Journal of Computer-Aided Molecular Design (2022) 36:341–354
https://doi.org/10.1007/s10822-021-00399-1

PERSPECTIVE

Progress on open chemoinformatic tools for expanding and exploring the chemical space

José L. Medina-Franco[1] · Norberto Sánchez-Cruz[1] · Edgar López-López[1,2] · Bárbara I. Díaz-Eufracio[1]

Received: 10 May 2021 / Accepted: 14 June 2021 / Published online: 18 June 2021

Abstract

The concept of chemical space is a cornerstone in chemoinformatics, and it has broad conceptual and practical applicability in many areas of chemistry, including drug design and discovery. One of the most considerable impacts is in the study of structure–property relationships where the property can be a biological activity or any other characteristic of interest to a particular chemistry discipline. The chemical space is highly dependent on the molecular representation that is also a cornerstone concept in computational chemistry. Herein, we discuss the recent progress on chemoinformatic tools developed to expand and characterize the chemical space of compound data sets using different types of molecular representations, generate visual representations of such spaces, and explore structure–property relationships in the context of chemical spaces. We emphasize the development of methods and freely available tools focusing on drug discovery applications. We also comment on the general advantages and shortcomings of using freely available and easy-to-use tools and discuss the value of using such open resources for research, education, and scientific dissemination.

Keywords Chemoinformatics · Drug discovery · Molecular representation · Open-source · Structure–activity relationships · Webserver

Introduction

Chemical space is a cornerstone concept in chemoinformatics. It serves as a framework to study the chemical compounds that populate or might do so, the "chemical universe" i.e., all compounds that can exist. Although it seems a straightforward idea (in particular, if one associates the idea of the chemical space with the chemical universe), it is not easy to define uniquely. Other subjective and general notions frequently used in chemoinformatics are "similarity" [1], or "diversity," "molecular or structural complexity" [2], "chemical beauty" [3], "descriptors' usefulness", to name a few examples.

✉ José L. Medina-Franco
medinajl@unam.mx; jose.medina.franco@gmail.com

[1] DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, 04510 Mexico City, Mexico

[2] Departamento de Química y Programa de Posgrado en Farmacología, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Apartado 14-740, 07000 Mexico City, Mexico

The notion of chemical space has numerous practical applications. In drug discovery, chemical space has provided a solid conceptual framework to guide diversity analysis, structure classification, library design, compound selection, and assessment of structure–property and structure–activity relationships (SPR, SAR or SP(A)R) that is a fundamental practice in drug discovery [4]. As commented hereunder, the notion of chemical space is also related to computational chemogenomics, where one aims to predict (and then validate experimentally) the intersection between the chemical and biologically relevant space. Indeed, in the early '60 s, the quantitative analysis of the SAR marked a significant milestone in the history of chemoinformatics and computer-aided drug design [5].

This Perspective aims to discuss advances in the development of chemoinformatic resources to characterize the chemical space of compound data sets using different types of molecular representations, generate visual representations of such spaces, and explore SP(A)R in the context of chemical spaces. In addition to analyzing the currently known chemical space, we comment on recent trends to augment the number of molecules that could be made. We emphasize the development of open tools focused on applications relevant

to drug discovery. As part of the discussion, we comment briefly on the advantages and shortcomings of using freely available and user-friendly tools and comment on the value of using such tools in research, education, teaching, and scientific dissemination. This manuscript is organized into six main sections. After this introduction, Sect. 2 presents an overview of the concept of chemical space, providing examples of different definitions proposed in the literature. Section 3 covers advances on open resources to expand and describe the chemical space, e.g., augmenting the number of compounds either on-stock or virtually available and calculating chemical descriptors. Section 4 presents advances on the concept, methods for the visual representation of the chemical space, including free web servers. The section after that discusses progress on the exploration of SP(A)R in the context of chemical space, including the exploration of "StARs" (Structure–Activity Relationships) in chemical space. Section 6 presents the conclusions and future directions.

The concept of chemical space

Chemical space is a subjective concept and different definitions have been proposed, which has been reviewed elsewhere [4, 6]. For instance, Virshup et al. define chemical space as "An M-dimensional cartesian space in which compounds are located by a set of M physicochemical and/or chemoinformatic descriptors" [7]. Along the same lines, Arús-Pous et al. describe it as "a concept to organize molecular diversity by postulating that different molecules occupy different regions of a mathematical space where the position of each molecule is defined by its properties" [8]. Based on these notions, Fig. 1 shows what can be considered a "chemical space table," where the rows are the *N* number of chemical compounds themselves (identified by, for instance, a text identifier). The columns are an *M* number of descriptors that describe the compounds, defining the "M-dimensional cartesian space" of Virshup's definition.

A common pitfall is that chemical space itself is frequently taken as equivalent to an image, aka, a visual representation. Although in many practical uses of chemical space, data visualization plays a major role, the chemical space itself is a subjective and general notion that depends primarily on the choice of the number and type of the descriptors that define the M-dimensional space. When a visualization method is not well suited to analyze a particular set of compounds and descriptors, it is always possible to analyze and extract information (and knowledge) from the chemical space using the full set (or relevant subsets) of the initial *M*-dimensions. Unless there are only two or three descriptors that define the *M*-dimensions in Fig. 1 ($M=2$ or 3, in which case the chemical space could be represented visually with a scatter plot), it is required a method to portray the *M*-dimensional space into two- or three-dimensions (2D/3D). Advances on the approaches to generate a visual representation of the chemical space, including the chemical space networks (that are coordinate-free) are addressed and cited in Sect. 4.

Suppose one adds one or more columns to the table in Fig. 1, representing the values of biological activity evaluations. In that case, one can produce a data format to perform

Chemical compound type	ID	Physicochemical properties		Topological descriptors		Molecular fragments		Similarities values based on shapes		Others
		D1	D2	D3	D4	D5	D6	D7	D8	Dm
Approved drugs	M1									...
Natural products	M2									...
Food chemicals	M3									...
Virtual compounds	M4									...
Synthesizable compounds	M5									...
Organometallic compounds	M6									...
Peptides	M7									...
Others	Mn	...	...	...	...	...	...	...	...	...

Fig. 1 Schematic representation of the chemical space concept as an *M*-dimensional descriptor space

SAR studies, reminiscent of a QSAR table or SAR matrix. In light of the concept of polypharmacology and multi-target drug design, it is possible to explore structure multiple-activity Relationships, e.g., "get SmARt" [9]. The "QSAR tables" have been the starting points to perform from simple QSAR linear regression studies to complex multivariate models used now in machine learning. Furthermore, QSAR tables are the basis of computational chemogenomics that is a strategy to navigate the chemical and biologically relevant chemical space [10, 11].

Type of molecules

The molecules (e.g., rows in the chemical space table in Fig. 1) typically used in drug discovery projects are *small* organic molecules (loosely defined with a molecular weight below 1,000 Da although could be bigger). These include natural products that have a significant impact on drug discovery [12] and semi-synthetic compounds. However, other types of molecules are also of interest in drug discovery, such as therapeutic peptides and proteins [13, 14], antibodies, and metallodrugs [15, 16]. The representation of these types of compounds, particularly metallodrug and organometallic molecules, is a major challenge in chemoinformatics. The representation and descriptors for (short) peptides and proteins are borderline between chemoinformatics and bioinformatics. For this *Perspective,* we will focus on the efforts to visualize the chemical space of mostly *small organic molecules*.

Type of descriptors

The descriptors (columns in the chemical space table in Fig. 1) can be *any* set of numbers that defines the space in an orderly (logical and rational manner). The type of descriptors can be suited to define the desired space and apply the concept for an array of applications, depending on the project's goals. Molecular description and the type of descriptors are distinctive of the different informatic disciplines in such a way that they somehow contribute to shape disciplines such as bioinformatics, chemoinformatics, biomedical informatics, etc. [17]. As commented in detail elsewhere in chemoinformatics common descriptors are calculated based on linear notations that are well-suited to manage many chemical compounds. It is also well-known that there is no single or a set of "best" descriptors as they should be selected based on their performance on a specific task [18]. This is associated with the inductive learning process used in chemoinformatics (as opposed to deductive learning used predominantly in quantum mechanics) [19].

