactive version of the plot was made freely available using DataWarrior,[84] rapidly identified “bright” and “dark” regions in chemical space, i.e., analog series with overall high and low activity, respectively, as inhibitors of tubulin.[92] The code to generate constellation plots is freely available at https://github.com/navejaromero/analog-series.

4. Hit Identification, optimization, and development of bioactive compounds

One of the most frequent approaches to identify active compounds from large compound libraries is through the computational filtering of possibly large or extremely large screening compound databases, followed by the relevant experimental validation. Certainly, virtual screening is a widely used tool in drug discovery.

4.1 Virtual screening

Virtual screening is a general approach devised to predict promising molecules, termed computational hits, that, upon experimental validation, have the potential to turn into lead molecules.[93, 94] Hits selected based on other features, such as physicochemical properties and smooth SAR, become leads ready to undergo optimization cycles. After intensive optimization, a drug-like molecule, termed a clinical candidate, might be considered for further preclinical and clinical development.[95] This process can be extremely costly and time-consuming and is bound to high attrition rates. Therefore, virtual screening could assist in hit identification at early preclinical stages.[94] Iterative rounds of virtual screenings [96] and scaffold-based analyses [97] could then aid in hit expansion and lead optimization. In a recent example, Steadman et al. [98] reported a docking-based virtual screening to identify new inhibitors of Notum, a negative regulator of Wnt signaling. They screened several successful series and found the [1,2,4]triazolo[4,3-b]pyradizin-3(2*H*)-one series as a new chemical class of Notum inhibitors.

With the rise of large and ultra-large chemical databases, virtual screening has evolved as a natural way to exploit their contents and diversity.[99, 100] Besides a database to search in, virtual screening requires additional information, for example, the receptor's structure and a force field for docking scoring (example of a structure-based approach) or known ligands and a system for assessing similarity (example of a ligand-based approach). Throughout this section, we will focus on ligand-based virtual screening, in particular, similarity-based virtual screening. This approach typically includes using one or more bioactive molecules as queries or references to compare against the database. The compounds most similar to the reference molecules are the computational hits.[101] Measuring similarity is, nonetheless, dependent on the operational definition of similarity: it might use, for example, chemical fingerprints, physicochemical properties, pharmacophoric features, and even combined approaches.[102] Several strategies for performing ligand-based virtual screening exist, but there is no consensus on the best method for every case.

Table 1 lists a few examples of successful cases of virtual screening.[2] This technique is gaining attention as the research community witnessed the applications identifying candidate compounds against COVID-19 from ultra-large compound databases.[103] Despite the extensive contributions and progress of virtual screening, this is one of the current grand challenges that face CADD. The hurdles include defining the search space, i.e., the type of chemical libraries to be explored, improving the algorithms to augment the hit rate, and developing or refining the current virtual screening processing tools to enhance the quantity and quality of the computational hits as well.[104]

Table 1. Examples of recent successful virtual screening campaigns including experimental validation.

Hit compound(s) activity	Virtual screening approach	Reference
11β-HSD1 inhibitors.	Growth-based screening of the ZINC database (1.8 million compounds).	[105]
Two inhibitors of SARS-CoV-2 M^{pro} inhibitors with IC_{50} values in the micromolar range.	Docking-based virtual screening of an in-house focused library.	[106]
Thirty-two inhibitors of Notum with IC_{50} values lower than 500 nM.	Docking-based virtual screening of 1.5 million compounds in a synthetic and commercial library (ChemDiv).	[98]
Four histone deacetylase inhibitors with nanomolar activity vs. HDAC 1, 3, and 6.	Pharmacophore model and docking of an in-house database of 22,700 molecules.	[107]
Six compounds with activity against *Mycobacterium tuberculosis* peptide deformylase.	Docking-based virtual screening of a commercial compound library with 7,120 small molecules.	[108]

4.2 VISAS: general approach that expands bioactive molecules

Sometimes a focused virtual screening approach might be desirable. One way of delimiting the chemical space implies carefully selecting the collection to screen. For instance, the database could be confined to drug-like molecules, natural products, synthesizable compounds, or datasets focused on a set of molecular targets, e.g., focused or targeted libraries. Thus, the hits will comply with relevant selection criteria, depending on the project. A second factor to consider is the similarity threshold to define hit compounds. A more flexible hit definition puts novelty and scaffold-hopping in the spotlight, whereas stringent criteria would be suitable for identifying lead compounds, SAR analysis, and hit expansion. We propose that the most extreme hit definition requires that the query and hit molecules are chemical analogs, i.e., molecules with a close synthetical relationship. Since query and hits might as well have arisen from an organic synthesis project, it might be understood as a “pseudo-optimization” algorithm enabling the rapid extraction of purchasable or readily available analogs for experimental SAR exploration. We term this approach ViSAS (Virtual Screening based on Analog Series) since the practical implementation builds upon the analog series formalism by Bajorath et al.[11, 12] Figure 3 depicts two exemplary analog series according to the definition presented by Naveja et al.[109] Briefly, the process of finding putative cores for a molecule begins with fragmenting the molecule (for instance, using RECAP retrosynthetic rules [110]) and subsequently filtering for relevant fully-connected fragments that include most of the original structure (we require that at least two-thirds of the heavy atoms from the molecule must be included in the fragment’s structure). Fragments obtained through this procedure are termed putative cores. Although this method allows every molecule to map to more than a single core, large analog series can be usually summarized in a few cores that comprehensively map all molecules in the series (see Figure 3). Nevertheless, keeping a record of all putative cores permits the later inclusion of new molecules, which is the principle on which we base the virtual screening approach proposed here.

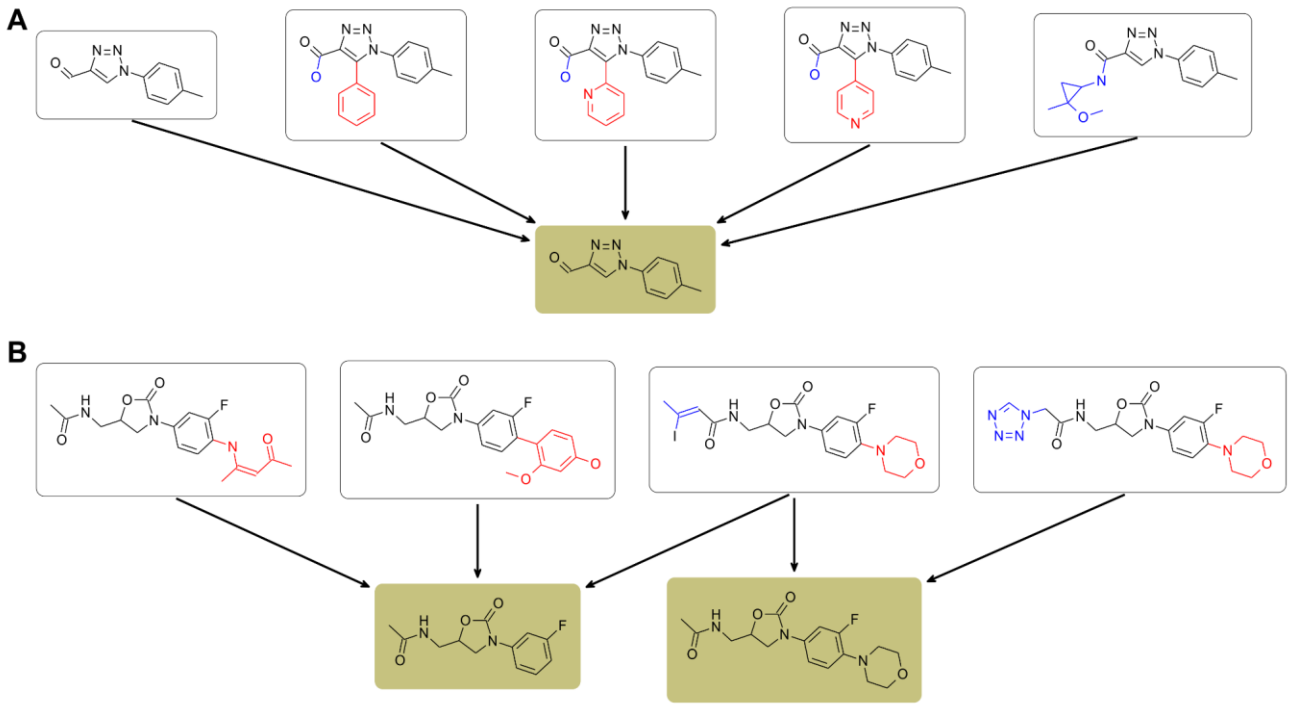


Figure 3. The general concept of analog series. All molecules in series A share a common core, which, for some applications, could be used to summarize it. Series B is somewhat more complex and requires at least two minimally overlapping cores for a comprehensive representation. Note that our definition of analog series allows every molecule to map to multiple cores. For clarity, not all putative cores are shown in this Figure. See reference [109] for more details on the fragmentation-and-indexing algorithm employed.

