Design and Diversity Analysis of Chemical Libraries in Drug Discovery

Dionisio A. Olmedo*[1,2], Armando A. Durant-Archibold[3,4], José Luis López-Pérez[5,6], José L. Medina-Franco*[7]

[1]Centro de Investigaciones Farmacognósticas de la Flora Panameña (CIFLORPAN), Apartado 0824-00178, Facultad de Farmacia, Universidad de Panamá, Ciudad de Panamá, Panamá. E-mail: ciflorp4@up.ac.pa
[2]Sistema Nacional de Investigación (SNI), Secretaria Nacional de Ciencia, Tecnología e Innovación (SENACYT), Ciudad del Saber, Clayton, Panamá.
[3]Centro de Biodiversidad y Descubrimiento de Drogas, Instituto de Investigaciones Científicas y Servicios de Alta Tecnología (INDICASAT AIP), Apartado 0843-01103, Panamá. E-mail: adurant@indicasat.org.pa
[4]Departamento de Bioquímica, Facultad de Ciencias Naturales, Exactas y Tecnología, Universidad de Panamá, Ciudad de Panamá, Panamá.
[5]CESIFAR, Departamento de Farmacología, Facultad de Medicina, Universidad de Panamá, Ciudad de Panamá, Panamá.
[6]Departamento de Ciencias Farmacéuticas, Facultad de Farmacia, Universidad de Salamanca, Avda. Campo Charro s/n,37071 Salamanca, España. E-mail: lopez@usal.es
[7]DIFACQUIM Grupo de Investigación, Departamento de Farmacia, Facultad de Química, Universidad Nacional Autónoma de México, Ciudad de México, Apartado 04510, México. E-mail: jose.medina.franco@gmail.com

*Address correspondence to this author: Centro de Investigaciones Farmacognósticas de la Flora Panameña (CIFLORPAN), Facultad de Farmacia, Universidad de Panamá, Apartado 0824-00178, Ciudad de Panamá, Panamá. Tel: +507-523-63070 and Grupo de Investigación DIFACQUIM, Departamento de Farmacia, Universidad Nacional Autónoma de México, Apartado 04510, Ciudad de México, México. Tel: +52 55 56223899, ext. 44458

Abstract

Chemical libraries and compound data sets are among the main inputs to start the drug discovery process at universities, research institutes, and the pharmaceutical industry. The approach used in the design of compound libraries, the chemical information they possess, and the representation of structures, play a fundamental role in the development of studies: chemoinformatics, food informatics, *in silico* pharmacokinetics, computational toxicology, bioinformatics, and molecular modeling to generate computational hits that will continue the optimization process of drug candidates. The prospects for growth in drug discovery and development processes in chemical, biotechnological, and pharmaceutical companies began a few years ago by integrating computational tools with artificial intelligence methodologies. It is anticipated that it will increase the number of drugs approved by regulatory agencies shortly.

Keywords: artificial intelligence; chemical libraries; chemoinformatics; chemical space; compound databases; natural products.

1. INTRODUCTION

Since the last century, research groups and commercial companies in the field of pharmaceuticals but also the agri-food industry and biotechnological products have been generating, collecting, and storing a large amount of chemical information in large chemical libraries [1–3]. These chemical libraries and collections of virtual compounds can be an exact reproduction of substances that exist in the scientist's laboratory or that can be acquired commercially, either natural products or synthetic compounds. Chemical libraries can also contain feasible virtual molecules according to simple chemical stability and synthetic feasibility rules generated by enumeration algorithms. The latter databases have the advantage of being able to synthesize substances on demand in case of need, for example after a virtual high-throughput screening.

Chemical libraries of real compounds are characterized by a large structural diversity. They contain not only their structural data and stereochemical information, but also their associated physicochemical, spectroscopic, and spectrometric properties. Some of this chemical information has been used in the drug discovery process [2, 3].

Chemical databases are very heterogeneous in terms of size, types of information, and organization. This means that the information stored in them on the properties of the substances is not the same, nor repeatable from one library to another due to the type of format in which the data is filed. This causes researchers to follow different guidelines or workflows in the design of chemical libraries, for which there are multiple facets in their scope of interpretation, analysis, and application in different computational studies [3, 4].

