Approaches for enhancing the analysis of chemical space for drug discovery

Fernanda I. Saldívar-González, José L.Medina-Franco*

DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, Avenida Universidad 3000, Mexico City 04510, Mexico

Contact author: medinajl@unam.mx; jose.medina.franco@gmail.com

Abstract

Chemical space is a powerful, general, and practical conceptual framework in drug discovery and other areas in chemistry that addresses the diversity of molecules and it has various applications. Moreover, chemical space is a cornerstone of chemoinformatics as a scientific discipline. In response to the increase in the set of chemical compounds in databases, generators of chemical structures, and tools to calculate molecular descriptors, novel approaches to generate visual representations of chemical space in low dimensions are emerging and evolving**.** Such approaches include a wide range of commercial and free applications, software, and open-source methods. Herein, the current state of chemical space in drug design and discovery is reviewed. The topics discussed herein include advances for efficient navigation in chemical space, the use of this concept in assessing the diversity of different data sets, exploring structure-property/activity relationships for one or multiple endpoints, and compound library design. Recent advances in methodologies for generating visual representations of chemical space have been highlighted, thereby emphasizing open-source methods. It is concluded that **q**uantitative and qualitative generation and analysis of chemical space require novel approaches for handling the increasing number of molecules and their information available in chemical databases (including emerging ultra-large libraries). In addition, it is of utmost importance to note that chemical space is a conceptual framework that goes beyond visual representation in low dimensions. However, the graphical representation of chemical space has several practical applications in drug discovery and beyond.

Keywords: artificial intelligence; chemical space; chemoinformatics, data visualization; drug discovery; structure-activity relationships.

List of abbreviations: 2D, two-dimensional; 3D, three-dimensional; AI, artificial intelligence; CLNs, chemical library networks; ECFPs, extended connectivity fingerprints; MACCS, molecular access system; ML, machine learning; MQN, molecular quantum number; NPs, natural products; PCA, principal component analysis; SOM, self-organizing map; SP(A)R, structure-property (activity) relationships; t-SNE, t-distributed stochastic neighbor embedding.

1. Introduction

Chemical space occasionally referred to in the literature as the “chemical universe” [1] is a concept that has become significant in chemoinformatics as an independent theoretical discipline [2]. Chemical space refers to all possible molecules and multi-dimensional conceptual spaces representing their structural and functional properties. In other words, chemical space is a contraction of the "chemical descriptor vector space" defined by the numerical vector D encoding molecular structure and/or property aspects as elements of the descriptor vector D. Therefore, and in contrast to cosmic space, chemical space is not a physical space and is not unique, because anyone is free to customize its vector space based on structural and functional properties. Indeed, structural and functional representation is arguably the most relevant feature in virtually all chemoinformatics or computational studies [3].

Applications of chemical space concept have progressed from drug discovery to other areas in chemistry, including organic synthesis, food chemistry, and material sciences, to name a few examples reviewed in the literature [4–6]. A key distinction between the different types of the systematic representations of chemical spaces in compound datasets lies in the type of properties or descriptors that are used to represent the compounds of interest. For instance, the nature of the descriptors used to represent small organic molecules is typically different from that describing chemicals with applications in material sciences. In some instances, the qualitative concept of chemical space is actively used to guide drug discovery projects; however, developing a consistent method to visually represent chemical space remains elusive because of the challenge in generating a consistent manner of representing chemical structures. A typical method employed in this area includes analyzing the chemical space of metal-containing compounds [7].

Initially, in drug discovery, chemical space concept proved useful to understand and generate knowledge of the pharmacokinetic properties and molecular diversity of biologically relevant compounds [8,9]. As the number of chemical compounds and their information in databases increased, more

sophisticated molecular descriptors and visualization techniques were developed to expand their applications. For instance, explorations of chemical space have considerably improved our comprehension of biology and led to the development of several tools for investigating structure-property and structure-activity relationships (SPR, SAR, and SP(A)R) [10]. In addition, this concept has raised interesting questions regarding the estimated size of chemical space, and has motivated several research groups to enumerate large libraries of virtual compounds [11,12]. Recently, the availability of software libraries and the rise of artificial intelligence (AI) [13] have led to the emergence of several tools that integrate machine learning (ML) methods as versatile tools to design, generate, and visualize the chemical space of small molecules [14].

Most chemoinformatics tools use two discrete procedures to represent chemical space: (i) calculation of molecular descriptors and (ii) projection from descriptor space into a two-dimensional (2D) plane or three-dimensional (3D) volume using one of the several known techniques [15]. The descriptors can be selected from the structure (constitution, configuration, and conformation) or properties (physical, chemical, and biological) of the molecules present. The types of descriptors guide the interpretations and predictions that can be made [16]. Therefore, descriptors based on physicochemical properties have been widely used to encode absorption, distribution, metabolism, and excretion properties that play an important role in determining the characteristics of therapeutic agents, such as absorption, solubility, and permeability through the membrane [17]. Other commonly used molecular representations are fingerprint-based descriptors in which the Molecular Access System (MACCS) Keys [18] and Extended Connectivity Fingerprints (ECFPs) [19] are among the most widely used methods to assess the structural diversity of small organic molecules. To improve the visual representation of chemical space and expand its application to larger compounds such as peptides, oligonucleotides, and complex carbohydrates, Capecchi et al. recently proposed the MAP4 (MinHashed Atom-Pair fingerprint up to four bonds) molecular fingerprint that, in principle, can encode compounds of virtually *any* size [20]. MAP4 combines substructure and atom-pair concepts to capture global and specific characteristics of the molecular size and shape, which are captured by the bond distance information encoded into the MAP4.

To generate graphical representations of chemical space, coordinate- [16] and cell-based [21] approaches have been developed. Recently, molecular networks have been recommended for

addressing the dimensionality problem [22,23]. Because it is complicated to visualize multidimensional spaces, coordinate-based approaches usually rely on dimensionality reduction techniques to transform high-dimensional data into two or three dimensions. Over the past two decades, several research groups have implemented different dimensionality reduction techniques to analyze chemical space. Such advances were extensively reviewed in a previous study [16]. The most common techniques include principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) [24], and self-organizing map (SOM) [25]. Previous studies have discussed the exploration of SPR in the context of chemical spaces [10].