Common types of descriptors that have been used to define the chemical space of *small organic molecules* include whole molecular properties that are aimed at encoding the so-called "drug-like," "lead-like," ADME (absorption, distribution, metabolism, and Excretion), toxicity, and other pharmaceutical-relevant characteristics. Other major molecular representations are fingerprint-based descriptors of different designs (dependent and independent of the molecule [20], and descriptors associated with sub-structures. Also, it has been approached using combined representations (e. g., hybrid fingerprints or combined molecular representations in general).

Beyond drug discovery, a recent application of physicochemical properties and molecular fingerprints to explore SPRs is to generate models that predict the smell of odorant molecules [21].

As further commented below, a novel type of descriptors that have been used to explore chemical spaces is the ISIDA descriptors, used to navigate the chemical space of natural products [22, 23].

Capecchi et al. recently proposed the molecular fingerprint MAP4 (MinHashed atom-pair fingerprint up to a diameter of four bonds). MAP4 has shown good performance in similarity searching and visual representation of the chemical space for small molecules and larger molecules such as peptides [24]. Reymond et al. recently used the MAP4 fingerprint to visualize the chemical space of natural products and [25] and peptides libraries in the public domain [26].

Recently the in silico acid-based profile of small molecules has been used to explore the chemical space of small molecules with epigenetic activity [27] and natural products from different sources [28].

Open resources to expand and describe the chemical space

There are reviews of open chemoinformatics resources for numerous applications [29, 30]. For instance, Singh et al. recently reviewed online web servers to perform virtual screening of small molecules and docking [31]. The authors reported 68 web applications in that review and classified them into target-fishing, ligand-based, and structure-based virtual screening. The review also covered compound databases that provide different information relevant to drug discovery, such as approved drugs, patented molecules or small molecules commercially available. Wu et al. surveyed databases and software commonly used to predict ADME/Tox-related properties [32].

Regarding the use of free web servers, Table 1 outlines the advantages and disadvantages of using open-source programs and freely accessible web servers. Overall, a clear benefit and advantage over commercial software are that they provide resources for research groups with a limited budget [33] and support open science. Also, the correct use of open-source programs advocates data reproducibility

Table 1 Overview of advantages and disadvantages of using open tools, including web servers

	Advantages	Disadvantages
Web servers	Increased accessibility Budget, cost-effective Experts and non-experts in chemoinformatics can use them They are good resources for the education of beginners and teaching They are convenient tools for distance learning (provided the servers are correctly used) They contribute to the generation of multi- and transdisciplinary science	They could be used as black boxes Limited access to parameters Potential issues of intellectual property Sensitive to proprietary data They are usually limited to a given (relatively short) number of compounds to analyze
Stand alone software	No need for programming or previous experience in programming is not mandatory Ready to use and apply to a research project No sensitivity to proprietary data	It might depend on the operating system Cost–benefit increases
Scripts and programs	Broadly widely customizable It can be implemented onto web servers Faster data processing speed	Experience in programming required Support might not be easily accessible. Depend on the experts The learning curve can be steep, not necessarily read to use

and facilitates cross-comparisons. A general disadvantage or caution of free web servers and "easily accessible" software is that they can be used as black boxes if they are used with no knowledge of the limitations of the tools and might lead to poor interpretation. Also, "easy-to-use" software has the associated risk of being used to generate only data and not knowledge and might promote the practice of irrational use of computers for drug discovery. Herein, we not aimed to fully discuss these points that are beyond the main goal of this manuscript that is focused on the chemical space. Instead, we want to give a brief comment about this topic that has been discussed openly in more detail elsewhere [34].

Resources for generating and organizing chemical structures

In the last few years, the chemical space has been growing rapidly: the number of compounds available in stock or that could be synthesized increases. Based on the Virshups' concept of chemical space (vide supra), generating compounds could be graphically represented as incrementing the number of rows in the "chemical space table" of Fig. 1. Chemical databases systematically organize the information of chemical compounds, and such databases have played a key role in drug discovery [35]. Progress on the development of compound databases in the public domain for drug discovery applications has been reviewed recently, and the interested reader is directed to these publications [36, 37].

In-stock and on-demand libraries

Virtual and make-on-demand libraries are having a significant impact on drug discovery. As pointed out by Walters, progress on the computer capabilities for generating and storing chemical compounds has increased the number of organic molecules that potentially could be synthesized [38].

A prominent example of a freely available and large library is the Generated Databases (GDB) developed in the group of Reymond et al. [39]. The most recent version is GDB-17 that contains 166.4 billion compounds up to 17 non-hydrogen atoms that include molecules not seen in the traditional medicinally relevant chemical space but have promising features to identify novel hit molecules [40].

Another recent development of an open resource to access purchasable or on-demand chemical libraries is ZINC20 that contains more than 9 million in-stock molecules and billions of new on-demand molecules [41]. Large-scale virtual screening of make-on-demand collections has led to discovering compounds with novel chemical scaffolds and submicromolar bioactivity [42]. Notably, the newest version of ZINC20 includes resources to generate a visual representation of the chemical space of the so-called "ultra large-scale chemical database [41].

Interestingly, the collection of compounds, so-called "dark chemical matter," represents a particular region of the chemical space that is mostly inactive [43].

Another recent development is the increase in the availability of natural product collections in the public domain that surpasses the half-million molecules [44]. A notable advance in this area is the assembly of the public database COCONUT (COlleCtion of Open NatUral producTs) [45]. In response to the COVID-19 pandemic, large and small collections and data sets of natural products have been virtually screened to identify potential compounds active in a number of molecular targets of SARS-CoV2. In most cases, however, experimental validation of the computational hits has to be performed as many publications were the result

of a "hype" and easy access to resources to conduct virtual screening.

De novo design and structure generation

Beyond the significant increase of chemical compounds that can be accessed (either in-stock or readily accessible after synthesis) a common trend now is the generation of chemical compounds designed de novo *using* machine learning. This has been reviewed recently in excellent review papers [8, 46].

There have also been advances in the automated generation of short peptides for drug discovery applications. A recent example in this area is the development of the free web server D-Peptide Builder that enumerates linear and cyclic combinatorial peptide libraries (Fig. 2) [47]. The server computes physicochemical properties of the newly enumerated peptides and provides tools to perform quantitative analysis of the structural diversity. D-Peptide builder also enables a visual representation of the chemical space of the libraries and compares it with the chemical space of five preloaded compound data sets (including small molecules and peptides approved for clinical use, natural products, macrolides and non peptide protein- protein interaction modulators).

PepCoGen is also a free web server for generating peptides with a specific physicochemical profile [48]. In particular, the server generates all possible combinations of peptides by modifying the amino acids having a comparable physicochemical property profile at a given position.

On a separate work, the code of the Peptide Design Genetic Algorithm (PDGA) was made publicly available. PDGA is designed to generate peptide sequences of different topologies so that the generated sequences are similar to a given reference molecule (as measured considering macromolecule extended atom-pair fingerprint (MXFP) (an atom-based fingerprint that considers the shape and pharmacophore features of the molecules [49]. The research group of Reymond has reviewed computational methods to design, generate and visualize the chemical space of peptides [26].

In order to support teaching in chemoinformatics, a tutorial that describes how to enumerate virtual libraries was published recently [50]. The tutorial describes a step-by-step procedure for anyone interested in designing and building chemical libraries with or without experience in using computational tools.