Algorithms and applications related to the automatic identification of analog series in large data sets have been reviewed.[12] For over a decade, the analog series algorithms derived from matched molecular pair analysis have demonstrated a compelling balance between chemical interpretability and scalability.[111] Recent developments have emphasized the ability of analog series for SAR and activity cliffs rationalization.[11, 112, 113] However, other industrial applications, such as the evaluation of progress in lead optimization,[114] highlight the potential for analog series analyses to assist drug discovery teams dealing with organic synthesis and biological evaluation.[12]

The logical formulation of virtual screening from the analog series emerges from the definition of chemical analogs: two molecules are considered analogs if they share a common core structure. Therefore, a typical fragment-and-index approach lists all possible matching cores for molecules in a dataset. Any new molecule that could be reduced to a fragment matching the fragments list would be an analog of the molecule(s) in the dataset indexed to this fragment. It remains only to define a fragmentation procedure and the requirements of a fragment to be considered a valid core. Many different such approaches have been reviewed elsewhere.[12, 111] For instance, exhaustive methods may consider every possible substructure to be a valid core. Nonetheless, such strategies might lead to practical limitations. For instance, even relatively small libraries of somewhat complex molecules might lead to a combinatorial explosion while exhaustive substructure enumeration. Furthermore, synthetic interpretability is not prioritized in this approach, thus leading to a harder

rationalization of the results. Therefore, matched molecular pairs obtained through retrosynthetic fragmentation [115] gradually developed into several applications relying on analog series computational identification,[11] such as analog series-based scaffolds [116], compact chemical space representations of analog series in constellation plots,[90] and the novel SAR rationalization approaches.[109, 113]

Another application of analog series yet to be fully harnessed is virtual screening in ultra-large libraries. While most virtual screening methods focus on identifying single molecules with a desired predicted property, working with analog series up front has the potential of readily identifying a whole family of compounds to be prioritized for additional computational analysis or tested experimentally for a richer and in-depth SAR analysis. In essence, ViSAS is a substructure search algorithm (see Figure 4). However, the valid substructures to search are delimited before a direct comparison between queries and compounds in the database to search occurs. This allows the fragmentation of the databases to be computed in advance, thus reducing the substructure search to a text-matching problem. Moreover, the inherent hierarchical structure of analog series can be represented as scaffold networks and R-group tables allowing prompt local SAR analyses early on.

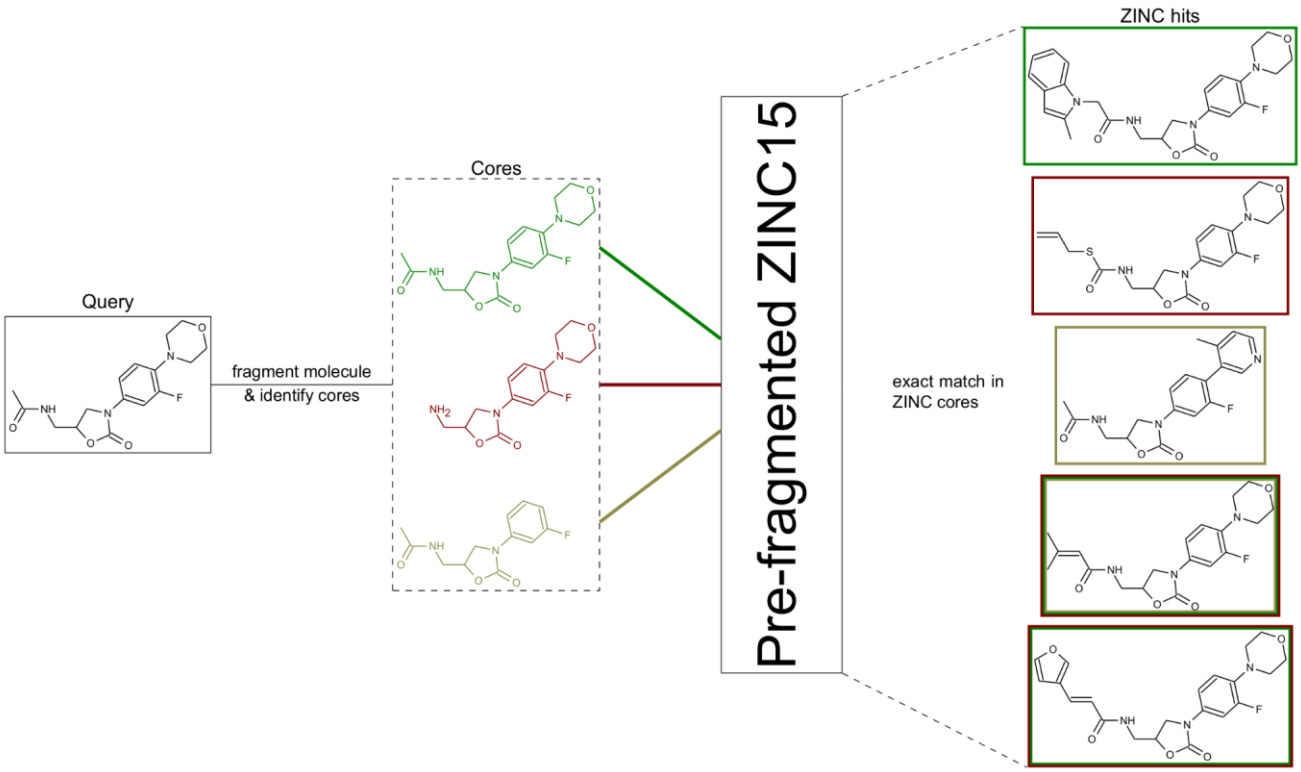


Figure 4. Virtual screening of analog series (ViSAS) concept. In this example, one query molecule is fragmented through RECAP rules, and only fragments retaining at least two-thirds of the heavy atoms in the query are considered cores. The cores are then used for searching for exact matches in the precomputed cores of the ZINC database. This allows searching for chemical analogs in ultra-large libraries (in this case, >740 million unique molecules). For each core, an R-group table with the matching compounds can be computed.

We fragmented ZINC15 (an ultra-large database of commercial compounds containing >740M unique chemical structures) to prepare it for virtual screening. Although fragmentation was time-consuming (about one month even with parallel processing), fragment-and-index approaches require fragmenting each molecule only once. This implies that updates would be faster, as only new molecules have to be processed and added to the dictionary. Any new molecule that is processed undergoes a standard washing procedure consisting of salt removal, extraction of the largest fragment, charge neutralization, and removal of stereochemistry information. Afterward, the washed molecule is searched in the list of processed SMILES, to avoid processing a compound twice. This list maps every unique washed SMILES to the identifiers - IDs - of the compounds mapping to it after the washing procedure. Any new SMILES are fragmented as described in [109, 117]. The fragmentation procedure is easy to run in parallel, as every molecule can be processed independently. We provide bash scripts for downloading and processing ZINC in https://github.com/navejaromero/analog-series. Also, the post-fragmentation ZINC library can be downloaded from Zenodo (10.5281/zenodo.6562818).

To further show the application of the analog series in virtual screening, in the next section we discuss a case study using public data to address a global health issue.