The objective of this manuscript is to review recent advances in methodologies for generating low-dimensional visual representations of chemical spaces. We emphasize on freely available and open-source methods. Despite the concept of chemical space having broad applicability in several areas of chemistry, including in organic and inorganic molecules (for instance, metallodrugs used in drug discovery [7]), this review focuses on the development and applications of chemical space to small organic compounds. It is expected that some of these methods can be extended or adapted to explore chemical space of other types of compounds. Using an analogy with the concept of a *multiverse* in cosmology, regions in the universe detached from one another exhibit distinct properties [26], and the systematic description of different types of chemical compounds with varying properties (metal-containing molecules, larger chemical compounds relevant in polymers, material science, and biochemistry) can be increased to *chemical multiverses*.

2. State-of-the-art applications of chemical space

The concept of chemical space has several practical applications. In this study, we organized the applications into four categories: selection of molecules from existing compound libraries, analysis of molecular diversity, SP(A)R, and library design (i.e., to assist the expansion of the chemical libraries).

2.1. Navigation of chemical space: selection of compounds from existing libraries

The identification of biologically relevant starting points within a vast chemical space is a particularly relevant task in designing compound collections and selecting compounds from existing libraries for computational and/or experimental screening. Although it is not an easy task, it is possible to utilize the

fact that the physicochemical and biological properties of molecules are associated with their molecular structures. This is known as “chemical similarity principle,” which states that if two molecules share similar structures, then they will likely have similar bioactivities. Thus, the distribution of the compounds in chemical space guides the search for compounds with a specific set of properties. The choice of descriptors to define chemical space is crucial, however, it is not unique; different from cosmic space, chemical space is not invariant. Therefore, molecular representation is the cornerstone of chemical space (and basically any other computational approach).

In this context, different cartographic methods have been proposed to efficiently navigate chemical spaces once a set of descriptors has been selected [27]. Most navigation methods involve positioning a reference query molecule and scanning a large database to identify the adjacent molecules, which are molecules with properties significantly similar to those of the reference structures. Notably, the adjacent molecules to the reference compound can be identified using the full set of descriptors that define chemical space, and this can be performed independently of the visualization method to project the full-dimensional space into a 2D/3D graph.

ChemGPS-NP was one of the first chemographic models used to comprehensively describe chemical space of natural products (NPs) using physicochemical properties and has proven to be useful in various applications [28]. ChemGPS-NP is a PCA-based model of physicochemical properties, defined by a training set of carefully selected compounds that act as "satellites" or reference structures with extreme properties. ChemGPS-NP projects or “positions” new molecules into the chemical space by comparing their physicochemical properties with those of the reference structures. Although PCA-based mapping is fast and easy to compute, it omits nonlinear interactions and some map regions are overloaded with data.

Some non-linear algorithms that have been implemented for chemical space visualization are t-SNE [24], and more recently, uniform manifold approximation and projection (UMAP) [29]. These types of algorithms effectively visualize clusters or groups of data points and their relative proximities. Another frequently used method is SOM [30], a grid-based method that has been used to support lead discovery efforts and target prediction. Examples of the latter include SOM-based prediction of drug equivalence relationships [31] and target inference generator [32].

Generative topographic mapping (GTM) represents a probabilistic alternative to SOMs [33]. This approach has been applied to visualize, analyze and model large collections of data sets for drug design and was also successfully used for large-scale SAR scanning [34].

As discussed in the Introduction section, within non-coordinate-based approaches, chemical space networks (CSNs) were proposed by Bajorath et al. to address the problem of dimensionality [22,23]. CSNs transform a multidimensional chemical space into a graph with the nodes representing chemical compounds and edges connecting compounds within a specific similarity boundary. These graphs provide immediate visualization that can be easily interpreted. CSNs can also be adequately characterized and compared using generally applicable statistical measures from network science. However, visualization becomes increasingly difficult as the number of compounds increases. Therefore, this method is not directly designed for diversity analysis. Recently, networks have been used as the basis for developing chemical library networks (CLNs) that can be used to explore the diversity of large and ultralarge molecular libraries. In general, representations of a tree-like nature, such as Tree MAP (TMAP), are more suitable for analyzing and interpreting large datasets [35].

To navigate through chemical and biological spaces more intuitively, several researchers have developed methods that seek to improve the interpretation by representing molecules beyond individual data points. An example is the scaffold tree approach that graphically represents chemical space as a tree, where the leaves represent individual chemical compounds and the intermediate nodes represent scaffolds and sub-scaffolds [36]. These representations allow a more consistent scaffold analysis in an SAR/SPR context and facilitate the identification of analog collections [37]. To facilitate the visualization of large analog series Constellation plots have been proposed (see section 2.3) [38].

2.2. Molecular diversity

In drug design, the concept of chemical similarity (or chemical diversity) has been addressed using different approaches, and its applications are mainly found in ligand-based design, for instance, in identifying bioactive compounds when some active compounds are known. Similarly, chemical similarity/diversity analysis provides useful information for projects that seek to prioritize the selection of potentially active compounds for experimental evaluation. Another application is the profiling and selection of compound collections with chemically diverse structures to increase the probability of

identifying new scaffolds that can lead to specific biological targets [39]. Similarity and diversity analyses have also been integrated into *de novo* design strategies to evaluate the structural and molecular novelty of chemical libraries, which play an important role in fairly comparing generative approaches [40].

Several studies reported thus far focus on the use of chemical space as an approach to assess the diversity of different datasets and explore the relationships between compound collections, from which valuable conclusions or interpretations have been obtained. For instance, the chemical space of natural compounds has been compared with other collections of compounds such as drugs approved for clinical use, synthetic molecules, and food chemicals [41]. In general, NPs are characterized by covering a region of chemical space more extensively than synthetic compounds and approved drugs, and they also populate areas in the chemical space that are generally not synthetically accessible [41–43]. The structural uniqueness and complexity of NPs have encouraged the continued use of these compounds to identify bioactive compounds for further development, optimization, or inspire the synthesis of compounds with unique scaffolds [44,45].

Recent representative molecular diversity studies include the analysis of novel libraries, such as compounds applied in the food industry [5,46], peptides [47], focused libraries [48], *de novo virtual* libraries [49], and commercially available fragments libraries for medicinal chemistry [50]. The results are summarized in Table 1. For these analyses, new molecular representations and visualization techniques were implemented. For instance, the chemical space of food compounds stored in FooDB was analyzed using ChemMaps, an approach based on reference or “satellite” compounds, that is, molecules whose distance (or similarity) to all other molecules in the chemical space yield sufficient information to produce a visual representation of the space [51,52]. In principle, it is possible to generate a 3D visual representation of chemical space using satellite structures.

TMAP was used in the global analysis of the peptide chemical space, whereas MAP4 was employed as the molecular representation of peptides [47]. A similar approach was used to visualize the chemical space of NPs in the public domain [53].

Table 1. Recent and representative studies of chemical space.