Resources for calculating descriptors freely available

In parallel to recent developments to enumerate, generate (synthesize), and make available chemical compounds (e.g., increase the number of rows in the "chemical space table" of Fig. 1 (vide supra), there has been a lot of progress in the development of descriptors, e.g., augment the number of *M*-dimensions or "columns" in Fig. 1. Of note, depending on the project's goals, one can generate a given

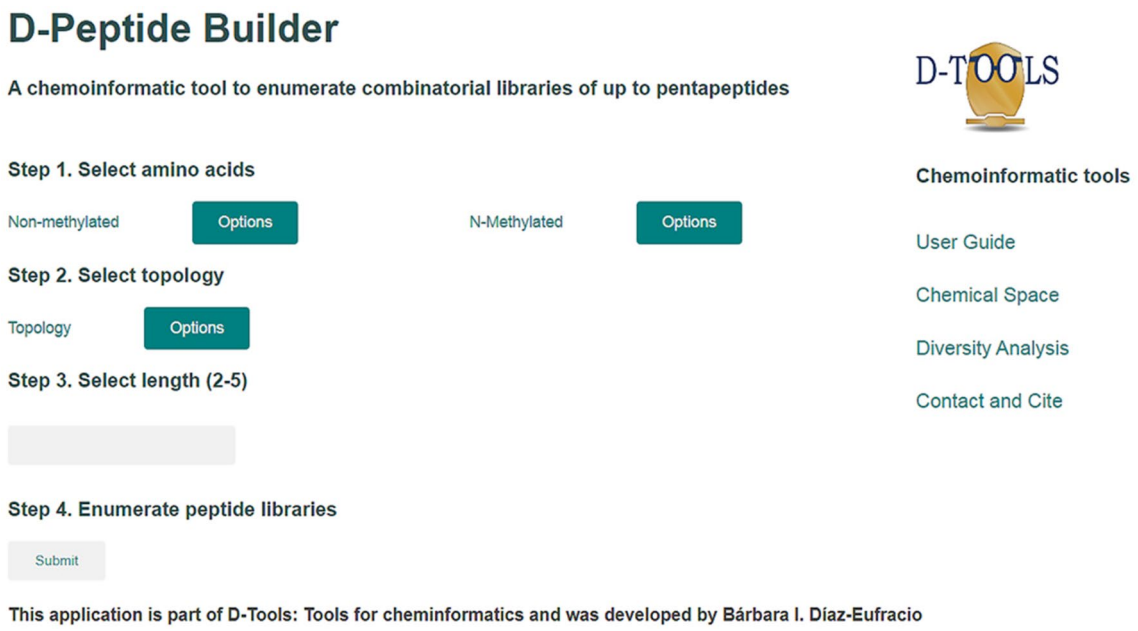

Fig. 2 The graphical user interface of D-Peptide builder: an example of a recent free webserver to generate compounds. D-Peptide builder enumerates combinatorial peptide libraries

finite set of descriptors to define the chemical space of the compounds under study. Thus, one can develop "different types of chemical spaces," e.g., defined by different sets of *M*-descriptors (Fig. 1). Arguably, it has been commented that "different chemical spaces" are associated by different types of molecules (small molecules, biologics, polymers, materials, etc. [46]). Under the later notion, molecules with different nature (like polymers, materials, etc.) would require a particular set of *M*-descriptors.

To define or generate the *M*-descriptors and define the chemical space using open-source and freely available software, there are several tools that have been available in the public domain for several years now. Typical examples include MayaChemTools (chemistry toolkit) [51], PaDEL-Descriptors [52], and the 3D descriptors implemented in QuBiLs-MIDAS [53], which was updated recently [54]. Additional free resources recently developed are briefly commented on hereunder.

PyDescriptors is a set of freely available 11,145 molecular descriptors easily interpretable and thus appropriate for QSAR studies [55]. PyDescriptors include 1D, 2D, and 3D descriptors that encode atomic fragments, pharmacophoric patterns, and diverse fingerprints. The PyDescriptors is a Python-based plugin that is implemented in PyMOL.

Mordred package for Python contains 1,800 2D and 3D descriptors freely available and promising for chemoinformatic studies and SPR analysis [56]. The descriptors can be used for large molecules (e.g., maitotoxin, a large non-polymer natural product with a molecular weight of 3,422). The Python package can be installed and used on different platforms (Linux, Windows, macOS). In the original publication [56] the Mordred descriptors were compared with the PaDEL-Descriptors [52] and turned out to be faster.

Another recent development in descriptors calculations is ChemDes [57]. This is a public integrated web-based platform that calculates 2D and 3D descriptors and molecular fingerprints. It calculates 3,679 descriptors (BlueDesc, Chemopy, CDK, RDKit, and PaDEL) and 59 types of molecular fingerprints for small (drug type) molecules. ChemDes is freely accessible via a previous registration, at http://www.scbdd.com/chemdes/ (accessed May 1st, 2021).

Overall, a critical and controversial point of chemical descriptors is their interpretability and physical meaning. In predictive models, it is open for discussion if the descriptors do not only show how a good statistical association between the chemical structure and the property (e.g., biological activity) of interest but if the descriptors can actually *explain* or contribute to the *causality* of the activity as encoded by the chemical descriptors [58, 59].

Resources for the visualization of chemical space

Visualization of chemical space plays a key role in communicating and disseminating information with experts and non-experts within a research group, an organization, community, and the research community on the large. In practice, chemical space is commonly studied accompanied by a graphical representation of the descriptors, typically a low-dimensional graph (2D or 3D). Formally speaking the chemical space (Fig. 1) could be unidimensional (1D), 2D, 3D and can be represented straightforwardly using scatter plots. The challenge comes when the *M*-dimensions are four or more. To this end, different mathematical approaches to reduce dimensions and techniques for data visualization have been applied to project chemical information in low dimensions and then map another property, such as biological activity, on that low-dimensional representation. In the past few years, progress on data visualization has been reviewed by different authors [6, 60, 61]. However, generating meaningful, interpretable, and useful graphical representations of chemical space is not trivial. Visualization of the chemical space (in particular in light of the rapid expansion of the compounds that might populate the space) is an area of active research to develop or improve methods [62]. Representative novel developments in the visual representation of the chemical space using open-source and freely available resources are discussed hereunder.

The research group of Varnek et al. generated the so-called "Universal REACH map, and application of the Generative Topographic Mapping (GTM) [63] to visualize the chemical space of chemicals from the Registration Evaluation Authorization and restriction of Chemicals (REACH) [64]. GTM produces 2D graphs on which each compound is represented with a data point. Ecotoxicological properties were mapped onto the 2D graph. The Universal REACH map was then used to classify and evaluate the property of new chemicals projected onto the map with a balanced accuracy from 0.60 to 0.78. In independent work, GTM was used to visualize a large library of 40 million fragment-like molecules [65] and the entire ZINC database of purchasable compounds, relative to 1.6 million biologically relevant molecules in ChEMBL [66]. A similar chemography approach using GTM was implemented to navigate the chemical space of 800 million organic molecules and identify "anti-CoV" regions [67]. More recently, GTM was used as a framework to visualize interactively the chemical space of a large database of natural products (COCONUT, vide supra) and ChEMBL [22]. The GTM maps were implemented into a freely available intuitive online tool called Natural Products Navigator (vide infra).

ChemMaps is a methodology for the visual representation of chemical space. It is based on the similarity matrix of compound data sets generated with the similarity computed with fingerprints and a similarity coefficient. ChemMaps is based on a reference or satellite approach implemented in ChemGPS [68] with the working hypothesis that satellites are, in principle, molecules whose similarity to the rest of the molecules in the database provides sufficient information for generating a visualization of the chemical space. The code to generate ChemMaps is freely available [69].

Another methodological advance in the visualization of chemical space is given by virtual reality. Probst and Reymond developed a virtual reality chemical space of DrugBank where the user can interactively explore the contents of this database. The source code of the application is publicly available [70].

Chemical space networks (CSNs) represent another major conceptual advance to generate visual representations of the chemical space, as discussed in detail by Maggiora and Bajorath [71, 72]. A major feature of CSNs is that they are coordinate-free representations of the chemical space. An algorithm to transform a multidimensional chemical space into CSNs readily has been developed that is further useful to explore SARs [73]. CSNs have been used in many applications, including the assessment of the molecules from patents [74].

DataWarrior is a free stand-alone program that is being increasingly used for diverse chemoinformatics tasks, including data visualization [75, 76]. Datawarrior in a recent version (number 5.00) implemented t-SNE [77]. At the time of writing this manuscript (May 2021) the latest release of DataWarrior is 5.5.0.