4.2.1 ViSAS on an antituberculosis chemical dataset

The process of hit expansion itself using the processed ZINC database has been recently described by Madariaga-Mazón et al. [118] In this case study, for the purpose of illustrating ViSAS in a real-life example, we used 118 recently published antituberculosis compounds as queries.[119] The fragmenting procedure identified 261 cores, which were then used for the text search in the preprocessed ZINC database. 3091 computational hits were identified. 67 cores matched at least one molecule in ZINC, however, only seven minimally-overlapping cores resulted in more than two hits (Figure 5). Note that this method only finds analogs by matching cores; it is not designed to directly add more cores to the chemical space, but only to enrich those that are already represented in the queries. Nonetheless, the results might be fragmented again and used for another round of virtual screening: this would increase the coverage of the chemical space at the cost of adding more diverse analogs. In this example, the total size of the database increased ~26-fold. However, a significant number of hits could be found only for a few cores, whose SAR could be characterized. For instance, the most represented core had over a thousand hits and two substitution sites. A selection of these hits might be readily acquired and tested experimentally (Figure 6).

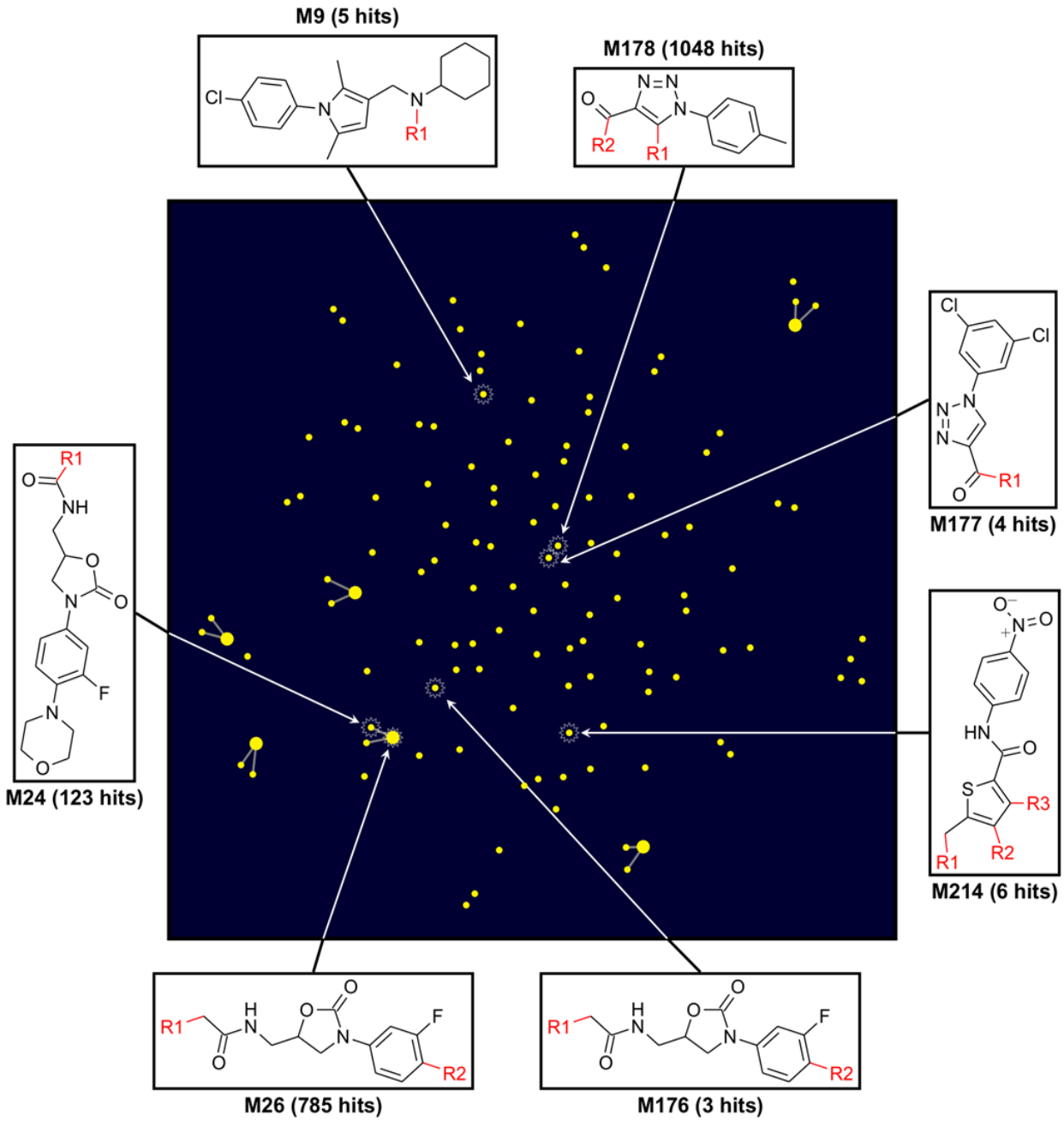
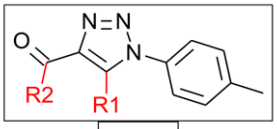


Figure 5. Constellation plot depicting the cores chemical space of a collection of 118 molecules with antituberculosis activity from the cores viewpoint. Every dot represents a valid retrosynthetic core. Larger points represent cores to which two molecules are mapped. Six complex analog series were found, forming constellations in the original data set. ZINC15 was searched for analogs of any of the cores, successfully finding more than a single molecule for seven of them (structures shown and dots highlighted with a clear halo). For simplicity, only 124 cores summarizing the whole core space are plotted; these were selected for minimal overlapping as described in [109].



M178

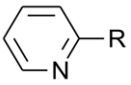
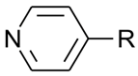
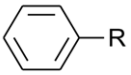
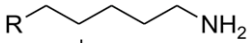
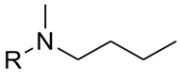
ID	R1	R2	Price (USD)
ZINC000065288225		[R]OH	$8.00
ZINC000011730479		[R]OH	$6.30
ZINC000011730472		[R]OH	$8.30
ZINC000013355576	[R]H		$1.79
ZINC000065036590	[R]H		$87.00
ZINC000040871814	[R]H	[R]OC	$7.80

Figure 6. R-group table showing a selection of the 1048 analogs matching **M178**, the most populated core from the antituberculosis collection in [119] matching the processed ZINC database. Prices as of May 2022, according to the ZINC Express website.[120]

4.3 *De novo* design libraries

Automated *de novo* design and virtual screening represent the *in silico* methods for chemical synthesis of new molecules and high throughput screening. The main goal of *de novo* design is the proposal of novel chemical entities. This computational approach considers a series of constraints to construct new molecular structures. These constraints could include: desired biological effect (primary constraint), drug-likeness, pharmacokinetic properties, toxicity or chemical feasibility (secondary constraints). The desired biological effect is considered the primary constraint for the reason that all programs contemplate this objective.[121] In addition, *de novo* design software has to address three tasks: the assembly of the new molecules, molecule scoring and optimization of the molecules.[122]

More than forty algorithms for *de novo* design have been published since the early 1990s. Taking into account the available information, structure-based or ligand-based approaches can be selected. The three-dimensional coordinates of the receptor are fundamental for the first and active binders for the second.[121] A recurrent ligand-based strategy is the definition of a pharmacophore model from an ensemble of known actives. PhDD [123] is a pharmacophore-based *de novo* design method, it incorporates the assessment of synthetic accessibility and the bioactivity of the proposed molecules is estimated with a fit value to the pharmacophore model. Another example of ligand-based *de novo* design is the reaction-based software DOGS,[[124] this program recommends a synthetic route for each compound. The scoring function of DOGS calculates the similarity of the new molecules with a known bioactive reference. The molecule scoring is another stage where the knowledge of active modulators can be exploited for automated *de novo* design.

Most recent software considers fragments above atoms as building blocks. Fragments' databases come from different sources including external inputs as catalogs or fragment-like compounds. Databases can also be constructed from virtual fragmentation of complete drug molecules to increase the probability of obtaining a drug-like molecule with synthetic accessibility.[125] This last strategy can be exploited not only with complete drugs but also with known active compounds against our target of interest.To follow the strategy of using bioactivity data the selected program has to admit this type of focused fragments, examples of softwares with this characteristic are LUDI (available with Discovery Studio) [126] and LigBuilder.[127]

It is possible that the software incorporates the fragmentation step from complete molecules entered by the user, like alvaBuilder (Alvascience, alvaBuilder (software for de novo molecular design) version 1.0.6, 2021, https://www.alvascience.com). In case the virtual fragmentation has to be made before entering the information to the program it is necessary to construct a database of active modulators. The threshold value to separate actives and inactives can be established in 10 **μ**M as suggested by previous studies.[128] Another strategy is to analyze the bioactivity data of the selected compounds for the target of interest, for example the median can be calculated to set a different limit to the particular database. Once the active set is ready, the molecular fragmentation can be done with algorithms like RECAP [110] and the resulting fragments with the suitable properties are selected as input.