In recent years, we have witnessed an explosion in the number of chemical databases with quite different purposes. In addition to corporate, private, and commercial databases, there are many freely accessible databases on the web. These databases, in addition to containing a wealth of chemical information, are equipped with chemoinformatics and bioinformatics tools that enhance their usefulness in drug discovery processes. [5]. **Table 1** summarizes some companies that offer the service of designing virtual chemical libraries and syntheses on demand.

Name	Services	Reference
ChemDiv 3D-Biodiversity Library	Virtual Library on-demand. Stock 1.6M screening compounds and 75K building blocks.	https://www.chemdiv.com/catalog/sets/3d-biodiversity-library/. Accessed February, 03,2022
Greenpharma	This has developed synthesis routes and offers chemical services and solutions to develop focused libraries based on scaffolds and synthesis is made on-demand.	https://www.greenpharma.com/home/. Accessed February, 03,2022
TargetMol	This offers over 120 types of compound libraries, 10,000 small molecules, and 16,000 natural products. Chemical synthesis, virtual screening, pharmacophore-based virtual screening, and molecular docking-based virtual screening on demand.	https://www.targetmol.com/home/. Accessed February, 03,2022
Enamine	This is a leader in the market of building blocks on demand. Stock 210 million of novel building blocks make on demand. Screening dataset of 2.9 million small molecules.	https://enamine.net/home/. Accessed February, 03,2022

Table 1 Examples of chemical companies that offer services of construction of virtual libraries and on-demand synthesis.

2. DESIGN OF CHEMICAL LIBRARIES

Chemical libraries can be generated using an approach based on selecting data subsets from large libraries and performing a partition and selection of various data subsets. Another way of designing chemical libraries is by constructing compound libraries based on a set of chemical fragments, which can be obtained through retrosynthetic analysis of a dataset of original compounds [6].

Different parameters have been used in the development of chemical libraries based on diversity [7–10]. Some of the concepts and representations used to generate these libraries are based on physicochemical properties (drug-like or lead-like) [11–13], 2D descriptors and molecular fingerprints (14), chemical space [15–17], molecular shape [18, 19], and pharmacophore models [20].

Virtual chemical libraries can be generated using a scheme of known synthetic reactions and reagents available [2, 4], molecular graph model [21], diversity-oriented synthesis library [22–24], target-focused libraries [25–26], and chemical libraries designed *de novo*.

A wide range of artificial intelligence (AI) techniques have been used in the design of de novo libraries to obtain various libraries of organic compounds [27–32]. Many companies have generated organic compound libraries focused on low molecular weight molecules for a wide range of computational chemistry applications [33].

Target-focused libraries have made it possible to obtain a broad variety of chemical libraries developed "in-house" by commercial companies against multiple biological targets. These types of libraries currently have a remarkable interest in drug discovery and development processes. An example is the company TargetMol Chemicals Inc. [3].

On-demand virtual chemical libraries contain a certain number of compounds synthesized by well-known chemical reactions and carried out in a stock building block and that are tailored to a specific biological target [34] Within the virtual chemical libraries, one of the most important is the in-house compound libraries (generated by universities, institutes, and chemical companies), which have had an exponential increase in their use and number of citations by the scientific community, since the beginning of the COVID19 pandemic.

A few libraries of organic compounds with $^1H/^{13}C$ spectroscopic [35–38] or spectrometric (MS) [39–41], information that can be of great utility in metabolomics and replication during the structural elucidation processes of natural products have been developed.

2.1 Data information in chemical libraries

The chemical libraries contain distinct types of compounds that, depending on their primary source and/or obtention method, can be classified as synthetic or semi-synthetic compounds, natural products, and virtual compounds. The properties of these several types of compounds differ significantly. In general, naturally occurring substances, with a larger number of stereocenters, are more complex than synthetic ones. This complexity can lead to the development of more selective pharmaceuticals. In fact, in some pharmacological groups, such as anti-tumor drugs, many of the pharmaceuticals are of natural origin or based on them. In addition to structural information, these databases may contain information on their activity in different pharmacological domains. **Figure 1** shows the types of compounds included in the different chemical libraries, while **Figure 2** summarizes the data information existing in the same.