Web servers

Table 2 summarizes free web applications to visualize the chemical space of compound collections. The table includes ChemGPS-NP, one of the first free web applications developed to visualize the biologically relevant chemical space [78]. In addition to ChemGPS-NP, some of the web servers in the table are dedicated to the browsing and visualization of the chemical space of user-supplied compounds (e.g., ChemMap.com [79], tMAPs [80], Natural Products Navigator [22]. Other websites include other functionalities such as D-Peptide Builder [47], and the Platform for Unified Molecular Analysis (PUMA) [81]. D-Peptide Builder is an application to enumerate chemical spaces of peptide combinatorial libraries and visualize chemical spaces. PUMA is a server that integrates the calculation of descriptors and visual representation of the chemical space based on those descriptors. Both web servers are part of D-Tools, a set of free web applications for chemoinformatics (https://www.difacquim.com/d-tools/) [82]. The research group of Reymond has developed several free web applications in Table 2 for the interactive visualization of chemical space (https://gdb.unibe.ch/tools/).

Figure 3 shows an example of a visualization of chemical space using the free server PUMA (Table 2). The figure shows a principal component analysis based on six physicochemical properties of pharmaceutical interest of two focused libraries (targeting DNMT1 and epigenetic targets). The libraries represent commercial synthetic compounds that can be acquired from chemical vendors for experimental screening). In PUMA, the user supplies the SMILES strings of curated compound libraries, and the server computes the

Table 2 Examples of freely available web servers for the interactive visualization of chemical space

Web Server	Brief description	URL (accessed May 1, 2021)	Ref
AtlasCBS	Generates two-dimensional, dynamical representations of its contents in terms of Ligand Efficiency Indices	https://www.ebi.ac.uk/chembl/atlascbs/intro.jsp	[83]
ChemMaps	Webserver developed to navigate throughout chemical and environmental chemical space	https://sandbox.ntp.niehs.nih.gov/chemmaps/	[79]
ChemGPS-NP	ChemGPS-NP Web is a system for computing the eight principal components (dimensions) describing physical–chemical properties for a reference set of compounds	https://chemgps.bmc.uu.se	[78]
Natural Products Navigator	Visualization and navigation through the chemical space of NPs and NP-like molecules	https://infochm.chimie.unistra.fr/npnav/chematlas_userspace/	[22]
tMAP	Visualization library for large, high-dimensional data sets	https://tmap.gdb.tools/	[80]
Faerun	Chemical space accessible by the PDGA with an interactive 3D map of the MXFP property space	http://faerun.gdb.tools/	[49]
PDB Explorer	Interactive visualization and similarity search of the RSCB Protein Databank in shape space	http://www.cheminfo.org/pdbexplorer/	
D-Peptide Builder	Enumerate chemical spaces of peptide combinatorial libraries and visualize chemical spaces	http://dpeptidebuilder.quimica.unam.mx:4000/	[47]
Platform for Unified Molecular Analysis	Online server to visualize the chemical space and compute the molecular diversity of your data sets	http://132.248.103.152:3838/PUMA/	[81]

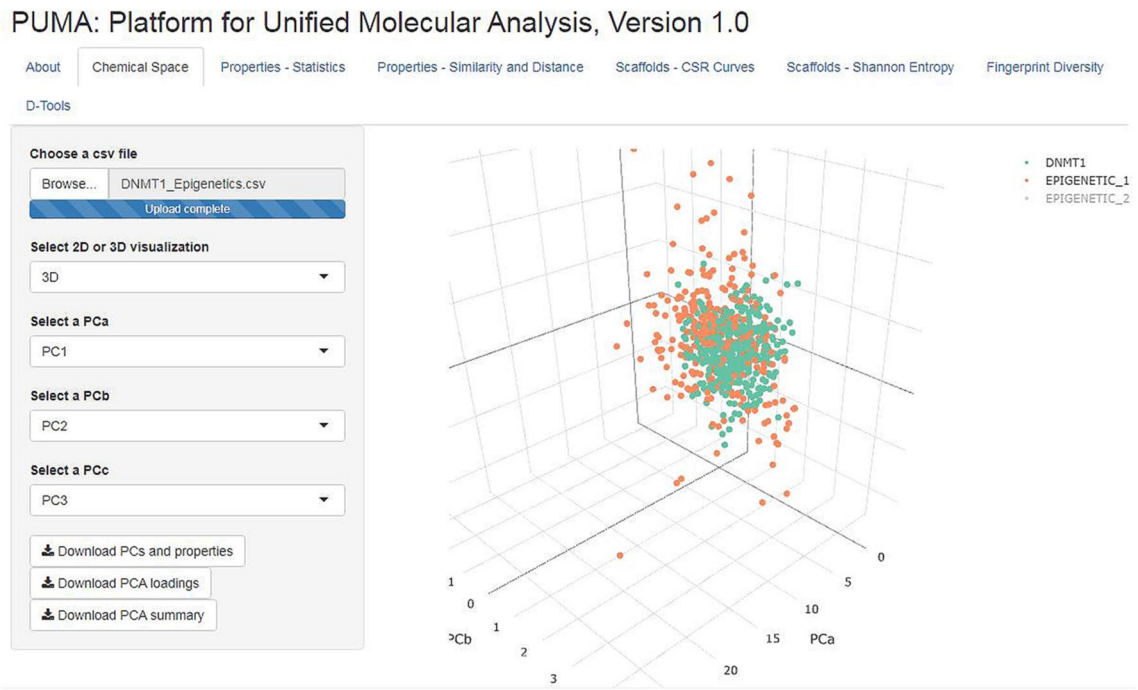

Fig. 3 Visual representation of the chemical space of user-supplied chemical structures using the free server Platform for Unified Molecular Analysis (PUMA). The figure shows the visual representation of the chemical space of two synthetic commercial libraries targeted for epigenetic targets (709 compounds in total). The principal component analysis is based on six physicochemical properties of pharmaceutical interest as described in [81]. On the free web server, the 2D or 3D plot is interactive

physicochemical properties internally (e.g., the descriptors) and then performs the principal component analysis. The user chooses to plot the first two or three principal components. From the lower left part of the graphical user interface (Fig. 3), the user can download from the sever the raw data and the loadings and a summary of the analysis. Full details of the server are described in [81].

Exploring for structure–activity relationships (StARs) in chemical space

As commend above, since chemical space is defined by a set of *M* descriptors (Fig. 1), that encode the structural or other characteristics of the molecules, it can serve as a basis to analyze SPRs and SmARTs if one adds one or more dimensions that describe the property (e.g., biological activity) of the compounds (i.e., the biological profile). Visually, the property (including the biological "activity") is usually mapped in the chemical space using a color (continuous color scale or categorical scheme) (Fig. 1) but could be visually represented in different forms (e.g., shapes for categorical variables). The visualization of SP(A)R and "STaRs in chemical space) has been commented on in the literature [61, 84]. Herein we emphasize exemplary most recent advances in this area.

Activity landscapes

Prof. Gerald Maggiora was one of the first investigators that kicked off the research on a general concept with high relevance in drug discovery: activity landscape modeling with his founding Editorial on activity cliffs [85]: pair of compounds with high structure similarity but unexpectedly large potency differences. Over the past few years, the concept, interpretation, and applications of activity cliffs have evolved, as reviewed by Bajorath et al. [86–88]. One of the most recent developments in the activity landscape concept has been the extension to model other properties of general interest beyond drug discovery [89].