To deal with the tasks of molecular generation and the increasing amount of available bioactivity data, artificial intelligence has been applied to automated *de novo* design. Taking into account the scoring of molecules, ML approaches like target prediction that classifies compounds into active and inactive or quantitative structure-activity relationships (QSAR) could be applied [129] Inverse QSAR or inverse quantitative structure-property relationships (QSPR) are also related with *de novo* design. These methodologies seek to correlate desired properties, including biological activity, to molecular structural feature. [130]

Research work from 2018 proposed an approach based on a generative model that made use of a recurrent neural network for *de novo* drug design. The model was trained with a large molecular set from the ChEMBL database, which annotates biological activity data and chemical structures. With this training the model learned the grammar of SMILES, the chosen molecular representation for the molecules. To generate focused libraries the model was fine-tuned with active modulators of a specific target. This was another strategy that took advantage of bioactivity data to generate novel molecules.[129]

4.3.1 Case study: DNMT focused libraries

This case study is centered in the discovery of new hits against DNMT. Automated *de novo* design was employed for the proposal of compounds. We hypothesized that *de novo* design could lead to molecules that expand the epigenetic-relevant chemical space, due to the expected novelty of the compounds. We selected alvaBuilder (Alvascience, alvaBuilder (software for de novo molecular design) version 1.0.6, 2021, https://www.alvascience.com), a ligand-based *de novo* design program, to construct the molecules. This software selects fragments as construction blocks and incorporates a genetic algorithm to optimize the search

of suitable compounds. To create the fragments and linkers the user has to select a database of entire molecules. The user enters data to create the training set, the molecules are then fragmented into ring systems, linkers, and lateral chains. This approach allows taking advantage of bioactivity data since active modulators could constitute the training set.

For the selection of the training set, we constructed a database of DNMT1 inhibitors with an IC_{50} of 10 **μ**M or less. Bioactivity data was obtained from ChEMBL database version 29 (2021). Since it has been previously observed that bioactivity data could differ between experimental techniques,[131] we only maintained 422 molecules with IC_{50} values (48% of those with annotated bioactivity). After data curation, we had 259 unique compounds with the desired inhibitory concentration.

The scoring function is also customized by the user, the score is a conglomeration of a set of rules. The first rule finds molecules that have a target value for the selected descriptor, from the 91 available. The second rule calculates the similarity to a reference, and the third evaluates if the molecules contain or not a molecular pattern. The result of the scoring aggregates with either arithmetic or geometric mean and exhibits values from zero (worst) to one (best).

To establish the scoring function, we selected seven descriptors: molecular weight, donor atoms for H-bonds, acceptor atoms for H-bonds, consensus LogP model, LogS aqueous solubility, synthetic accessibility score (SAscore) and topological polar surface area (TPSA). Synthetic accessibility is one of the major concerns about *de novo* design. Therefore, we included this quantitative estimation in addition to other physicochemical properties.

We calculated the descriptors of the active molecules with alvaDesc (Alvascience, alvaDesc (software for molecular descriptors calculation) version 2.0.10, 2021, https://www.alvascience.com). This program has the same algorithms for the computation of descriptors as alvaBuilder. With this information, we set up the donor atoms for H-bonds to be ≥ 2 and the SAscore to ≤ 5.979. For the rest of the descriptors, the range was designated to the mean ± the standard deviation of the calculated numerical values for the active inhibitors. The final score was aggregated with the arithmetic mean of the selected rules. We defined a population size of 65 and a maximum number of iterations of 100 for the genetic algorithm. With the same training set and scoring function we obtained 10 sets of new molecules.

With the ten different sets, we computed similarity matrices with PUMA server [132]. The results confirmed that predicted physicochemical properties are highly similar, with Tanimoto coefficients between 0.969 and 0.983. The similarity results were expected due to the definition of the scoring function. Since we confirmed that molecular properties were alike, we also wanted to compute structural similarity between the compounds.

We calculated two different fingerprints: MACCS keys (166-bits) and Morgan radius 2 with RDKit node for KNIME. Preliminary results showed that the similarity inter and intraset is lower than the calculated with molecular properties. Cumulative distribution functions computed with PUMA showed median similarity values from 0.471 to 0.590 with MACCS keys and 0.114 - 0.149 with Extended Connectivity Fingerprints radius 4, both results present interset similarity. Overall, the results showed that new molecules exhibit highly similar properties. These molecular properties were established as secondary constraints by the scoring function.

Nevertheless, the sets exhibit less structural similarity according to the selected fingerprints. The calculated structural diversity is expected for a *de novo* design. In this case, it could also be influenced by the initial diversity of the training set. This is encouraging due to the probability that the desired bioactivity could also be transferred to the novel molecules.

5. Extended similarity methods

Binary similarity indices [133] and metrics are core elements of the machinery used to explore chemical space, classify molecules, design new drugs, and screen molecular libraries in search for promising compounds. However, as well-studied and ubiquitous as they are, these indices have a fundamental drawback, given by the fact that they can only compare two molecules at a time. This means that if we want to estimate the similarity of N compounds, we will need $O(N^2)$ operations, which greatly limits the scaling of these algorithms and restricts them to narrow sections of chemical space. Motivated by these issues, a new family of similarity indices [55, 56] (extended or *n*-ary similarity indices) was recently proposed, that can compare multiple molecules at the same time. In this section we briefly review the characteristics of these indices, and some exemplary applications.

5.1 The extended similarity framework

The defining characteristic of the extended similarity indices [55, 56] is that they are capable of comparing any number of molecules at the same time. Remarkably, the procedure leading to this generalization is extremely simple, which contributes to the ease of implementation of these indices. The starting point is having all the molecules that are going to be analyzed in a suitable representation. For now, all the cheminformatic-related applications of the *n*-ary indices have mostly relied on binary fingerprints (e.g., any type of fingerprint, including MACCS keys, RDKit fingerprints, and circular or Morgan fingerprints), but there has been extensive work on generalizing the domain of definition of these indices, so one could also use arbitrary sequence representations,[134] latent-space descriptor-based approaches,[135] or even coordinate-based 3D representations.[136] The key is that all the molecules will be encoded by equal-length “vectors”. Then, we just need to form a cumulative vector, σ, with components σ_k equal to the sum of each of the components of the molecules to be analyzed (if the molecular representations are aligned in a matrix-like array, then σ will correspond to the sums of the columns of this matrix). This is the most time-consuming step in the calculation of the *n*-ary indices, however, it is easy to see that it will scale as $O(N)$, so it is dramatically more efficient than using any binary comparison (e.g., recent benchmarks have been able to compare tens of billions of molecules using extended indices in a regular laptop). The next step is to define a coincidence threshold, γ, that is going to be used to determine which of the components of σ are going to be identified as corresponding to similarity or dissimilarity descriptors. This classification is extremely easy to perform, since we just need to notice that $|2*\sigma_k - n| > \gamma$ indicates that σ_k will be a similarity, while $|2*\sigma_k - n| \leq \gamma$ indicates dissimilarity (with n being the number of molecules to be compared). We can gain even more detail into the nature of the similarity descriptors: if $2*\sigma_k - n > \gamma$, the similarity will be given by the dominant contribution of “on” bits in the fingerprints

(they mostly share the presence of the same feature), but if $n - 2*\sigma_k > \gamma$, then the similarity will be given by the preponderance of "off" bits in the fingerprints (the given feature is mostly absent from the molecules). Of course, this means that unless we select $\gamma = n - 1$ (which is an extreme choice), we are going to consider similar descriptors that do not necessarily correspond to a perfect coincidence of the features. To properly account for this, we have to introduce weight functions in the formalism, which will penalize these partial coincidences. Then, with all these ingredients, we just need to substitute the (weighted) 1- and 0-similarity, together with the dissimilarity descriptors in the same expressions used to define the binary similarity indices, and we have their corresponding extended version.