Libraries	Molecular representation	Visualization technique	Application	Reference
133 Natural compounds with TAS2R activity.	Molecular fingerprints	t-SNE	Investigation of chemical similarity of bitter food compounds and identification of associations within the chemical space (chemistry-driven and/or receptor-driven).	[46]
40,531 peptides.	MAP4 molecular fingerprint	TMAP	Overview of the established peptide chemical space in the form of an interactive map.	[47]
11 Commercial libraries focused on epigenetic targets (53,443 compounds in total).	RDKit fingerprints	Constellation plots and CLNs	Select compound libraries for further virtual screening or compound acquisition.	[48]
130 Million stratified sample of GDB-13.	42-D MQN	PCA	Rapid inspection of the generated molecules.	[49]
Nearly 24,000 food chemicals stored in FooDB.	Physicochemical properties, molecular complexity, and scaffold content	ChemMaps	Quantification of the diversity and chemical complexity of the chemical compounds stored in FooDB.	[51]

To assist the processes of decision-making and selecting compound libraries for further virtual screening or compound acquisition for high- or medium-throughput screening for epigenetic drug discovery, Flores-Padilla et al. reported a comprehensive analysis of 11 commercial libraries of varying sizes focused on epigenetic targets (with 53,443 compounds in total) [48]. Analysis of the chemical diversity and coverage of chemical space was conducted with Constellation plots based on the chemical core scaffolds and CLNs [54]. The latter is based on structural fingerprints and facilitates the visual representation of the chemical space of compound datasets with a significant number (millions) of compounds in an efficient manner. The analysis highlighted a commercial library with an extensive coverage of chemical space (despite low intra-molecular diversity) and identified compound collections that cover unique regions of the chemical space not populated by other epigenetic-focused libraries.

As previously discussed, diversity analysis of chemical space can be used to evaluate and compare different generative approaches. For instance, Arús-Pous et al. used PCA plots of molecular quantum number (MQN) fingerprints to assess the quality of the training process in generative models [49]. In that study, MQN PCA plots allowed the following up and improvement of the comprehension of the varying architectures of molecular generative models. Another recent and representative example of the use of chemical space to analyze diversity was performed with more than 400,000 purchasable building blocks (PBBs) provided by eMolecules (Zabolotna et al. 2021). Visualization of the chemical space of these PBBs using GTM allowed the identification of the most represented and underrepresented classes of PBBs. The results can be focused to improve PBB libraries in a way that allows efficient synthesis in a relevant medicinal chemistry space.

2.3. Structure-property (activity) relationships

As mentioned in the Introduction section, one of the major practical applications of visual representation of chemical space in drug discovery is SAR analysis [55] where the concept of chemical space provides a solid and consistent framework for representing the structural data. When activity data are added (e.g., mapped) into a visual representation of chemical space, it is possible to navigate through the chemical space and exploring (qualitatively or quantitatively) variations in activity upon changes in chemical structures. The massive amount of data stored in chemical databases, including incomplete chemogenomic data or activity data obtained at single concentrations, makes visualization SAR difficult; however, is can be aided by the power of visualization tools. Previous studies highlighted advances in methodologies that explore SAR of compound data sets and screening collections [10,55].

Recent developments in analyzing SP(A)R include constellation plots. Briefly, constellation plots are 2D graphs that combine the clustering of compound datasets based on chemical scaffolds (in particular, analog series) and the distributions or mutual relationships of analog series based on fingerprint representations. Recently constellation plots were used to analyze the SAR of a large dataset of small molecules tested in a panel of cell lines using high-throughput screening. The authors identified a consistent cell-selective analog series of chemical compounds and proposed statistics to quantify cell promiscuity and consistency [56].

In a separate and recent analysis, constellation plots were used to uncover a promising analog series of inhibitors of tubulin-microtubules. In that study [57], the authors analyzed the SAR of a curated dataset of 851 compounds with anticancer activity targeting tubulin-microtubules. In particular, the constellation plots identified at least six analog series of compounds with high average activity (known as “bright regions” in chemical space). The plot also indicates an analog series with predominantly inactive molecules (“dark regions” in chemical space). In recent developments, constellation plots have been implemented in DataWarrior [58] such that the user can explore the chemical space interactively.

Another recent example of the application of chemical space to SP(A)R analysis lies at the interface of drug discovery and food chemistry [46]. Bayer et al. explored the associations between the chemical structures of 133 compounds with known biological activities and extra-oral bitter taste receptors, which belong to the superfamily of G-protein-coupled receptors. As part of the analysis, the authors represented the chemical space of the compounds using t-SNE as a visualization tool; the compounds were represented using MACCS key fingerprints. It was observed that the visual representation of chemical space grouped chemical compounds with similar functional groups, even though the compounds can belong to different classes (depending on the type of receptors they are related to).

2.4. Compound library design

Over the last few decades, medicinal chemistry has made major breakthroughs in increasing the accessible chemical space, which is estimated to contain approximately 10^{63} molecules [11,59]. In this context, having access to more regions of the chemical space can, in principle, augment the probability of finding something “interesting” and valuable. Thus, algorithms and methods to augment and search these spaces can focus on the generation of new molecules to compounds with desirable properties for drug design or discovery projects. In this regard, it remains to determine the medicinally relevant chemical space as the number of therapeutic targets is evolving [60]. A related challenge is to establish the intersection of chemical space with the biological space. These questions are being addressed by computational chemogenomics and have been noted as one of the major challenges in computer-aided drug design [61].

Computational approaches to facilitate the design of functional molecules include the development of *de novo* algorithms that explore chemical spaces to generate new compounds. For instance, the *de novo*

design algorithm for exploring chemical space scans the space and generates structures in a specific area on a user-selected pane [62]. Similarly, Capecchi et al. developed the peptide design genetic algorithm (PDGA), a computational tool that generates highly-similarity analogs of bioactive peptides with various peptide chain topologies in a chemical space defined by the macromolecule extended atom pair fingerprint [63]. Recently, Aspuru et al. proposed the superfast traversal, optimization, novelty, exploration, and discovery (STONED) algorithm to perform exploration and interpolation in chemical space to obtain novel molecules [64]. STONED uses self-referencing embedded strings [65], a molecular representation that is more suitable for ML. This algorithm reduces the long training times, large datasets, and handcrafted rules.

In general, deep generative models can operate over large spaces of molecular structures and embed the chemical properties of these structures into a vector space. These models can generate new and previously unidentified chemical compounds by decoding from this 'latent' space of chemical structures. Recent reviews of the *de novo* design have examined progress in generative model architecture and evaluated their efficiency with reference to experimentally validated test cases in the literature [66–68].