To illustrate this point, Fig. 4a shows the Structure–Property Similarity (SPS) map for tubulin inhibitors generated with the free website Activity Landscape Plotter [90]. Each data point represents a pairwise comparison that shows the relationship between the difference in Topological Surface Area (TPSA) and the molecular similarity. The data points are further distinguished by the SALI value [91], using a continuous color scale from a low value (green) to a high value (red). In this context, higher SALI values represent a higher relationship between TPSA values and similarity between each pair of compounds. In contrast, Fig. 4b shows a Dual-Property Difference (DPD) map, plotting all pairwise activity differences of tubulin inhibitors with A-549 cell-line

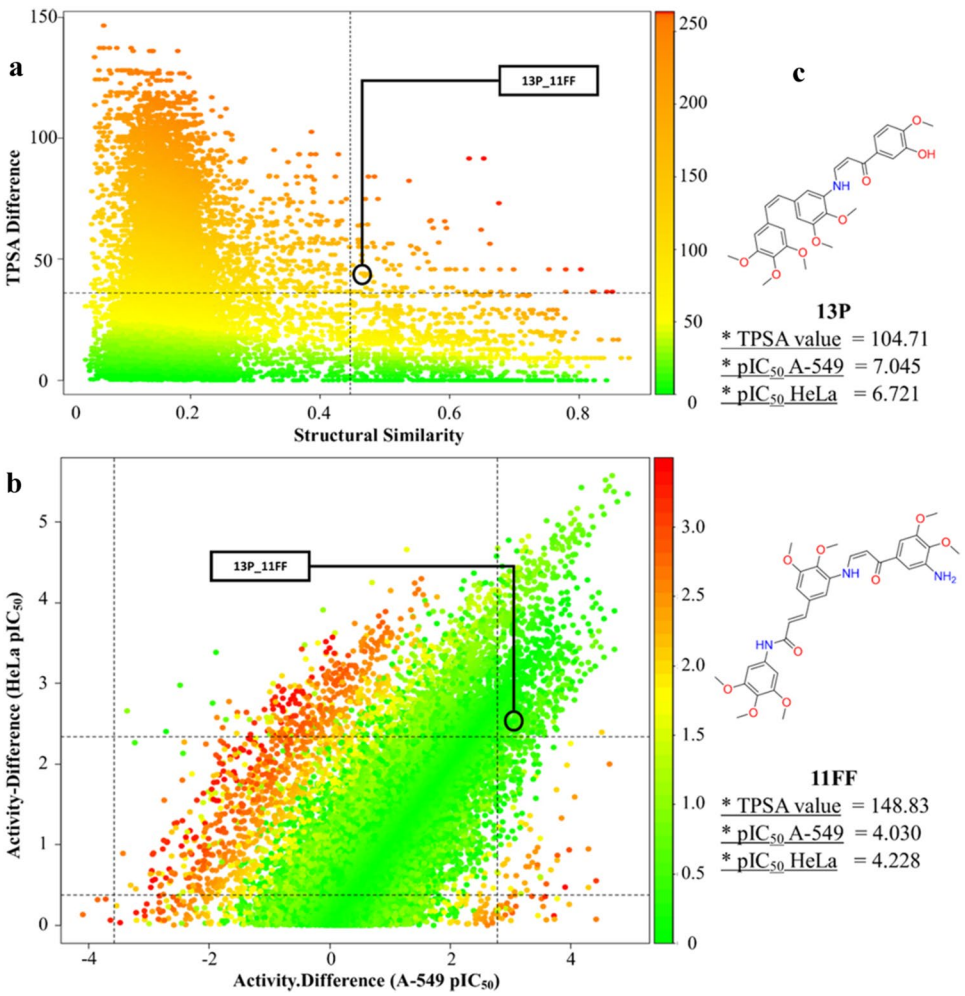

Fig. 4 Property Landscapes of compounds with activity against Tubulin using cell-based inhibition data. **a** Structure–Property Similarity (SPS) map of 188 tubulin inhibitors that correspond to 17,578 pairwise comparisons. The property cliffs are displayed in the upper-right zone. Each data point was colored using a SALI value scale from green (low) to red (high); **b** Dual Property Difference (DPD) map of tubulin inhibitors. The dual active compounds are displayed in the upper right zone. Each data point was colored using a selectivity score from green (low) to red (high); **c** Example of a property and dual activity cliff

(X-axis) and HeLa cell-line (Y-axis). Therefore, DPD maps facilitate the identification of compounds with selective and dual activity.

Using SPR graphs allows us to relate chemical structures with their properties, bioactivities, or other characteristics. For example, Fig. 4 shows a property and dual activity cliffs (**13P** and **11FF**) pair. These compounds are structurally similar (0.470—using ECFP6 and the Tanimoto coefficient). However, their TPSA is different (property cliff). It is well documented that TPSA values > 140 (like that of compound **11FF** in Fig. 4C) lose their ability to cross membranes, unlike compounds with TPSA values < 140 (like that of compound **13P**) that retain this ability [92]. This is a case study that illustrates the similarity-property-activity relationship.

Constellation plots

Constellation plots were developed to combine a substructure-based representation and classification of compounds with a coordinate-based representation of chemical space [93]. Constellation plots are 2D graphs that combine substructure-based clustering of compounds with a fingerprint-based similarity classification of the chemical scaffolds. The substructure-based clustering of the molecules is based on the concept of analog series-based scaffolds [94, 95]. Since the biological activity data (or any other property) can be mapped into a Constellation plot, these 2D representations of the chemical space enable identifying whole regions in chemical space rich in SPR annotations: groups of molecules, aka "constellations" in chemical space. The groups of molecules rich in biological activity would be light "bright StARs" in chemical space and be different from 'dark regions': groups of molecules with no biological activity [61].

Additionally, in the constellation plots, the analog series with similar chemical structures are closely ordered because they share similar X and Y coordinates in the 2D plots. In contrast, analog series with more different structures are far apart. Recently, López-López E. et al. proposed a methodology to navigate interactively/dynamically in the chemical space using constellation plots [96]

by implementing the DataWarrior software [76]. All this allows applying filters for compounds, analogous series, biological activity, and other properties of pharmaceutical interest using an intuitive platform that is well suited for all users (expert or non-experts on chemoinformatics tools). Figure 5 illustrates an example of a Constellation plot for a series of tubulin inhibitors. The plot shows 147 data points, each one representing an analog series. The size of the data point indicates the relative number of compounds in each analog series, and the color is the average activity of the compound in the series so that green-to-red colored dots point to analog series enriched with active molecules, hence more promising for further development. In contrast, cyan-to-blue colored dots indicate analog series with mostly inactive molecules. Full details of the study are described elsewhere [96].

Constellation plots have been used to navigate the chemical space of high-throughput screening data of compounds consistently tested against the same panel of cell lines. In that work, Naveja et al. proposed a proof-of-concept of a method for finding a consistent cell-selective analog series of chemical compounds and identified the so-called "luminaries in chemical space" [97].

Conclusions and perspectives

For years the subjective but fundamental notion of chemical space has assisted drug discovery projects. Chemical space is also a cornerstone concept in chemoinformatics. In the past few years, we have witnessed an expansion of the chemical space regarding the number of compounds that are known or can be synthesized in principle. As commented on this Perspective, it is growing how the chemical compounds can be represented and the number of public tools to compute descriptors. Open-source codes can be implemented in other public web servers, chemoinformatics suits, and desktop programs. In any case, the ready availability of compound libraries that are expanding the chemical space and the ready availability of tools to conduct virtual screening: e.g., in silico bioactivity profiling (or computer-assisted compound selection of the chemical