These new indices give more freedom at the time of performing any similarity-based analysis, because now we can study correlations between an arbitrary number of molecules.[56] Reassuringly, it has been already shown that they are internally and externally consistent with respect to the newly introduced hyper-parameter γ.[133, 137, 138] The former implies that they will rank multiple datasets in the same way, largely independently of the value of γ. The latter reflects the fact that the ranking obtained from extended indices and the ranking obtained from standard binary indices will also be the same over most γ values.

5.2 Global description: chemical diversity, Chemical Library Networks, and clustering

The ability of the extended similarity indices of requiring O(*N*) operations to compare *N* objects makes them immediately attractive to explore large sections of chemical space, potentially dealing with numbers of molecules out of reach for current approaches. The first direct application related to this is quantifying the chemical diversity of large molecular libraries.[84] The key insight here is that while raw similarity values can be hard to interpret (except in the trivial cases when they are 0 or 1), we can easily determine the relative degree of diversity with respect to a given reference dataset. This makes it easier to readily interpret what one means when saying that libraries are more or less diverse. Several detailed benchmarks showed that the combination of RDKit fingerprints and extended Tanimoto index provides the most robust measure of chemical diversity.

The hyper-efficiency of the *n*-ary indices motivated their use in representations of chemical space spanning millions of molecules. The inspiration here came from Maggiora and Bajorath's Chemical Space Networks (CSNs). The CSNs start from a set of molecules and use these molecules as nodes/vertices in a graph. Then, one decides whether to connect these nodes with edges depending on the (binary) similarity between the involved molecules. As powerful and intuitive as this process can be, it has the same problem as other binary-based approaches, since it demands $O(N^2)$ operations. The extended similarity indices, on the other hand, can be used to define Chemical Library Networks (CLNs),[84] which borrow inspiration from the CSNs, but now the nodes of the graph correspond to complete libraries, and the edges are associated to the extended similarity resulting from comparing any two such datasets. Preliminary studies have applied this methodology to representing chemical spaces with more than 18 million molecules, which is several orders of magnitude more than what has been represented using CSNs. This provides an unprecedented opportunity to map extremely large sections of chemical space, study how the connectivity patterns depend on factors

like the molecular representation, and potentially see how the relations between large libraries evolve in a dynamic when elements are added or removed to them. This versatility strongly points out to the potential of CLNs to be used in polypharmacology and drug repurposing.

The structure of the extended similarity indices naturally leads to a new way of performing hierarchical clustering.[56, 139] Notice that if in any given moment (e.g., k^{th} iteration) we have clusters $c_1(k)$, $c_2(k)$, …, $c_N(k)$, we can proceed to the $(k + 1)^{th}$ iteration by combining the two clusters that maximize their joint extended similarity. In other words, we have a new linkage criterion. While the scaling of this algorithm is the same as those relying on standard linkage criteria like single, average, complete, etc., the *n*-ary clustering has two key advantages. On one hand, this new clustering algorithm has proven to be more robust than current methods, as quantified by the V-measure.[139] Moreover, with the extended clustering, we can provide a very convenient estimate of the number of clusters in the data, without any extra computational cost. Recent studies have shown that this new method is capable of readily classifying various JAK inhibitors derived from different scaffolds.[56]

5.3 Local description: diversity selection and medoid calculation

A less obvious advantage of the *n*-ary indices is their ability to shed light into the local structure of large chemical spaces by singling out a molecule, or sampling a few species. The first of such applications that was explored in detail was the use of extended indices in diversity selection [56] (e.g., selecting a maximally diverse subset from a given library). There are several ways to do this using binary comparisons (like the MaxMin and MaxSum algorithms), but they scale as $O(N^2)$. On the other hand, with the extended similarity indices one can just directly maximize the diversity of the selected set by minimizing the extended similarity of the molecules that are going to be picked (e.g., the Max_nDis [56] or ECS_MeDiv [136] algorithms). This simple procedure scales linearly, so it is in perfect position to handle very large datasets. Moreover, several benchmarks have shown that the Max_nDis algorithm can result in sets that are more than 3 times more diverse than those selected by MaxMin or MaxSum.

Perhaps even more surprisingly, the extended indices allow us to find the most representative elements of a set with great ease.[139] This is a key task in several fields, known as the medoid problem. However, the usual solutions either scale as $O(N^2)$, or use several approximations and stochastic tools to get down to $O(N\log N)$. The main insight to approach this problem using extended indices is to introduce the concept of complementary similarity. That is, given an element in the set, the complementary similarity is the extended similarity of all the elements in the set, except for the chosen one. It is clear then that the medoid will correspond to the element in the set that has the lowest possible value of complementary similarity. This simple recipe provides excellent results when applied to the analysis of biological ensembles,[139] and has also been used to explore epigenetic-focused libraries.[140] An enticing possibility of this algorithm is classifying all the molecules in a library depending on how “central” or “outlier” they are, and information that we can use to select either “stars” or “satellites” in order to represent chemical space regions with more detail.

6. Conclusion and outlook

CADD has made clear contributions to identify and advance drug candidates that are now in clinical use. Over the past few decades, and more remarkably in recent years, AI (arguably better called augmented intelligence) altogether with even richer databases is advancing drug research at a tremendous speed. AI has made clear and dramatic advances in SBDD and LBDD. However, the scientific community should employ AI correctly and for the right reasons beyond hype or fashion: ultimately, “fashion vanishes with time but quality (reasoning) is timeless”.

As part of the large data sources currently available to train AI models, large and ultra large compound databases are emerging, many of them being designed with the aid of *de novo* design. Consequently, the chemical space is expanding, boosting the proposal and evolution of novel visual graphical representations. Similarly, the concept of chemical multiverse has recently emerged as an approach to capture alternative chemical spaces of a compound data set given by different molecular representations. In this regard, the constellation plots, which are based on the concept of analog series, are visual representations of chemical multiverses that facilitate SP(A)R analysis, including the analysis of virtual screening campaigns. We would like to emphasize the advantage to look beyond the traditional chemical space typically composed of small organic compounds obtained from medicinal chemistry and computational efforts such as *de novo* design. The community should look into “neglected or underrepresented chemical spaces” such as metal-containing compounds, food chemicals, and other molecules from natural sources.

Systematic computational searches in chemical and biological spaces -virtual screening or target fishing- is a common and useful practice in drug discovery with several documented successful cases. To this end, ViSAS is a general approach that expands bioactive molecules' chemical space by finding analogs in libraries of purchasable compounds. The hits can be arranged in an R-group table, similar to a molecule optimization campaign; however, this is probably a first in virtual screening. Although the ViSAS concept is formalized in this review, its principles and applications have been published and discussed. We anticipate that local SAR analysis with ML can help select the most relevant analog series to test experimentally. Similarly, the ViSAS concept can be extended to the analysis of *de novo* libraries, to focus on novel analog series and identify those that have commercially available precursors.

In order to speed and facilitate the navigation of the large and ultra large compound libraries, the extended or *n*-ary similarity indices were recently proposed. These novel indices have found a broad range of applications such as quantifying the chemical diversity of (large) molecular libraries, the graphical representation of chemical space through Chemical Library Networks, clustering, diversity selection, and identifying the most representative compound of a compound data set (the medoid).

A practical perspective to the continued improvement of CADD and AI, besides the methodological challenges and their application in drug discovery projects, is the convenience of formal training and education at the undergraduate and graduate level. Several practitioners learn and practice CADD “on the fly” as the research needs emerge. Formal training at early ages would prepare better researchers and practitioners of CADD, AI and it will be advantageous for the continued improvement of communication between

multidisciplinary research teams so that experts in CADD can communicate more effectively with medicinal chemists and other drug discovery team members.

Acknowledgments

F.I.S-G and D.L.P.-R are thankful to CONACyT for the granted scholarship number 848061 and 888207, respectively. JJN is grateful to the Alexander von Humboldt Foundation for a postdoctoral scholarship and to CONACYT for the National Researchers Program. Authors thank grant support from the General Direction of Academic Staff Affairs (DGAPA), UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (UNAM-DGAPA-PAPIIT), grants IN201321 and IV200121. R.A.M.-Q. acknowledges support from the University of Florida in the form of a startup grant and a UFII SEED award.