3. Novel approaches

3.1. Meta-analysis of applications of chemical space

To understand the evolution of the concept of chemical space and its applications, a meta-analysis of the literature has been performed using the search terms "chemical space" and "drug design" in PubMed (https://pubmed.ncbi.nlm.nih.gov/). In total, the search yielded 1538 articles (November 2021) that were analyzed using VOSviewer [69]. The results of the meta-analysis indicated that the main concurrent terms associated with the keywords used were SAR analysis and small-molecule library design (Figure 1a). Visualization of chemical space has been used frequently to support the analysis of antineoplastic agents (76 articles), protein kinase inhibitors (60), antibacterials (51), antimalarials (28), and antiviral compounds (21). Similarly, a notable number of articles related to the concept of chemical space are associated with drug repurposing (20). In particular, using network-based representations to predict drug-

target interactions and more complex interactions, including drug-disease, protein-disease, and drug-side effect associations, to name a few [70].

According to the author's keywords (Figure 1b), the most recent articles (see color scale) are focused on ML methods such as deep learning. It is also highlighted that the concept of chemical space has had recent applications in Alzheimer's disease and in emerging diseases such as COVID-19. Particularly for COVID-19, chemical space visualization proved to be a fast way to analyze and describe the huge chemical space of known antiviral compounds [71,72]. For instance, GTM is one of the methods used to represent the chemical space of compounds obtained from medicinal chemistry efforts against coronaviruses (CoVs) [72]. In particular, GTMs helped highlight the structural relationship between antivirals of different categories, predict their polypharmacological profiles, and emphasize frequently encountered chemotypes. Similarly, chemical space concept was very helpful in finding attractive compounds for repositioning [73] and guiding the identification of potent and selective scaffolds with anti-COVID activity [74].

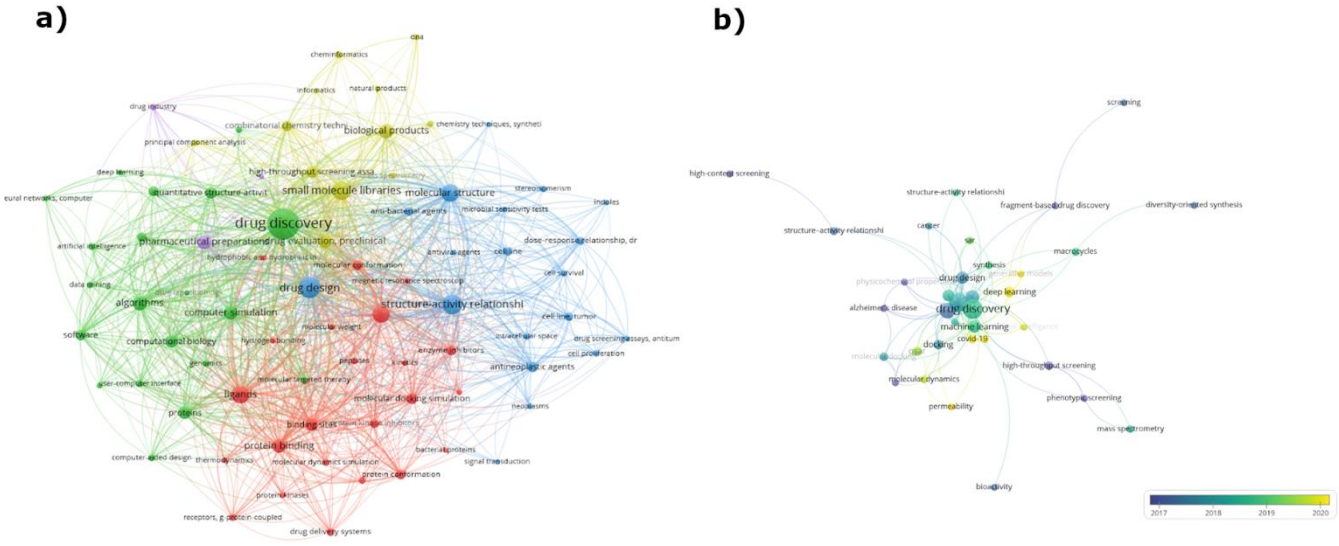

Figure 1. Meta-analysis of the literature: network analysis that reveals the main concurrent terms related with the keywords “chemical space” and “drug discovery” in PubMed (1538 results). The network maps were constructed on VOSviewer [69]. a) Network analysis based on all results (occurrence > 20, number of keywords selected: 90) and b) network analysis based on the keywords of the study (occurrence > 5, number of keywords selected: 35).

Advances in AI and availability of software libraries have resulted in ML methods, such as deep learning and versatile tools for exploring chemical space for drug discovery applications [14]. Table 2

summarizes the novel approaches using ML methods, some of which have been mentioned previously. Recent advances have focused in identifying molecules with desirable properties in large chemical spaces. To this end, genetic algorithms (GAs) [75,76], methods using variational autoencoders (VAEs) [77,78], recurrent neural networks (RNNs) [79,80], and generative antagonistic networks (GANs) [81,82] have been developed. In several instances, these algorithms are associated with the generation of new molecules and have exhibited the ability to traverse chemical space more effectively, reaching optimal chemical solutions while considering fewer molecules than allowed by the brute-force screening of large chemical libraries. Similarly, several evolutionary and RNN selection mechanisms have proven successful in multi-objective optimization problems [83,84]. Emerging approaches in chemical enumeration incorporate chemical reactions into ML-based generation to design novel compounds in a synthetically accessible chemical space [85].

As mentioned, similarity-based compound networks such as CSNs allow the visualization of SAR patterns. To increase the number of practical applications of network-based chemical space representations and decrease biases in ML, it is necessary to incorporate amounts of data from chemical interactomes. In this regard, it is also necessary to improve network visualization to obtain reasonable representations of networks containing thousands of nodes. Addressing these difficulties will be useful in SAR analysis and drug repurposing.

Another application of the field of neural networks has been to solve address problems related to big data and visual representation of datasets with a large number of compounds [54]. It is anticipated that more researchers will integrate ML methods to speed up chemical space analysis and realize more efficient outcomes.

Table 2. Novel approaches using machine learning methods.

Application	Novel approaches using ML methods	Reference(s)
Navigation of chemical space: Selection of compounds from existing libraries	• GAs, VAEs, RNN and GAN as search algorithms	[71-78]
Chemical diversity	• Dimensionality reduction methods (t-SNE, TMAP, and GTM) • Chemical library networks (CLN)	[24,34,35] [54]
Structure-property/activity relationships	• Chemical space networks (CSNs)	[22,23]

	• Constellation plots	[38]
Design novel compound libraries	• Chemical reactions in ML-based generation • Multi-objective optimization algorithms	[12,85] [83,84]

3.2. Novel implementations for visualization

The interactive visualization of 2D and 3D representations of chemical spaces, in particular of large and ultra-large data sets, has been an active area of investigation. The interactive visualization of chemical space was performed using an open-source code and is freely available on websites.