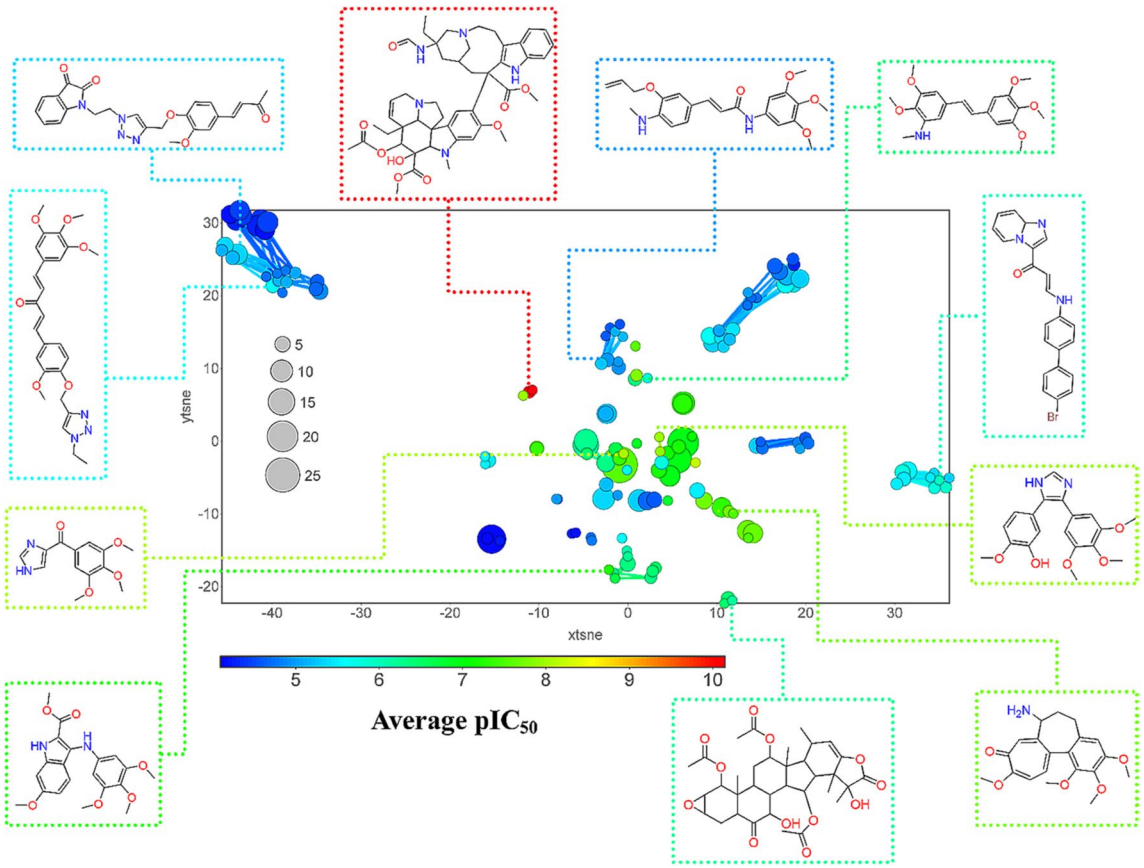

Fig. 5 Constellation plot of compounds with activity against Tubulin using cell-based inhibition data. The plot shows 147 data points, each one representing an analog series. The size of the data point indicates the relative number of compounds in each analog series, and the color is the average activity of the compound in the series. Linking lines represent shared molecules between two analog series. Figure was adapted from López-López E. et al. [96]

space), favor the potential identification of small molecules with therapeutically relevant targets.

Similar to the *expansion* of the chemical space (more compounds and more descriptors, e.g., enlarge the table in Fig. 1)), novel free applications and open-source methods to generate visual representations of the chemical space are emerging and evolving. Recent developments include CSNs, TMAPs, GTMs, Constellation plots, and ChemMaps. Virtual reality has started to facilitate the interactive exploration of chemical spaces. Some of these visualization tools have been implemented in freely available websites that enable the browsing of chemical spaces. Several methodologies aim to assist the analysis of SP(A)Rs and identity promising regions or clusters of compounds in chemical space.

Despite numerous open-source and easily accessible ways to calculate molecule descriptors, the user has to pay close attention (rational use) by preparing -curating—the compounds and then generating appropriate descriptors relevant to the problem in question. Considering the large chemical databases and large sets of descriptors available: one of the first and critical questions is defining the chemical space to be explored by focusing on the type of compounds of interest and the type of descriptors. In several drug discovery applications, the choice of compounds and descriptors is *dynamic*: an iterative process where one explores different compounds and various descriptors that best suit the work goals.

We also want to encourage students, newcomers to the field, and users of free and easy-to-use tools and websites to properly use and interpret the concept of chemical space. Based on the topics discussed from this *Perspective*, chemical space is a subjective and complex notion and goes beyond nice and colorful graphs. Along these lines, we encourage that the newcomers to the field select the methods for the right reasons and not because they are "popular." Instead, because the methods are thoroughly validated and properly documented. The interested reader is referred to the *Opinion* manuscript "Rationality over fashion and hype in drug design," where this and related points are discussed in more detail, and it is open for discussion with the scientific community [34].

Acknowledgements E.L.-L. and B.I.D.-E. thank to the Consejo Nacional de Ciencia y Tecnología (CONACyT), Mexico, for the scholarships No. 762342 and 817896, respectively. Discussions with Jesús Naveja, Alexis Padilla and Rodrigo Gutiérrez are acknowledged. We thank the support of DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), grant IN201321.

Funding DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), grant IN201321.

Declarations

Conflicts of interest None.

References

1. Maggiora G, Vogt M, Stumpfe D, Bajorath J (2014) Molecular similarity in medicinal chemistry. J Med Chem. https://doi.org/10.1021/jm401411z
2. Méndez-Lucio O, Medina-Franco JL (2017) The many roles of molecular complexity in drug discovery. Drug Discov Today 22:120–126. https://doi.org/10.1016/j.drudis.2016.08.009
3. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4:90–98. https://doi.org/10.1038/nchem.1243
4. Medina-Franco JL, Martínez-Mayorga K, Giulianotti MA, Houghten RA, Pinilla C (2008) Visualization of the chemical space in drug discovery. Curr Comput Aided Drug Des 4:322–333. https://doi.org/10.2174/157340908786786010
5. Gasteiger J (2020) Chemistry in times of artificial intelligence. ChemPhysChem 21:2233–2242. https://doi.org/10.1002/cphc.202000518
6. Osolodkin DI, Radchenko EV, Orlov AA, Voronkov AE, Palyulin VA et al (2015) Progress in visual representations of chemical space. Exp Opin Drug Discov 10:959–973. https://doi.org/10.1517/17460441.2015.1060216
7. Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN (2013) Stochastic Voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. J Am Chem Soc 135:7296–7303. https://doi.org/10.1021/ja401184g
8. Arús-Pous J, Awale M, Probst D, Reymond JL (2019) Exploring chemical space with machine learning. Chimia (Aarau) 73:1018–1023. https://doi.org/10.2533/chimia.2019.1018
9. Saldívar-González FI, Naveja JJ, Palomino-Hernández O, Medina-Franco JL (2017) Getting smart in drug discovery: chemoinformatics approaches for mining structure–multiple activity relationships. RSC Adv 7:632–641. https://doi.org/10.1039/C6RA26230A
10. Jacoby E, Mozzarelli A (2009) Chemogenomic strategies to expand the bioactive chemical space. Curr Med Chem 16:4374–4381. https://doi.org/10.2174/092986709789712862
11. Bajorath J (2013) A perspective on computational chemogenomics. Mol Inf 32:1025–1028. https://doi.org/10.1002/minf.201300034
12. Atanasov AG, Zotchev SB, Dirsch VM, Orhan IE, Banach M et al (2021) Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov 20:200–216. https://doi.org/10.1038/s41573-020-00114-z
13. Fosgerau K, Hoffmann T (2015) Peptide therapeutics: current status and future directions. Drug Discovery Today 20:122–128. https://doi.org/10.1016/j.drudis.2014.10.003
14. Pandya AK, Patravale VB (2021) Computational avenues in oral protein and peptide therapeutics. Drug Discov Today. https://doi.org/10.1016/j.drudis.2021.03.003
15. Mjos KD, Orvig C (2014) Metallodrugs in medicinal inorganic chemistry. Chem Rev 114:4540–4563. https://doi.org/10.1021/cr400460s
16. Anthony EJ, Bolitho EM, Bridgewater HE, Carter OWL, Donnelly JM et al (2020) Metallodrugs are unique: opportunities and challenges of discovery and development. Chem Sci 11:12888–12917. https://doi.org/10.1039/D0SC04082G