References

1. Lee JW, Maria-Solano MA, Vu TNL, Yoon S, Choi S. Big data and artificial intelligence (AI) methodologies for computer-aided drug design (CADD). Biochem Soc Trans. 2022 Feb 28;50(1):241–52.
2. Sabe VT, Ntombela T, Jhamba LA, Maguire GEM, Govender T, Naicker T, et al. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. Eur J Med Chem. 2021 Nov 15;224:113705.
3. Zhao L, Ciallella HL, Aleksunes LM, Zhu H. Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. Drug Discov Today. 2020 Sep;25(9):1624–38.
4. Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: recent advances and future perspectives. Expert Opin Drug Discov. 2021 Sep;16(9):949–59.
5. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020 May;19(5):353–64.
6. Mak KK, Pichika MR. Artificial intelligence in drug development: present status and future prospects. Drug Discov Today. 2019 Mar;24(3):773–80.
7. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021 Aug;596(7873):590–6.
8. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.
9. Miljković F, Rodríguez-Pérez R, Bajorath J. Impact of Artificial Intelligence on Compound Discovery, Design, and Synthesis. ACS Omega. 2021 Dec 14;6(49):33293–9.
10. Bajorath J. Deep machine learning for computer-aided drug design. Front Drug Discov [Internet]. 2022 Feb 7;2. Available from: https://www.frontiersin.org/articles/10.3389/fddsv.2022.829043/full.
11. Stumpfe D, Dimova D, Bajorath J. Computational Method for the Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. J Med Chem. 2016 Aug 25;59(16):7667–76.

12. Naveja JJ, Vogt M. Automatic Identification of Analogue Series from Large Compound Data Sets: Methods and Applications. Molecules. 2021 Aug 31;26(17):5291.
13. González-Medina M, Jesús Naveja J, Sánchez-Cruz N, Medina-Franco JL. Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. RSC Adv. 2017;7(85):54153–63.
14. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 2019 Jan 8;47(D1):D930–40.
15. Masoudi-Sobhanzadeh Y, Omidi Y, Amanlou M, Masoudi-Nejad A. Drug databases and their contributions to drug repurposing. Genomics. 2020 Mar;112(2):1087–95.
16. Kunimoto R, Bajorath J, Aoki K. From traditional to data-driven medicinal chemistry: A case study. Drug Discov Today. 2022 Aug;27(8):2065–70.
17. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol. 2008 Nov;4(11):682–90.
18. Nogales C, Mamdouh ZM, List M, Kiel C, Casas AI, Schmidt HHHW. Network pharmacology: curing causal mechanisms instead of treating symptoms. Trends Pharmacol Sci. 2022 Feb;43(2):136–50.
19. Jacoby E. Computational chemogenomics. Wiley Interdiscip Rev Comput Mol Sci. 2011 Jan;1(1):57–67.
20. Brown JB. Computational Chemogenomics. New York: Springer New York; 2018.
21. Saldívar-González FI, Lenci E, Trabocchi A, Medina-Franco JL. Exploring the chemical space and the bioactivity profile of lactams: a chemoinformatic study. RSC Adv. 2019 Aug 23;9(46):27105–16.
22. López-López E, Fernández-de Gortari E, Medina-Franco JL. Yes SIR! On the structure–inactivity relationships in drug discovery. Drug Discov Today. 2022 Aug 1;27(8):2353–62.
23. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021 Feb;26(2):511–24.
24. Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. Drug Discov Today. 2021 Apr;26(4):1040–52.
25. Bajorath J. Artificial intelligence in interdisciplinary life science and drug discovery research. Future Sci OA. 2022 Apr;8(4):FSO792.
26. Bajorath J. State-of-the-art of artificial intelligence in medicinal chemistry. Future Sci OA. 2021 Mar 29;7(6):FSO702.
27. Bajorath J, Chávez-Hernández AL, Duran-Frigola M, Gortari EF de, Gasteiger J, López-López E, et al. Chemoinformatics and Artificial Intelligence Colloquium: Progress and Challenges to Develop Bioactive Compounds. ChemRxiv [Internet]. 2022 Aug 9 [cited 2022 Sep 5]; Available from: https://chemrxiv.org/engage/chemrxiv/article-details/62f1a15d42ddf532a9b420af.
28. Definition of augmented intelligence - Gartner information technology glossary [Internet]. Gartner. [cited 2022 Sep 13]. Available from: https://www.gartner.com/en/information-technology/glossary/augmented-

intelligence.
29. Warr WA, Nicklaus MC, Nicolaou CA, Rarey M. Exploration of Ultralarge Compound Collections for Drug Discovery. J Chem Inf Model. 2022 May 9;62(9):2021–34.
30. Medina-Franco JL, López-López E, Andrade E, Ruiz-Azuara L, Frei A, Guan D, et al. Bridging informatics and medicinal inorganic chemistry: Toward a database of metallodrugs and metallodrug candidates. Drug Discov Today. 2022 May;27(5):1420–30.
31. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J Nat Prod. 2020 Mar 27;83(3):770–803.
32. Medina-Franco JL, Saldívar-González FI. Cheminformatics to Characterize Pharmacologically Active Natural Products. Biomolecules. 2020 Nov 17;10(11):1566.
33. Kirchmair J. Molecular Informatics in Natural Products Research. Mol Inform. 2020 Nov;39(11):e2000206.
34. Saldívar-González FI, Aldas-Bulos VD, Medina-Franco JL, Plisson F. Natural product drug discovery in the artificial intelligence era. Chem Sci. 2022 Feb 9;13(6):1526–46.
35. Yongye AB, Waddell J, Medina-Franco JL. Molecular scaffold analysis of natural products databases in the public domain. Chem Biol Drug Des. 2012 Nov;80(5):717–24.
36. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of Open Natural Products database. J Cheminform. 2021 Jan 10;13(1):2.
37. Medina-Franco JL. Towards a unified Latin American Natural Products Database: LANaPD. Future Sci OA. 2020 Jun 19;6(8):FSO468.
38. Gómez-García A, Medina-Franco JL. Progress and impact of Latin American natural product databases. Biomolecules. 2022 Aug 30;12(9):1202.
39. Barenie R, Darrow J, Avorn J, Kesselheim AS. Discovery and Development of Pregabalin (Lyrica): The Role of Public Funding. Neurology. 2021 Oct 26;97(17):e1653–60.
40. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021 Jan;26(1):80–93.
41. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018 Jan 14;9(2):513–30.
42. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing Learned Molecular Representations for Property Prediction. J Chem Inf Model. 2019 Aug 26;59(8):3370–88.
43. Minnich AJ, McLoughlin K, Tse M, Deng J, Weber A, Murad N, et al. AMPL: A Data-Driven Modeling Pipeline for Drug Discovery. J Chem Inf Model. 2020 Apr 27;60(4):1955–68.
44. Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low Data Drug Discovery with One-Shot Learning. ACS Cent Sci. 2017 Apr 26;3(4):283–93.
45. Wang F, Liu D, Wang H, Luo C, Zheng M, Liu H, et al. Computational screening for active compounds targeting protein sequences: methodology and experimental validation. J Chem Inf Model. 2011 Nov 28;51(11):2821–8.

46. Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, et al. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. PLoS One. 2012 May 30;7(5):e37608.
47. Li Z, Li X, Liu X, Fu Z, Xiong Z, Wu X, et al. KinomeX: a web application for predicting kinome-wide polypharmacology effect of small molecules. Bioinformatics. 2019 Dec 15;35(24):5354–6.
48. Amendola G, Cosconati S. PyRMD: A New Fully Automated AI-Powered Ligand-Based Virtual Screening Tool. J Chem Inf Model. 2021 Aug 23;61(8):3835–45.
49. Cyclica launches ligand express [Internet]. Cyclica. [cited 2022 Sep 22]. Available from: https://cyclicarx.com/press-releases/cyclica-launches-ligand-express-a-disruptive-cloud-based-platform-to-revolutionize-drug-discovery/.
50. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. Chem Rev. 2019 Sep 25;119(18):10520–94.
51. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. Front Environ Sci [Internet]. 2016 February 2;3. Available from: https://www.frontiersin.org/articles/10.3389/fenvs.2015.00080
52. Collins KD, Glorius F. A robustness screen for the rapid assessment of chemical reactions. Nat Chem. 2013 Jul;5(7):597–601.
53. Hessler G, Baringhaus KH. Artificial Intelligence in Drug Design. Molecules. 2018 Oct 2;23(10):2520.
54. Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses. Science. 1969 Oct 10;166(3902):178–92.
55. Miranda-Quintana RA, Bajusz D, Rácz A, Héberger K. Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 1: Theory and characteristics†. J Cheminform. 2021 Apr 23;13(1):32.
56. Miranda-Quintana RA, Rácz A, Bajusz D, Héberger K. Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 2: speed, consistency, diversity selection. J Cheminform. 2021 Apr 23;13(1):33.
57. Yoshimori A, Bajorath J. Iterative DeepSARM modeling for compound optimization. Artificial Intelligence in the Life Sciences. 2021 Dec 1;1:100015.
58. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers. 2021 Aug;25(3):1315–60.
59. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021 Aug 20;373(6557):871–6.
60. Smith JS, Isayev O, Roitberg AE. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. Chem Sci. 2017 Mar 28;8(4):3192–203.
61. Zanette C, Bannan CC, Bayly CI, Fass J, Gilson MK, Shirts MR, et al. Toward Learned Chemical Perception of Force Field Typing Rules. J Chem Theory Comput. 2019 Jan 8;15(1):402–23.
62. Noé F, Olsson S, Köhler J, Wu H. Boltzmann generators: Sampling equilibrium states of many-body