Web servers available in the public domain for enabling interactive visualization of chemical spaces have been reviewed recently [10]. This review includes classical and early developments such as Chem-GPS (*vide supra*), a significant set of public tools developed by Reymond et al. such as Ferun, and PDB Explorer. In the past few months, progress has been made in the interactive analysis of chemical space. A notable example is the “magic rings” developed by Ertl: a freely available web page with an interactive clustering of rings and Bemis-Murcko scaffolds present in compounds in ChEMBL [86] (28 release) with a biological activity value of 10 microM [87]. The interactive clustering available at https://bit.ly/magicrings enables users to quickly identify the main substructures of the major target classes of relevance in drug discovery.

Another recent development is the NP navigator [88], which further bridges the application of cheminformatics in NP research [89,90]. The NP navigator, publicly available at https://infochm.chimie.unistra.fr/npnav/chematlas_userspace/, is an implementation of the visualization algorithm GTM maps that explores interactively the chemical space of COCONUT (Collection of Open Natural Products database) [91] (currently the largest collection of NPs in the public domain), bioactive molecules in ChEMBL, and purchasable compounds from the ZINC database [92]. Interactive navigation can be used to explore chemical compounds based on different representations such as physicochemical properties, scaffold distribution, commercial availability, and biological activity.

In a recent study, Chávez-Hernández et al. implemented an interactive visualization of the chemical space of a newly generated library of HIV-1 viral protease inhibitors assembled from NP fragments. Visual representation of the chemical space was based on TMAPs [20] and molecular fingerprints. The

interactive representation of the chemical space enables the user to navigate through a synthetic compound library of pseudo-NPs [93] designed *de novo*.

Figure 2 illustrates examples of visual representations of chemical space using freely available online resources. Figure 2a shows a TMAP with antidiabetic compounds of different origins: 38 approved drugs from DrugBank, 337 antidiabetic compounds from medicinal plants (DiaNAT DB) [94], 201 compounds from ChEMBL with experimental evaluation of DMT2, and 20 compounds designed and synthesized by Navarrete-Vazquez et al. [95–99]. It is observed that, structurally, the compounds tend to group according to the database to which they belong, with the DiaNAT database being the most diverse. In this graph, it is also possible to identify antidiabetic approved drugs and compounds from DiaNAT and ChEMBL, similar to those designed by Navarrete-Vazquez et al. This supports the molecular design employed and, in turn, may help expand this collection of compounds. Figure 2b was obtained from the NP navigator and projects NPs from DiaNAT (right) and the compounds synthesized by Navarrete-Vazquez et al. (left) onto a comparative landscape, where the black colored background of the map corresponds to the library (libraries) that were selected as a basis of the landscape (black regions correspond to the NPs, red regions correspond to the NP-like ZINC compounds, and white areas correspond to the empty regions of chemical space). The NP-Umap illustrates the preferential locations of the compounds in the DiaNAT database in regions corresponding to NPs and the antidiabetic compounds from synthesis in regions of ZINC compounds. Furthermore, it is possible to perform further structural analyses based on the maximum common substructures (MCS) to identify NP and NP-like analogs of selected compounds.