17. López-López E, Bajorath J, Medina-Franco JL (2021) Informatics for chemistry, biology, and biomedical sciences. J Chem Inf Model 61:26–35. https://doi.org/10.1021/acs.jcim.0c01301
18. David L, Thakkar A, Mercado R, Engkvist O (2020) Molecular representations in AI-driven drug discovery: a review and practical guide. J Cheminf 12:56. https://doi.org/10.1186/s13321-020-00460-5
19. Varnek A, Baskin II (2011) Chemoinformatics as a theoretical chemistry discipline. Mol Inf 30:20–32. https://doi.org/10.1002/minf.201000100
20. Maggiora G (2014) Introduction to molecular similarity and chemical space. In: Martinez-Mayorga K, Medina-Franco J (eds) Foodinformatics. Springer, Cham
21. Sharma A, Kumar R, Ranjta S, Varadwaj PK (2021) Smiles to smell: decoding the structure-odor relationship of chemical compounds using the deep neural network approach. J Chem Inf Model 61:676–688. https://doi.org/10.1021/acs.jcim.0c01288
22. Zabolotna Y, Ertl P, Horvath D, Bonachera F, Marcou G, et al. (2021) NP Navigator: a new look at the natural product chemical space. ChemRxiv. https://doi.org/10.26434/chemrxiv.14236457.v1
23. Ruggiu F, Marcou G, Varnek A, Horvath D (2010) Isida property-labelled fragment descriptors. Mol Inf 29:855–868. https://doi.org/10.1002/minf.201000099
24. Capecchi A, Probst D, Reymond J-L (2020) One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminf 12:43. https://doi.org/10.1186/s13321-020-00445-4
25. Capecchi A, Reymond JL (2020) Assigning the origin of microbial natural products by chemical space map and machine learning. Biomolecules. https://doi.org/10.3390/biom10101385
26. Capecchi A, Reymond J-L (2021) Peptides in chemical space. Med Drug Discov 9:100081. https://doi.org/10.1016/j.medidd.2021.100081
27. Santibáñez-Morán MG, Medina-Franco JL (2021) The acid/base characterization of molecules with epigenetic activity. ChemMedChem. https://doi.org/10.1002/cmdc.202001009
28. Santibáñez-Morán MG, Medina-Franco JL (2020) Analysis of the acid/base profile of natural products from different sources. Mol Inform 39:e1900099. https://doi.org/10.1002/minf.201900099
29. Villoutreix BO, Lagorce D, Labbé CM, Sperandio O, Miteva MA (2013) One hundred thousand mouse clicks down the road: selected online resources supporting drug discovery collected over a decade. Drug Discovery Today 18:1081–1089. https://doi.org/10.1016/j.drudis.2013.06.013
30. Gonzalez-Medina M, Naveja JJ, Sanchez-Cruz N, Medina-Franco JL (2017) Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. RSC Adv 7:54153–54163. https://doi.org/10.1039/C7RA11831G
31. Singh N, Chaput L, Villoutreix BO (2020) Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace. Briefings Bioinf 22:1790–1818. https://doi.org/10.1093/bib/bbaa034
32. Wu F, Zhou Y, Li L, Shen X, Chen G et al (2020) Computational approaches in preclinical studies on drug discovery and development. Front Chem 8:726. https://doi.org/10.3389/fchem.2020.00726
33. Willems H, De Cesco S, Svensson F (2020) Computational chemistry on a budget: supporting drug discovery with limited resources. J Med Chem 63:10158–10169. https://doi.org/10.1021/acs.jmedchem.9b02126
34. Medina-Franco JL, Martinez-Mayorga K, Fernández-de Gortari E, Kirchmair J, Bajorath J (2021) Rationality over fashion and hype in drug design. F1000Research 10(Chem Inf Sci):397. https://doi.org/10.12688/f1000research.52676.1
35. Miller MA (2002) Chemical database techniques in drug discovery. Nat Rev Drug Discov 1:220–227. https://doi.org/10.1038/nrd745
36. Scior T, Bernard P, Medina-Franco JL, Maggiora GM (2007) Large compound databases for structure-activity relationships studies in drug discovery. Mini-Rev Med Chem 7:851–860. https://doi.org/10.2174/138955707781387858
37. Moura Barbosa AJ, Del Rio A (2012) Freely accessible databases of commercial compounds for high- throughput virtual screenings. Curr Top Med Chem 12:866–877. https://doi.org/10.2174/156802612800166710
38. Walters WP (2019) Virtual chemical libraries. J Med Chem 62:1116–1124. https://doi.org/10.1021/acs.jmedchem.8b01048
39. Blum LC, Reymond J-L (2009) 970 Million drug-like small molecules for virtual screening in the chemical universe database Gdb-13. J Am Chem Soc 131:8732–8733. https://doi.org/10.1021/ja902302h
40. Meier K, Bühlmann S, Arús-Pous J, Reymond JL (2020) The Generated Databases (GDBs) as a source of 3d-shaped building blocks for use in medicinal chemistry and drug discovery. Chimia (Aarau) 74:241–246. https://doi.org/10.2533/chimia.2020.241
41. Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR et al (2020) ZINC20-a free ultralarge-scale chemical database for ligand discovery. J Chem Inf Model 60:6065–6073. https://doi.org/10.1021/acs.jcim.0c00675
42. Lyu J, Wang S, Balius TE, Singh I, Levit A et al (2019) Ultra-large library docking for discovering new chemotypes. Nature 566:224–229. https://doi.org/10.1038/s41586-019-0917-9
43. Wassermann AM, Lounkine E, Hoepfner D, Le Goff G, King FJ et al (2015) Dark chemical matter as a promising starting point for drug lead discovery. Nat Chem Biol 11:958–966. https://doi.org/10.1038/nchembio.1936
44. Chen Y, de Bruyn KC, Kirchmair J (2017) Data resources for the computer-guided discovery of bioactive natural products. J Chem Inf Model 57:2099–2111. https://doi.org/10.1021/acs.jcim.7b00341
45. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C (2021) COCONUT Online: collection of open natural products database. J Cheminform 13:2. https://doi.org/10.1186/s13321-020-00478-9
46. Coley CW (2021) Defining and exploring chemical spaces. Trends Chem 3:133–145. https://doi.org/10.1016/j.trechm.2020.11.004
47. Díaz-Eufracio BI, Palomino-Hernández O, Arredondo-Sánchez A, Medina-Franco JL (2020) D-Peptide Builder: a web service to enumerate, analyze, and visualize the chemical space of combinatorial peptide libraries. Mol Inf 39:e2000035. https://doi.org/10.1002/minf.202000035
48. Ali N, Shamoon A, Yadav N, Sharma T (2020) Peptide combination generator: a tool for generating peptide combinations. ACS Omega 5:5781–5783. https://doi.org/10.1021/acsomega.9b03848
49. Capecchi A, Zhang A, Reymond JL (2020) Populating chemical space with peptides using a genetic algorithm. J Chem Inf Model 60:121–132. https://doi.org/10.1021/acs.jcim.9b01014
50. Saldívar-González FI, Huerta-García CS, Medina-Franco JL (2020) Chemoinformatics-based enumeration of chemical libraries: a tutorial. J Cheminf 12:64. https://doi.org/10.1186/s13321-020-00466-z
51. Sud M (2016) Mayachemtools: an open source package for computational drug discovery. J Chem Inf Model 56:2292–2297. https://doi.org/10.1021/acs.jcim.6b00505
52. Yap CW (2011) Padel-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474. https://doi.org/10.1002/jcc.21707
53. García-Jacas CR, Marrero-Ponce Y, Acevedo-Martínez L, Barigye SJ, Valdés-Martiní JR et al (2014) Qubils-Midas: a parallel free-software for molecular descriptors computation based