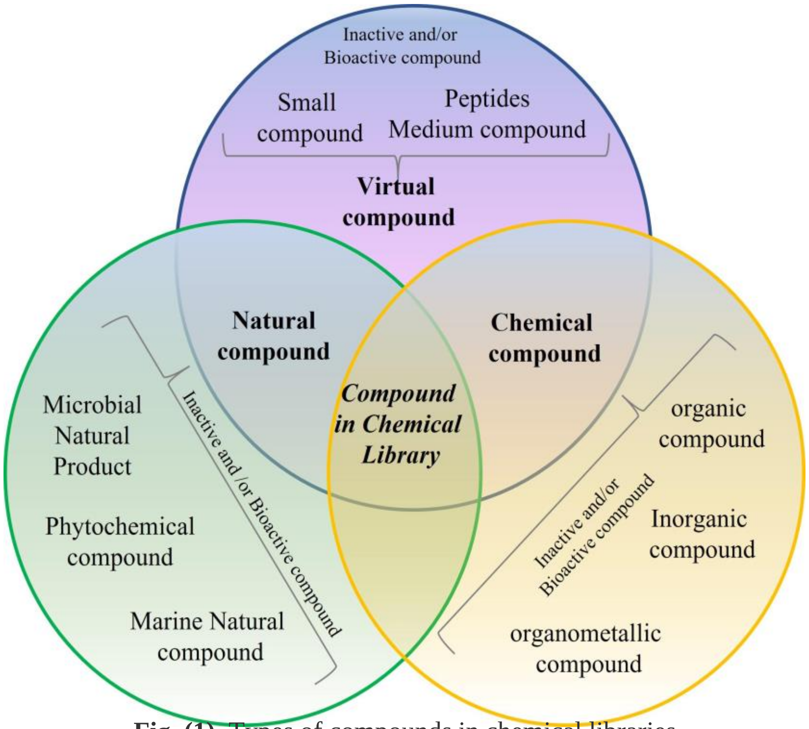


Fig. (1). Types of compounds in chemical libraries

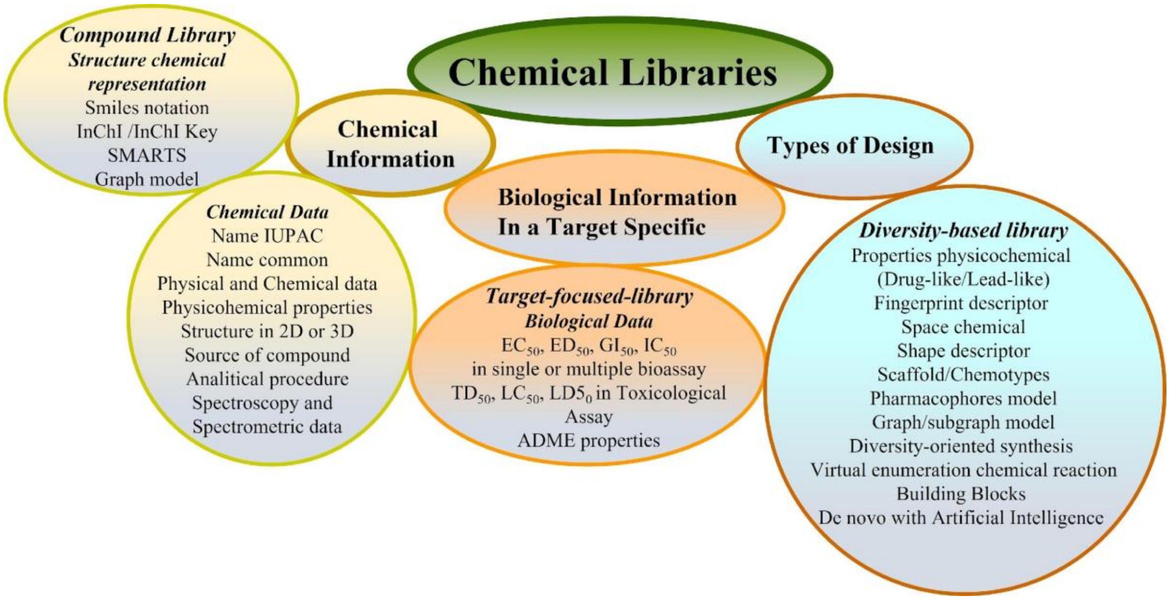


Fig. (2). Overview of data information in chemical libraries.