systems with deep learning. Science [Internet]. 2019 Sep 6;365(6457). Available from: http://dx.doi.org/10.1126/science.aaw1147.

63. Brewerton SC. The use of protein-ligand interaction fingerprints in docking. Curr Opin Drug Discov Devel. 2008 May;11(3):356–64.
64. Da C, Kireev D. Structural protein-ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. J Chem Inf Model. 2014 Sep 22;54(9):2555–61.
65. Sánchez-Cruz N, Medina-Franco JL, Mestres J, Barril X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. Bioinformatics. 2021 Jun 16;37(10):1376–82.
66. McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, et al. GNINA 1.0: molecular docking with deep learning. J Cheminform. 2021 Jun 9;13(1):43.
67. Méndez-Lucio O, Ahmad M, del Rio-Chanona EA, Wegner JK. A geometric deep learning approach to predict binding conformations of bioactive molecules. Nature Machine Intelligence. 2021 Dec 2;3(12):1033–9.
68. Medina-Franco JL, Martinez-Mayorga K, Fernández-de Gortari E, Kirchmair J, Bajorath J. Rationality over fashion and hype in drug design. F1000Res [Internet]. 2021 May 18;10. Available from: http://dx.doi.org/10.12688/f1000research.52676.1
69. Ruddigkeit L, Blum LC, Reymond JL. Visualization and virtual screening of the chemical universe database GDB-17. J Chem Inf Model. 2013 Jan 28;53(1):56–65.
70. Medina-Franco JL, Chávez-Hernández AL, López-López E, Saldívar-González FI. Chemical Multiverse: An Expanded View of Chemical Space. Mol Inform. 2022 Aug 2;e2200116.
71. Varnek A, Baskin II. Chemoinformatics as a Theoretical Chemistry Discipline. Mol Inform. 2011 Jan 17;30(1):20–32.
72. Maggiora GM. Introduction to Molecular Similarity and Chemical Space. In: Martinez-Mayorga K, Medina-Franco JL, editors. Foodinformatics: Applications of Chemical Information to Food Chemistry. Cham: Springer International Publishing; 2014. p. 1–81.
73. Chuang KV, Gunsalus LM, Keiser MJ. Learning Molecular Representations for Medicinal Chemistry. J Med Chem. 2020 Aug 27;63(16):8705–22.
74. Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. Wiley Interdiscip Rev Comput Mol Sci. 2022 Feb 18;12:e1603.
75. Polinsky A. Chapter 12 - Lead-Likeness and Drug-Likeness. In: Wermuth CG, editor. The Practice of Medicinal Chemistry (Third Edition). New York: Academic Press; 2008. p. 244–54.
76. Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol. 2004 Dec;1(4):337–41.
77. Warr W. Report on an NIH workshop on ultralarge chemistry databases [Internet]. ChemRxiv. 2021 [cited 2022 Sep 12]. Available from: https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/60c75883bdbb89984ea3ada5/original/report-on-an-nih-

workshop-on-ultralarge-chemistry-databases.pdf
78. Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. Nature. 2004 Dec 16;432(7019):855–61.
79. Medina-Franco JL, Naveja JJ, López-López E. Reaching for the bright StARs in chemical space. Drug Discov Today. 2019 Nov;24(11):2162–9.
80. Medina-Franco JL, Sánchez-Cruz N, López-López E, Díaz-Eufracio BI. Progress on open chemoinformatic tools for expanding and exploring the chemical space. J Comput Aided Mol Des. 2021 Jun 18; 36:341–54.
81. Osolodkin DI, Radchenko EV, Orlov AA, Voronkov AE, Palyulin VA, Zefirov NS. Progress in visual representations of chemical space. Expert Opin Drug Discov. 2015 Jun 22;10(9):959–73.
82. Saldívar-González FI, Medina-Franco JL. Approaches for enhancing the analysis of chemical space for drug discovery. Expert Opin Drug Discov. 2022 Jul;17(7):789–98.
83. Wawer M, Lounkine E, Wassermann AM, Bajorath J. Data structures and computational tools for the extraction of SAR information from large compound sets. Drug Discov Today. 2010 Aug;15(15-16):630–9.
84. Dunn TB, Seabra GM, Kim TD, Juárez-Mercado KE, Li C, Medina-Franco JL, et al. Diversity and Chemical Library Networks of Large Data Sets. J Chem Inf Model. 2022 May 9;62(9):2186–201.
85. Everett H. Hugh Everett Theory of the Universal Wavefunction. Thesis, Princeton University; 1957.
86. Ren X, Shi YS, Zhang Y, Liu B, Zhang LH, Peng YB, et al. Novel Consensus Docking Strategy to Improve Ligand Pose Prediction. J Chem Inf Model. 2018 Aug 27;58(8):1662–8.
87. Willett P. Combination of similarity rankings using data fusion. J Chem Inf Model. 2013 Jan 28;53(1):1–10.
88. Medina-Franco JL, Maggiora GM, Giulianotti MA, Pinilla C, Houghten RA. A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. Chem Biol Drug Des. 2007 Nov;70(5):393–412.
89. Medina-Franco JL, Martínez-Mayorga K, Bender A, Marín RM, Giulianotti MA, Pinilla C, et al. Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. J Chem Inf Model. 2009 Feb;49(2):477–91.
90. Naveja JJ, Medina-Franco JL. Finding Constellations in Chemical Space Through Core Analysis. Front Chem. 2019 Jul 16;7:510.
91. Naveja JJ, Medina-Franco JL. Consistent cell-selective analog series as constellation luminaries in chemical space. Mol Inform. 2020 Dec;39(12):e2000061.
92. López-López E, Cerda-García-Rojas CM, Medina-Franco JL. Tubulin Inhibitors: A Chemoinformatic Analysis Using Cell-Based Data. Molecules. 2021 Apr 24;26(9):2483.
93. Muegge I, Oloff S. Advances in virtual screening. Drug Discov Today Technol. 2006 Dec 1;3(4):405–11.
94. Schneider G. Virtual screening: an endless staircase? Nat Rev Drug Discov. 2010 Apr;9(4):273–6.