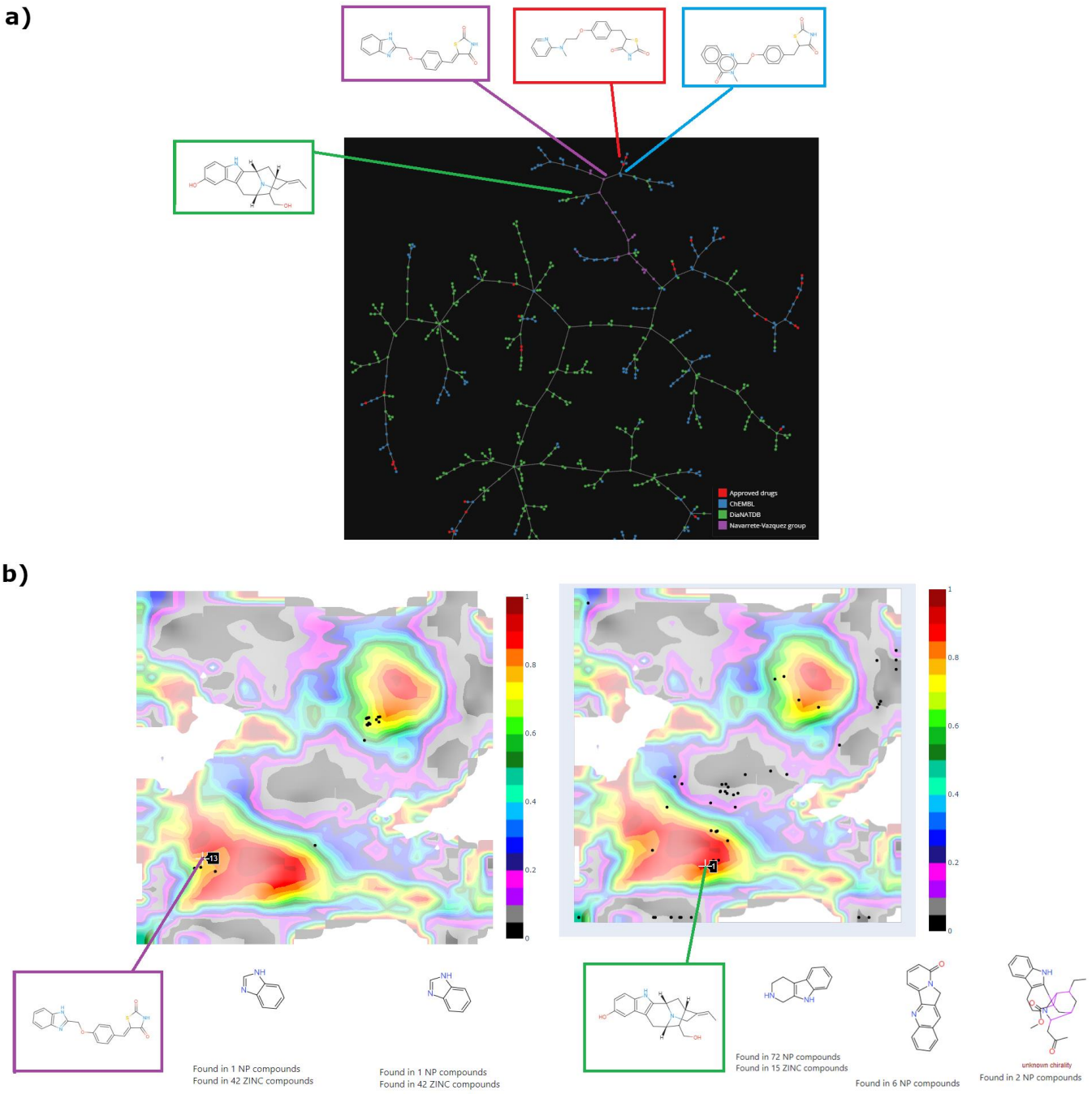

Figure 2. Examples of visual representations of chemical space using freely available online resources. a) Chemical space visualization of antidiabetic compounds using TMAP based on molecular fingerprints. Compound libraries represented in colors: FDA-approved drugs (red), DiaNAT DB (green), compounds from ChEMBL (blue), and compounds designed and synthesized by Navarrete-Vazquez et al. (purple). b) NPs from DiaNAT (right) and compounds synthesized in Navarrete-Vazquez's group (left) projected onto class landscape comparing COCONUT natural products (black) with NP-like ZINC compounds (red). Maximum common substructures (MCS) found in COCONUT and NP-like ZINC of selected molecules are illustrated.

Expert opinion

Chemical space is a core concept in chemoinformatics with several practical applications in drug discovery and other areas in chemistry. Typically, chemical space is used for selecting specific sets of compounds for further computational or experimental screening, diversity, and SP(A)R analysis, and to guide the design of novel molecules. The latter application is intended such that the newly generated compounds are at the intersection of the biologically relevant chemical space. In any application, compound representation is a key variable in qualitative or quantitative chemical space analysis (including visual representation); it has to be in line with the objective of the study as it will guide the interpretation of the analysis.

Currently, ML methodologies continue to open new possibilities for generating hundreds and thousands of new molecules from an exhaustive search in chemical space. To perform the search in the chemical space faster and more efficiently, in particular for large data sets, the visualization methods should scale well with the number of molecules (“haystack size”); find the most relevant compounds (e.g., find the “needle,” irrespective of the size of the haystack); and be affordable to run on standard hardware.

In recent years, with a significant an increasing number of molecules to be analyzed, novel methods to generate visual representations of chemical space have been developed. While interpreting such visualizations, one should consider that they are approximations and that the “true” chemical space is defined by the complete set of descriptors used. Because it is challenging to select the appropriate method according to the expected qualities of the visualization, it is advisable to complement the visual (e.g., qualitative) analysis of chemical space with a quantitative analysis considering the entire multidimensional space. In this regard, it is advisable to consider consensus approaches: multiple representations of chemical space (at least more than one), because each visualization will capture part of the “true” chemical space.

As part of the progress in method development, there have been notable developments in the implementation of freely available online resources. In this manner, the user can interactively explore the chemical space of compound datasets.

There are still challenges in exploring the chemical space for drug discovery, such as developing consistent representations of metal-containing compounds. Other challenges include consistently representing the chemical space of non-traditional small- and medium-sized biologically relevant compounds such as peptides, macrocycles, and metal-containing clinical candidates.

Funding

We thank funding support from DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (UNAM-DGAPA-PAPIIT), grant IN201321. FI.S-G is thankful to CONACYT for the granted scholarship number 848061.

Declaration of interest

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

References

[1] Ruddigkeit L, Blum LC, Reymond J-L. Visualization and virtual screening of the chemical universe database GDB-17. J Chem Inf Model. 2013;53:56–65.

[2] Varnek A, Baskin II. Chemoinformatics as a Theoretical Chemistry Discipline. Mol Inform. 2011;30:20–32.

[3] Chuang KV, Gunsalus LM, Keiser MJ. Learning Molecular Representations for Medicinal Chemistry. J Med Chem. 2020;63:8705–8722.

[4] Grygorenko OO, Volochnyuk DM, Ryabukhin SV, et al. The Symbiotic Relationship Between Drug Discovery and Organic Chemistry. Chemistry. 2020;26:1196–1237.

[5] Naveja JJ, Rico-Hidalgo MP, Medina-Franco JL. Analysis of a large food chemical database: chemical space, diversity, and complexity. F1000Res: published online 3 July 2018, http://dx.doi.org/10.12688/f1000research.15440.2.

[6] Pollice R, Dos Passos Gomes G, Aldeghi M, et al. Data-Driven Strategies for Accelerated Materials Design. Acc Chem Res. 2021;54:849–860.

[7] Meggers E. Exploring biologically relevant chemical space with metal complexes. Curr Opin Chem Biol. 2007;11:287–292.

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

[9] Paolini GV, Shapland RHB, van Hoorn WP, et al. Global mapping of pharmacological space. Nat Biotechnol. 2006;24:805–815.

[10] Medina-Franco JL, Sánchez-Cruz N, López-López E, et al. Progress on open chemoinformatic tools for expanding and exploring the chemical space. J Comput Aided Mol Des. 2022; in press. doi: 10.1007/s10822-021-00399-1.

**** Recent review of chemical space focused on open resources to analyze the space.**

[11] Walters WP. Virtual chemical libraries. J Med Chem. 2019;62:1116–1124.

[12] Hoffmann T, Gastreich M. The next level in chemical space navigation: going far beyond enumerable compound libraries. Drug Discov Today. 2019;24:1148–1156.

[13] Miljković F, Rodríguez-Pérez R, Bajorath J. Impact of artificial intelligence on compound discovery, design, and synthesis. ACS Omega. 2021; 49:33293–33299.

**** Recent review of artificial intelligence.**

[14] Arús-Pous J, Awale M, Probst D, et al. Exploring Chemical space with machine learning. Chimia. 2019;73:1018–1023.

[15] Schneider G. Analysis of Chemical Space. Landes Bioscience; 2013.

[16] Osolodkin DI, Radchenko EV, Orlov AA, et al. Progress in visual representations of chemical space. Expert Opin Drug Discov. 2015;10:959–973.

*** Comprehensive expert opinion of approaches to visualize chemical space.**

[17] Cumming JG, Davis AM, Muresan S, et al. Chemical predictive modelling to improve compound quality. Nat Rev Drug Discov. 2013;12:948–962.

[18] Durant JL, Leland BA, Henry DR, et al. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci. 2002;42:1273–1280.

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

[20] Capecchi A, Probst D, Reymond J-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform. 2020;12:43.