2.2. Structural representation in chemical libraries

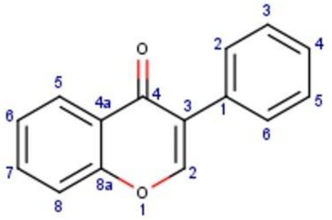
Chemical libraries have implemented different linear notation systems to describe 2D molecular models. Examples of linear notations are SMILES, SMARTS, InChI, InChI Key, and others.

2.2.1 SMILES string

Simplified Molecular Input Line Entry Specification (SMILES) [42–46] is a straightforward coding system for unambiguously describing molecular structures using alpha-numeric character strings in which atoms are represented by their atomic symbols and bonds are only represented when they are multiple in the same way as written according to IUPAC standards. Branches are specified in parentheses and/or square brackets. Most molecular editors can interpret these SMILES codes for conversion into two-or three-dimensional graphical molecular models [47–50]. There is a form of canonical representation, which is the one used in computational studies since it "guarantees to have unique molecules," facilitating rapid data mining in chemical library databases. SMILES is not only used for the representation of molecular models but also for performing similarity searches, in which a comparison of the physicochemical properties of molecules is carried out. SMILES is, therefore, useful in drug design studies based on physicochemical properties [11–13], molecular fingerprint [14], chemical space [15–17], and molecular scaffold [48]. Weininger, A., and Weininger, D. developed the original SMILES specification in the late 1980s and early 1990s [43, 44, 51]. In 2007, an open standard called "OpenSMILES" was developed by the open-source chemistry community.

2.2.2. SMARTS string

SMARTS (SMILES Arbitrary Target Specification) is an extension of SMILES useful for performing a search by substructure to find structural fragments paired with other matching substructures [46, 52]. SMARTS has applications in studies based on the search for structural fragments in chemical libraries, such as virtual screening and molecular docking studies. In addition, SMARTS notation has been used to determine Pan Assay Interference Compounds (PAINS) in chemical libraries for high throughput screening [52]. **Figure 3** shows SMILES and SMART representations of a natural product (benzopyranone) present in chemical libraries.



IUPAC Name	3-phenyl-4*H*-1-benzopyran-4-one
Smiles ACDChemSketch$_1$	O=C1c2ccccc2OC=C1c1ccccc1
Smiles ChemAxon$_2$	O=C1C(=COC2=C1C=CC=C2)C1=CC=CC=C1
Smiles JME editor$_3$	O=c2c(c1ccccc1)coc3ccccc23
Smiles JME editor$_3$	C=13(C(=CC=CC=1)C(C(C=2(C=CC=CC=2))=CO3)=O)
Smiles PubChem editor$_4$	C1=CC=CC2=C1C(C(=CO2)C3=CC=CC=C3)=O
SMART PubChem editor$_4$	c1cccc-2c1-[#6](-[#6](=[#6]-[#8]-2)-c3ccccc3)=[#8]

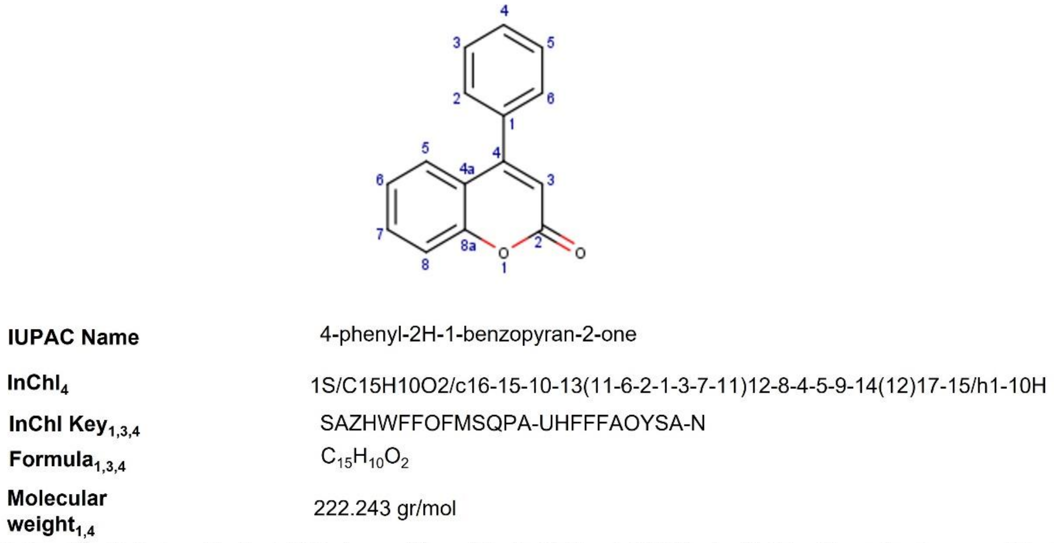
$_1$ACD/ChemSketch, version 2021.1.0, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, 2022.; $_2$MarvinSketch, version 21.11.0, ChemAxon Ltd., Budapest, Hungary. Www.chemaxon.com, 2022.; $_3$nJME editor 2017_11_16 http://www.cheminfo.org/flavor/malaria/Utilities/SMILES_generator___checker/index.html. Accessed Feb 16, 2022. ; $_4$PubChem Sketcher V2.4. https://pubchem.ncbi.nlm.nih.gov//edit3/index.html. accesed Feb 15, 2022