95. Zhao H. Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. Drug Discov Today. 2007 Feb;12(3-4):149–55.
96. Tanrikulu Y, Krüger B, Proschak E. The holistic integration of virtual screening in drug discovery. Drug Discov Today. 2013 Apr;18(7-8):358–64.
97. Zhu T, Cao S, Su PC, Patel R, Shah D, Chokshi HB, et al. Hit identification and optimization in virtual screening: practical recommendations based on a critical literature analysis. J Med Chem. 2013 Sep 12;56(17):6560–72.
98. Steadman D, Atkinson BN, Zhao Y, Willis NJ, Frew S, Monaghan A, et al. Virtual Screening Directly Identifies New Fragment-Sized Inhibitors of Carboxylesterase Notum with Nanomolar Activity. J Med Chem. 2022 Jan 13;65(1):562–78.
99. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, et al. Ultra-large library docking for discovering new chemotypes. Nature. 2019 Feb;566(7743):224–9.
100. Sadybekov AA, Sadybekov AV, Liu Y, Iliopoulos-Tsoutsouvas C, Huang XP, Pickett J, et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. Nature. 2022 Jan;601(7893):452–9.
101. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015 Jan;71:58–63.
102. Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? Drug Discov Today. 2002 Sep 1;7(17):903–11.
103. Muratov EN, Amaro R, Andrade CH, Brown N, Ekins S, Fourches D, et al. A critical overview of computational approaches employed for COVID-19 drug discovery. Chem Soc Rev. 2021 Aug 21;50(16):9121–51.
104. Medina-Franco. Grand Challenges of Computer-Aided Drug Design: The Road Ahead. Front Drug Des Discovery. [Internet 2021 July 28]. Available from: https://www.frontiersin.org/articles/10.3389/fddsv.2021.728551
105. Liu Z, Singh SB, Zheng Y, Lindblom P, Tice C, Dong C, et al. Discovery of Potent Inhibitors of 11β-Hydroxysteroid Dehydrogenase Type 1 Using a Novel Growth-Based Protocol of in Silico Screening and Optimization in CONTOUR. J Chem Inf Model. 2019 Aug 26;59(8):3422–36.
106. Amendola G, Ettari R, Previti S, Di Chio C, Messere A, Di Maro S, et al. Lead Discovery of SARS-CoV-2 Main Protease Inhibitors through Covalent Docking-Based Virtual Screening. J Chem Inf Model. 2021 Apr 26;61(4):2062–73.
107. Peng Z, Zhao Q, Tian X, Lei T, Xiang R, Chen L, et al. Discovery of Potent and Isoform-selective Histone Deacetylase Inhibitors Using Structure-based Virtual Screening and Biological Evaluation. Mol Inform. 2022 Feb 27;e2100295.
108. Li X, Jiang Q, Yang X. Discovery of Inhibitors for Mycobacterium Tuberculosis Peptide Deformylase Based on Virtual Screening in Silico. Mol Inform. 2022 Mar;41(3):e2100002.
109. Naveja JJ, Pilón-Jiménez BA, Bajorath J, Medina-Franco JL. A general approach for retrosynthetic

molecular core analysis. J Cheminform. 2019 Sep 24;11(1):61.
110. Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. J Chem Inf Comput Sci. 1998 May;38(3):511–22.
111. Wassermann AM, Dimova D, Iyer P, Bajorath J. Advances in computational medicinal chemistry: Matched molecular pair analysis. Drug Dev Res. 2012 Dec;73(8):518–27.
112. Kunimoto R, Dimova D, Bajorath J. Application of a New Scaffold Concept for Computational Target Deconvolution of Chemical Cancer Cell Line Screens. ACS Omega. 2017 Apr 30;2(4):1463–8.
113. Hu H, Bajorath J. Increasing the public activity cliff knowledge base with new categories of activity cliffs. Future Sci OA. 2020 Apr 15;6(5):FSO472.
114. Vogt M, Yonchev D, Bajorath J. Computational Method to Evaluate Progress in Lead Optimization. J Med Chem. 2018 Dec 13;61(23):10895–900.
115. de la Vega de León A, Bajorath J. Matched molecular pairs derived by retrosynthetic fragmentation. Medchemcomm. 2014;5(1):64–7.
116. Dimova D, Stumpfe D, Hu Y, Bajorath J. Analog series-based scaffolds: computational design and exploration of a new type of molecular scaffolds for medicinal chemistry. Future Sci OA. 2016 Dec;2(4):FSO149.
117. Naveja JJ, Vogt M, Stumpfe D, Medina-Franco JL, Bajorath J. Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound-Core Relationship Method. ACS Omega. 2019 Jan 31;4(1):1027–32.
118. Madariaga-Mazón A, Naveja JJ, Medina-Franco JL, Noriega-Colima KO, Martinez-Mayorga K. DiaNat-DB: a molecular database of antidiabetic compounds from medicinal plants. RSC Adv. 2021;11(9):5172–8.
119. Makarov V, Salina E, Reynolds RC, Kyaw Zin PP, Ekins S. Molecule Property Analyses of Active Compounds for Mycobacterium tuberculosis. J Med Chem. 2020 Sep 10;63(17):8917–55.
120. Bobrowski TM, Korn DR, Muratov EN, Tropsha A. ZINC Express: A Virtual Assistant for Purchasing Compounds Annotated in the ZINC Database. J Chem Inf Model. 2021 Mar 22;61(3):1033–6.
121. Hartenfeller M, Schneider G. Enabling future drug discovery by de novo design. Wiley Interdiscip Rev Comput Mol Sci. 2011 Sep;1(5):742–59.
122. Schneider G, Clark DE. Automated De Novo Drug Design: Are We Nearly There Yet? Angew Chem Int Ed Engl. 2019 Aug 5;58(32):10792–803.
123. Huang Q, Li LL, Yang SY. PhDD: a new pharmacophore-based de novo design method of drug-like molecules combined with assessment of synthetic accessibility. J Mol Graph Model. 2010 Jun;28(8):775–87.
124. Hartenfeller M, Zettl H, Walter M, Rupp M, Reisen F, Proschak E, et al. DOGS: reaction-driven de novo design of bioactive compounds. PLoS Comput Biol. 2012 Feb 16;8(2):e1002380.
125. Fischer T, Gazzola S, Riedl R. Approaching Target Selectivity by De Novo Drug Design. Expert Opin

Drug Discov. 2019 Aug;14(8):791–803.
126. Böhm HJ. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J Comput Aided Mol Des. 1992 Feb;6(1):61–78.
127. Yuan Y, Pei J, Lai L. LigBuilder V3: A Multi-Target de novo Drug Design Approach. Front Chem. 2020 Feb 28;8:142.
128. Ertl P. Magic Rings: Navigation in the Ring Chemical Space Guided by the Bioactive Rings. J Chem Inf Model. 2022 May 9;62(9):2164–70.
129. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Cent Sci. 2018 Jan 24;4(1):120–31.
130. Gantzer P, Creton B, Nieto-Draghi C. Inverse-QSPR for de novo design: A review. Mol Inform. 2020 Apr;39(4):e1900087.
131. Guianvarc'h D, Arimondo PB. Challenges in developing novel DNA methyltransferases inhibitors for cancer therapy. Future Med Chem. 2014 Jul;6(11):1237–40.
132. González-Medina M, Medina-Franco JL. Platform for Unified Molecular Analysis: PUMA. J Chem Inf Model. 2017 Aug 28;57(8):1735–40.
133. Miranda-Quintana RA, Cruz-Rodes R, Codorniu-Hernandez E, Batista-Leyva AJ. Formal theory of the comparative relations: its application to the study of quantum similarity and dissimilarity measures and indices. J Math Chem. 2010 May;47(4):1344–65.
134. Bajusz D, Miranda-Quintana RA, Rácz A, Héberger K. Extended many-item similarity indices for sets of nucleotide and protein sequences. Comput Struct Biotechnol J. 2021 Jun 16;19:3628–39.
135. Rácz A, Dunn TB, Bajusz D, Kim TD, Miranda-Quintana RA, Héberger K. Extended continuous similarity indices: theory and application for QSAR descriptor selection. J Comput Aided Mol Des. 2022 Mar;36(3):157–73.
136. Rácz A, Mihalovits LM, Bajusz D, Héberger K, Miranda-Quintana RA. Molecular Dynamics Simulations and Diversity Selection by Extended Continuous Similarity Indices. J Chem Inf Model. 2022 Jul 25;62(14):3415–25.
137. Miranda-Quintana RA, Kim TD, Heidar-Zadeh F, Ayers PW. On the impossibility of unambiguously selecting the best model for fitting data. J Math Chem. 2019 Aug;57(7):1755–69.
138. Miranda-Quintana RA, Bajusz D, Rácz A, Héberger K. Differential Consistency Analysis: Which Similarity Measures can be Applied in Drug Discovery? Mol Inform. 2021 Jul;40(7):e2060017.
139. Chang L, Perez A, Miranda-Quintana RA. Improving the analysis of biological ensembles through extended similarity measures. Phys Chem Chem Phys. 2021 Dec 22;24(1):444–51.
140. Flores-Padilla EA, Juárez-Mercado KE, Naveja JJ, Kim TD, Alain Miranda-Quintana R, Medina-Franco JL. Chemoinformatic Characterization of Synthetic Screening Libraries Focused on Epigenetic Targets. Mol Inform. 2021 Dec 20;e2100285.