[21] Pearlman RS, Smith KM. Novel software tools for chemical diversity. 3D QSAR in Drug Design. Dordrecht: Kluwer Academic Publishers; 2005. p. 339–353.

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

[23] Vogt M, Stumpfe D, Maggiora GM, et al. Lessons learned from the design of chemical space networks and opportunities for new applications. J Comput Aided Mol Des. 2016;30:191–208.

[24] van der Maaten L. Visualizing data using t-SNE. J Mach Learn Res. 2008; 1: 1-48.

[25] Digles D, Ecker GF. Self-organizing maps for in silico screening and data visualization. Mol Inform. 2011;30:838–846.

[26] Kragh H. Contemporary history of cosmology and the controversy over the multiverse. Ann Sci. 2009;66:529–551.

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

[28] Larsson J, Gottfries J, Muresan S, et al. ChemGPS-NP: tuned for navigation in biologically relevant chemical space. J Nat Prod. 2007;70:789–794..

[29] McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv [stat.ML]. 2018. Available at: http://arxiv.org/abs/1802.03426 [Last accessed 20 January 2022]

[30] Kohonen T. Exploration of very large databases by self-organizing maps. Proceedings of International Conference on Neural Networks (ICNN'97). 1997. p. PL1–PL6 vol.1.

[31] Reker D, Rodrigues T, Schneider P, et al. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. Proc Natl Acad Sci U S A. 2014;111:4067–4072.

[32] Schneider P, Schneider G. De-orphaning the marine natural product (±)-marinopyrrole A by computational target prediction and biochemical validation. Chem Commun. 2017;53:2272–2274.

[33] Bishop CM, Svensén M, Williams CKI. GTM: The generative topographic mapping. Neural Comput. 1998;10:215–234.

[34] Kireeva N, Baskin II, Gaspar HA, et al. Generative topographic mapping (GTM): Universal tool for data visualization, structure-activity modeling and dataset comparison. Mol Inform. 2012;31:301–312.

[35] Probst D, Reymond J-L. Visualization of very large high-dimensional data sets as minimum spanning trees. J Cheminform. 2020;12:12.

[36] Schuffenhauer A, Ertl P, Roggo S, et al. The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification. J Chem Inf Model. 2007;47:47–58.

[37] Hu Y, Stumpfe D, Bajorath J. Lessons learned from molecular scaffold analysis. J Chem Inf Model. 2011;51:1742–1753.

[38] Naveja JJ, Medina-Franco JL. Finding constellations in chemical space through core analysis. Front Chem. 2019;7:510.

[39] López-Vallejo F, Giulianotti MA, Houghten RA, et al. Expanding the medicinally relevant chemical space with compound libraries. Drug Discov Today. 2012;17:718–726.

[40] Jiménez-Luna J, Grisoni F, Weskamp N, et al. Artificial intelligence in drug discovery: recent advances and future perspectives. Expert Opin Drug Discov. 2021;16:949–959.

[41] Saldívar-González FI, Pilón-Jiménez BA, Medina-Franco JL. Chemical space of naturally occurring compounds. Physical Sciences Reviews. 2018;0:525.

[42] Ertl P, Schuffenhauer A. Cheminformatics analysis of natural products: lessons from nature inspiring the design of new drugs. Prog Drug Res. 2008;66:217, 219–235.

[43] Rodrigues T. Harnessing the potential of natural products in drug discovery from a cheminformatics vantage point. Org Biomol Chem. 2017;15:9275–9282.

[44] 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;83:770–803.

[45] Atanasov AG, Zotchev SB, Dirsch VM, et al. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov. 2021;20:200–216.

[46] Bayer S, Mayer AI, Borgonovo G, et al. Chemoinformatics view on bitter taste receptor agonists in Food. J Agric Food Chem. 2021;69:13916–13924.

[47] Capecchi A, Reymond J-L. Peptides in chemical space. Medicine in Drug Discovery. 2021;9:100081.

[48] Flores-Padilla A, Juárez-Mercado EK, Naveja JJ, et al. Chemoinformatic characterization of synthetic screening libraries focused on epigenetic targets. Mol Inf. 2022, in press. doi: 10.1002/minf.202100285.

[49] Arús-Pous J, Blaschke T, Ulander S, et al. Exploring the GDB-13 chemical space using deep generative models. J Cheminform. 2019;11:20.

[50] Zabolotna Y, Volochnyuk DM, Ryabukhin SV, et al. A Close-up Look at the Chemical Space of Commercially Available Building Blocks for Medicinal Chemistry. J Chem Inf Model: published online 3 December 2021, http://dx.doi.org/10.1021/acs.jcim.1c00811.

[51] Naveja JJ, Medina-Franco JL. ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds. F1000Research. 2017;6(Chem Inf Sci):1134.

[52] Borrel A, Kleinstreuer NC, Fourches D. Exploring drug space with ChemMaps.com. Bioinformatics. 2018;34:3773–3775.

[53] Capecchi A, Reymond J-L. Assigning the origin of microbial natural products by chemical space map and machine learning. Biomolecules 2020;10:1385.

[54] Dunn TB, Seabra GM, Kim TD, et al. Diversity and chemical library networks of large data sets. J Chem Inf Model. 2021, in press. doi: 10.1021/acs.jcim.1c01013.

[55] Wawer M, Lounkine E, Wassermann AM, et al. Data structures and computational tools for the extraction of SAR information from large compound sets. Drug Discov Today. 2010;15:630–639.

[56] Naveja JJ, Medina-Franco JL. Consistent cell-selective analog series as constellation luminaries in chemical space. Mol Inform. 2020;39:e2000061

[57] López-López E, Cerda-García-Rojas CM, Medina-Franco JL. Tubulin inhibitors: A chemoinformatic analysis using cell-based data. Molecules. 2021;26:2483.

[58] Sander T, Freyss J, von Korff M, et al. DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model. 2015;55:460–473.

[59] Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996;16:3–50

[60] Zdrazil B, Richter L, Brown N, et al. Moving targets in drug discovery. Sci Rep. 2020;10:20213.

[61] Medina-Franco JL. Grand challenges of computer-aided drug design: The road ahead. Front Drug Discov: 2021;1:728551.

[62] Takeda S, Kaneko H, Funatsu K. Chemical-space-based de novo design method to generate drug-like molecules. J Chem Inf Model. 2016;56:1885–1893.

[63] Capecchi A, Zhang A, Reymond J-L. Populating Chemical Space with Peptides Using a Genetic Algorithm. J Chem Inf Model. 2020;60:121–132.