on multilinear algebraic maps. J Comput Chem 35:1395–1409. https://doi.org/10.1002/jcc.23640
54. García-Jacas CR, Marrero-Ponce Y, Vivas-Reyes R, Suárez-Lezcano J, Martinez-Rios F et al (2020) Distributed and multicore Qubils-Midas Software V2.0: Computing Chiral, Fuzzy, Weighted and Truncated Geometrical Molecular Descriptors Based on Tensor Algebra. J Comput Chem 41:1209–1227. https://doi.org/10.1002/jcc.26167
55. Masand VH, Rastija V (2017) Pydescriptor: a new pymol plugin for calculating thousands of easily understandable molecular descriptors. Chemometrics Intell Lab Syst 169:12–18. https://doi.org/10.1016/j.chemolab.2017.08.003
56. Moriwaki H, Tian Y-S, Kawashita N, Takagi T (2018) Mordred: a molecular descriptor calculator. J Cheminf 10:4. https://doi.org/10.1186/s13321-018-0258-y
57. Dong J, Cao DS, Miao HY, Liu S, Deng BC et al (2015) Chemdes: an integrated web-based platform for molecular descriptor and fingerprint computation. J Cheminform 7:60. https://doi.org/10.1186/s13321-015-0109-z
58. Guha R (2008) On the interpretation and interpretability of quantitative structure-activity relationship models. J Comput-Aided Mol Des 22:857–871. https://doi.org/10.1007/s10822-008-9240-5
59. Chen CH, Tanaka K, Kotera M, Funatsu K (2020) Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. J Cheminform 12:19. https://doi.org/10.1186/s13321-020-0417-9
60. Reymond J-L (2015) The chemical space project. Acc Chem Res 48:722–730. https://doi.org/10.1021/ar500432k
61. Medina-Franco JL, Naveja JJ, López-López E (2019) Reaching for the bright stars in chemical space. Drug Discovery Today 24:2162–2169. https://doi.org/10.1016/j.drudis.2019.09.013
62. Rarey M, Nicklaus MC, Warr W (2021) Call for papers for the special issue: from reaction informatics to chemical space. J Chem Inf Model 61:1531–1532. https://doi.org/10.1021/acs.jcim.1c00321
63. Lin A, Baskin II, Marcou G, Horvath D, Beck B et al (2020) Parallel generative topographic mapping: an efficient approach for big data handling. Mol Inf 39:e2000009. https://doi.org/10.1002/minf.202000009
64. Lunghini F, Gilles M, Azam P, Enrici MH, Van Miert E et al (2021) Visualization and analysis of the reach-chemical space with generative topographic mapping. Mol Inf 40:e2000232. https://doi.org/10.1002/minf.202000232
65. Lin A, Horvath D, Afonina V, Marcou G, Reymond JL et al (2018) Mapping of the available chemical space versus the chemical universe of lead-like compounds. ChemMedChem 13:540–554. https://doi.org/10.1002/cmdc.201700561
66. Zabolotna Y, Lin A, Horvath D, Marcou G, Volochnyuk DM et al (2021) Chemography: searching for hidden treasures. J Chem Inf Model 61:179–188. https://doi.org/10.1021/acs.jcim.0c00936
67. Horvath D, Orlov A, Osolodkin DI, Ishmukhametov AA, Marcou G et al (2020) A chemographic audit of anti-coronavirus structure-activity information from public databases (ChEMBLl). Mol Inf 39:e2000080. https://doi.org/10.1002/minf.202000080
68. Rosen J, Lovgren A, Kogej T, Muresan S, Gottfries J et al (2009) ChemGPS-NPweb: chemical space navigation online. J Comput Aided Mol Des 23:253–259. https://doi.org/10.1007/s10822-008-9255-y
69. Naveja J, Medina-Franco J (2017) Chemmaps: towards an approach for visualizing the chemical space based on adaptive satellite compounds. F1000Research 6(Chem Inf Sci):1134. https://doi.org/10.12688/f1000research.12095.2
70. Probst D, Reymond JL (2018) Exploring Drugbank in virtual reality chemical space. J Chem Inf Model 58:1731–1735. https://doi.org/10.1021/acs.jcim.8b00402
71. Maggiora GM, Bajorath J (2014) Chemical space networks: a powerful new paradigm for the description of chemical space. J Comput Aided Mol Des 28:795–802. https://doi.org/10.1007/s10822-014-9760-0
72. Vogt M, Stumpfe D, Maggiora GM, Bajorath J (2016) Lessons learned from the design of chemical space networks and opportunities for new applications. J Comput Aided Mol Des 30:191–208. https://doi.org/10.1007/s10822-016-9906-3
73. de la Vega de León A, Bajorath J (2016) Chemical space visualization: transforming multidimensional chemical spaces into similarity-based molecular networks. Future Med Chem 8:1769–1778. https://doi.org/10.4155/fmc-2016-0023
74. Kunimoto R, Bajorath J (2017) Exploring sets of molecules from patents and relationships to other active compounds in chemical space networks. J Comput Aided Mol Des 31:779–788. https://doi.org/10.1007/s10822-017-0061-2
75. López-López E, Naveja JJ, Medina-Franco JL (2019) Datawarrior: an evaluation of the open-source drug discovery tool. Expert Opin Drug Discov 14:335–341. https://doi.org/10.1080/17460441.2019.1581170
76. Sander T, Freyss J, von Korff M, Rufener C (2015) Datawarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model 55:460–473. https://doi.org/10.1021/ci500588j
77. van der Maaten L, Hinton G (2008) Visualizing data using T-SNE. J Mach Learn Res 9:2579–2605
78. Larsson J, Gottfries J, Muresan S, Backlund A (2007) ChemGPS-NP: tuned for navigation in biologically relevant chemical space. J Nat Prod 70:789–794. https://doi.org/10.1021/np070002y
79. Borrel A, Kleinstreuer NC, Fourches D (2018) Exploring drug space with Chemmaps.com. Bioinformatics 34:3773–3775. https://doi.org/10.1093/bioinformatics/bty412
80. Probst D, Reymond J-L (2020) Visualization of very large high-dimensional data sets as minimum spanning trees. J Cheminf 12:12. https://doi.org/10.1186/s13321-020-0416-x
81. González-Medina M, Medina-Franco JL (2017) Platform for unified molecular analysis: PUMA. J Chem Inf Model 57:1735–1740. https://doi.org/10.1021/acs.jcim.7b00253
82. Naveja JJ, Oviedo-Osornio CI, Trujillo-Minero NN, Medina-Franco JL (2018) Chemoinformatics: a perspective from an academic setting in Latin America. Mol Divers 22:247–258. https://doi.org/10.1007/s11030-017-9802-3
83. Cortés-Cabrera A, Morreale A, Gago F, Abad-Zapatero C (2012) Atlascbs: a web server to map and explore chemico-biological space. J Comput Aided Mol Des 26:995–1003. https://doi.org/10.1007/s10822-012-9587-5
84. Wawer M, Lounkine E, Wassermann AM, Bajorath J (2010) Data structures and computational tools for the extraction of sar information from large compound sets. Drug Discov Today 15:630–639. https://doi.org/10.1016/j.drudis.2010.06.004
85. Maggiora GM (2006) On outliers and activity cliffs-why QSAR often disappoints. J Chem Inf Model 46:1535. https://doi.org/10.1021/ci060117s
86. Stumpfe D, de la Vega de León A, Dimova D, Bajorath J (2014) Advancing the activity cliff concept, Part II. F1000Research 3:375. https://doi.org/10.12688/f1000research.3788.1
87. Hu H, Bajorath J (2020) Increasing the public activity cliff knowledge base with new categories of activity cliffs. Future Sci OA 6:FSO472. https://doi.org/10.2144/fsoa-2020-0020
88. Stumpfe D, Hu H, Bajorath J (2020) Advances in exploring activity cliffs. J Comput-Aided Mol Des 34:929–942. https://doi.org/10.1007/s10822-020-00315-z
89. Maggiora G, Medina-Franco JL, Iqbal J, Vogt M, Bajorath J (2020) From qualitative to quantitative analysis of activity and property landscapes. J Chem Inf Model 60:5873–5880. https://doi.org/10.1021/acs.jcim.0c01249

90. González-Medina M, Méndez-Lucio O, Medina-Franco JL (2017) Activity landscape plotter: a web-based application for the analysis of structure-activity relationships. J Chem Inf Model 57:397–402. https://doi.org/10.1021/acs.jcim.6b00776
91. Guha R, VanDrie JH (2008) Structure-activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 48:646–658. https://doi.org/10.1021/ci7004093
92. Ritzén A, David L (2019) Physicochemical parameters of recently approved oral drugs. Success Drug Discov 4:35–53
93. Naveja JJ, Medina-Franco JL (2019) Finding constellations in chemical space through core analysis. Front Chem 7:510. https://doi.org/10.3389/fchem.2019.00510
94. Stumpfe D, Dimova D, Bajorath J (2016) Computational method for the systematic identification of analog series and key compounds representing series and their biological activity profiles. J Med Chem 59:7667–7676. https://doi.org/10.1021/acs.jmedchem.6b00906
95. Naveja JJ, Pilón-Jiménez BA, Bajorath J, Medina-Franco JL (2019) A general approach for retrosynthetic molecular core analysis. J Cheminf. https://doi.org/10.1186/s13321-019-0380-5
96. López-López E, Cerda-García-Rojas CM, Medina-Franco JL (2021) Tubulin inhibitors: a chemoinformatic analysis using cell-based data. Molecules. https://doi.org/10.3390/molecules26092483
97. Naveja JJ, Medina-Franco JL (2020) Consistent cell-selective analog series as constellation luminaries in chemical space. Mol Inf 39:e2000061. https://doi.org/10.1002/minf.202000061