Fig. (3). SMILES and SMART notation of the 3-phenyl-benzopyran-4-one core using different structure editors. Here it is observed that the representation of SMILES presents a different ordering of the characters in its linear notation, while the SMART better defines the structure, avoiding errors when using different structural editors. This difference is explained by the fact that in structural editors, for example, aromatic rings are represented in upper case (when they should be lower case), and this occurs because aromaticity has not been considered, and benzene is considered as tricyclohexene.

2.2.3 InChI and InChI Keys Identifier

InChI is the International Chemical Identifier developed by the International Union of Pure and Applied Chemistry (IUPAC) in collaboration with the U.S. National Institute of Standards and Technology (NIST) and the InChI Trust [53–56]. The InChI establishes a unique label for each chemical compound, facilitating the linking of diverse data compilations. This linear notation system resolves the inconvenient and chemical ambiguities of the SMILES language about stereocenter, tautomer, and valence [53–56].

The InChI Key is a fixed-length (27-character) condensed digital representation of an InChI, developed to make it easy to perform web searches for chemical structures. The first block of 14 characters for an InChI key encodes the core molecular constitution, as described by a formula, connectivity, hydrogen positions, and charge sublayers of the InChI main layer [53-56]. The other structural features complementing the core data, namely exact positions of mobile hydrogens, stereochemical, isotopic, and metal ligands, whichever are applicable, are encoded by the second block of InChI Key. (The InChI Key is described in detail here (https://www.inchi-trust.org/) [53–56]. **Figure 4** show an example of InChl and InChl Key representation of organic compound.



IUPAC Name	4-phenyl-2H-1-benzopyran-2-one
InChl$_4$	1S/C15H10O2/c16-15-10-13(11-6-2-1-3-7-11)12-8-4-5-9-14(12)17-15/h1-10H
InChl Key$_{1,3,4}$	SAZHWFFOFMSQPA-UHFFFAOYSA-N
Formula$_{1,3,4}$	$C_{15}H_{10}O_2$
Molecular weight$_{1,4}$	222.243 gr/mol

[1]ACD/ChemSketch, version 2021.1.0, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, 2022.; [2]MarvinSketch, version 21.11.0, ChemAxon Ltd., Budapest, Hungary. Www.chemaxon.com, 2022.; [3]nJME editor 2017_11_16 http://www.cheminfo.org/flavor/malaria/Utilities/SMILES_generator___checker/index.html. Accessed Feb 16, 2022. ; [4]PubChem Sketcher V2.4. https://pubchem.ncbi.nlm.nih.gov//edit3/index.html. accesed Feb 15, 2022