[64] Nigam A, Pollice R, Krenn M, et al. Beyond generative models: Superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. Chem. Sci. 2021;12:7079-7090

[65] Krenn M, Häse F, Nigam A, et al. SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. Mach Learn Sci Technol, 2020;1:045024

[66] Meyers J, Fabian B, Brown N. De novo molecular design and generative models. Drug Discov Today. 2021; 26:2707-2715.

[67] Elton DC, Boukouvalas Z, Fuge MD, et al. Deep learning for molecular design—a review of the state of the art. Mol Syst Des Eng. 2019;4:828–849.

[68] Schneider G, Clark DE. Automated de novo drug design: Are we nearly there yet? Angew Chem Int Ed Engl. 2019;58:10792–10803.

[69] Perianes-Rodriguez A, Waltman L, van Eck NJ. Constructing bibliometric networks: A comparison between full and fractional counting. J Informetr. 2016;10:1178–1195.

[70] Orlov AA, Berishvili VP, Nikitina AA, et al. Chapter 13 - Analysis of Chemical Spaces: Implications for Drug Repurposing. In: Roy K, editor. In Silico Drug Design. Academic Press; 2019. p. 359–395.

71] Kumar A, Loharch S, Kumar S, et al. Exploiting cheminformatic and machine learning to navigate the available chemical space of potential small molecule inhibitors of SARS-CoV-2. Comput Struct Biotechnol J. 2021;19:424–438.

[72] Horvath D, Orlov A, Osolodkin D, et al. A Chemographic audit of anti-coronavirus structure-activity information from public databases (ChEMBL). Mol Inform . 2020; 39:e2000080.

[73] Chakraborti S, Bheemireddy S, Srinivasan N. Repurposing drugs against the main protease of SARS-CoV-2: mechanism-based insights supported by available laboratory and clinical data. Mol Omics. 2020;16:474–491.

[74] Santana MVS, Silva-Jr FP. De novo design and bioactivity prediction of SARS-CoV-2 main protease inhibitors using recurrent neural network-based transfer learning. BMC Chem Biol. 2021;15:8.

[75] Virshup AM, Contreras-García J, Wipf P, et al. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. J Am Chem Soc. 2013;135:7296–7303.

*** Original paper that clearly defines chemical space as a conceptual framework, in particular as a chemical descriptor vector space.**

[76] Henault ES, Rasmussen MH, Jensen JH. Chemical space exploration: how genetic algorithms find the needle in the haystack. PeerJ Phy Chem. 2020;2:e11.

[77] Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4:268–276.

[78] Winter R, Montanari F, Steffen A, et al. Efficient multi-objective molecular optimization in a continuous latent space. Chem Sci. 2019;10:8016–8024.

[79] Li X, Xu Y, Yao H, et al. Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors. J Cheminform. 2020;12:42.

[80] Brown N, Fiscato M, Segler MHS, et al. GuacaMol: Benchmarking Models for de Novo Molecular Design. J Chem Inf Model. 2019;59:1096–1108.

[[81] Guimaraes GL, Sanchez-Lengeling B, Outeiral C, et al. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models]. arXiv [stat.ML]. 2017. Available from: http://arxiv.org/abs/1705.10843.

[82] Prykhodko O, Johansson SV, Kotsias P-C, et al. A de novo molecular generation method using latent vector based generative adversarial network. J Cheminform. 2019;11:74.

[83] Yasonik J. Multiobjective de novo drug design with recurrent neural networks and nondominated sorting. J Cheminform. 2020;12:14.

[84] Nicolaou CA, Brown N. Multi-objective optimization methods in drug design. Drug Discov Today Technol. 2013;10:e427–e435.

[85] Coley CW. Defining and exploring chemical spaces. Trends in Chemistry. 2021;3:133–145.

**** Recent review of methodologies to generate virtually chemical structures.**

[86] Mendez D, Gaulton A, Bento AP, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47:D930–D940.

[87] Ertl P. Magic Rings: Navigation in the ring chemical space guided by the bioactive rings. J Chem Inf Model. 2021, in press, doi: 10.1021/acs.jcim.1c00761.

[88] Zabolotna Y, Ertl P, Horvath D, et al. NP Navigator: A new look at the natural product chemical space. Mol Inf. 2021;40:e2100068.

[89] Chen Y, Kirchmair J. Cheminformatics in natural product-based drug discovery. Mol Inf. 2020;39:e2000171.

[90] Medina-Franco JL, Saldívar-González FI. Cheminformatics to characterize pharmacologically active natural products. Biomolecules. 2020;10:1566.

[91] Sorokina M, Merseburger P, Rajan K, et al. COCONUT online: Collection of Open Natural Products database. J Cheminform. 2021;13:2.

*** Comprehensive compendium of natural products in the public domain.**

[92] Irwin JJ, Tang KG, Young J, et al. ZINC20—A free ultralarge-scale chemical database for ligand discovery. J Chem Inf Model. 2020;60:6065–6073.

[93] Karageorgis G, Foley DJ, Laraia L, et al. Principle and design of pseudo-natural products. Nat Chem. 2020;12:227–235.

[94] Madariaga-Mazón A, Naveja JJ, Medina-Franco JL, et al. DiaNat-DB: a molecular database of antidiabetic compounds from medicinal plants. RSC Adv. 2021;11:5172–5178.

[95] Domínguez-Mendoza EA, Galván-Ciprés Y, Martínez-Miranda J, et al. Design, synthesis, and in silico multitarget pharmacological simulations of acid bioisosteres with a validated in vivo antihyperglycemic effect. Molecules. 2021;26:799

[96] Colín-Lozano B, Estrada-Soto S, Chávez-Silva F, et al. Design, synthesis and in combo antidiabetic bioevaluation of multitarget phenylpropanoic acids. Molecules.2018;23:340.

[97] Nava-Molina L, Uchida-Fuentes T, Ramos-Tovar H, et al. Novel CB1 receptor antagonist BAR-1 modifies pancreatic islet function and clinical parameters in prediabetic and diabetic mice. Nutr Diabetes. 2020;10:7.

[98] Gutierréz-Hernández A, Galván-Ciprés Y, Domínguez-Mendoza EA, et al. Design, synthesis, antihyperglycemic studies, and docking simulations of benzimidazole-thiazolidinedione hybrids. J Chem Chem Eng: published on line 15 Oct 2019, https://doi.org/10.1155/2019/1650145.

[99] Herrera-Rueda MÁ, Tlahuext H, Paoli P, et al. Design, synthesis, in vitro, in vivo and in silico pharmacological characterization of antidiabetic N-Boc-l-tyrosine-based compounds. Biomed Pharmacother. 2018;108:670–678.