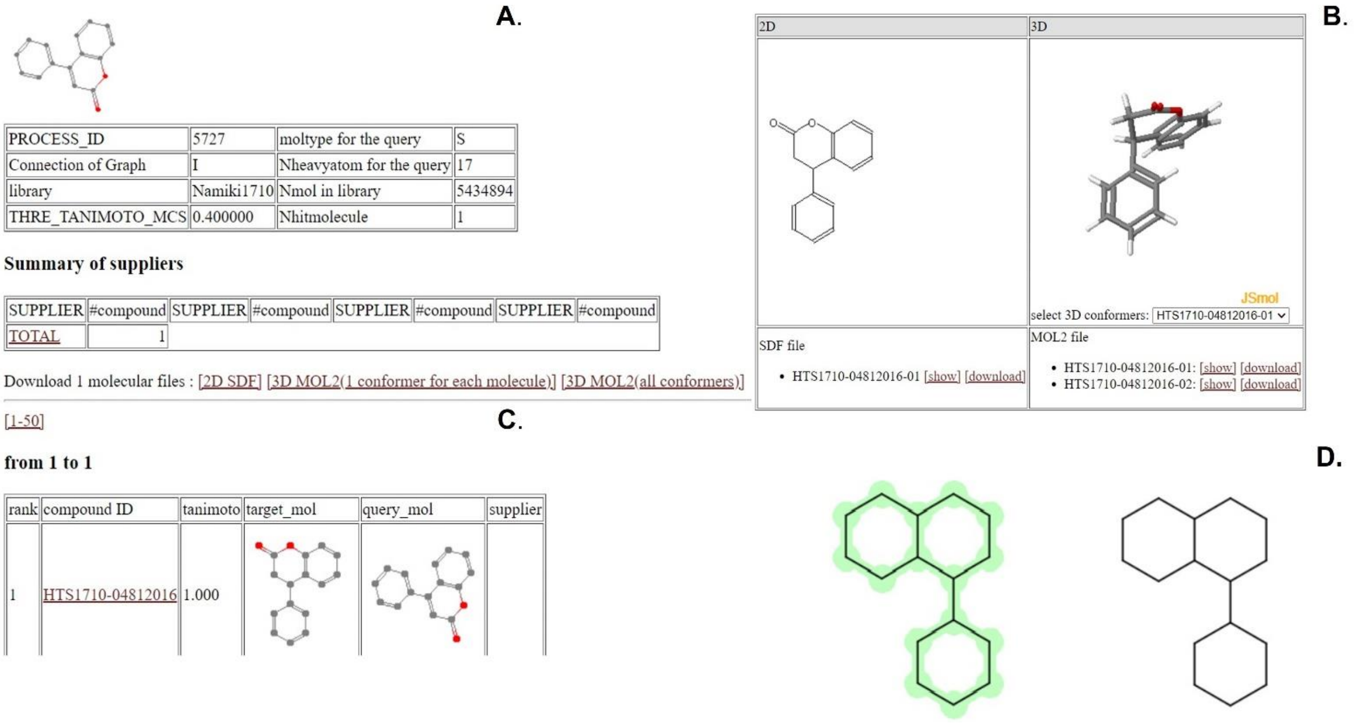
Fig. (4). InChl and InChl Key representation of the 4-phenyl-benzopyran-4-one using different structure editors. This notation lineal defines better the structure of a chemical compound.

2.2.4. Graph model

Chemical graphs

In the graphical representation of a chemical structure, the vertices represent the atoms, while the edges represent the bonds, and the order of the bonds corresponds to the multiplicity of edges. This graphical representation of vertices and edges describes a "chemical graph". The maximum number of bonds that an atom can form has been determined by the valence of the chemical elements involved in bond formation [57, 58].

The graph model has been utilized in analyzing based-fingerprint [59], based scaffold [60], machine learning (ML) [61], deep learning methods [62], AI [63], a fragment-based model for the construction of building blocks and reaction-based *de novo* design on demand [64]. The graph-based representations (**Figure 5**) were generated with the Kcombu program [65].



PROCESS_ID	5727	moltype for the query	S
Connection of Graph	I	Nheavyatom for the query	17
library	Namiki1710	Nmol in library	5434894
THRE_TANIMOTO_MCS	0.400000	Nhitmolecule	1

Summary of suppliers

SUPPLIER	#compound	SUPPLIER	#compound	SUPPLIER	#compound	SUPPLIER	#compound
TOTAL	1						

Download 1 molecular files : [2D SDF] [3D MOL2(1 conformer for each molecule)] [3D MOL2(all conformers)]

[1-50]

from 1 to 1

rank	compound ID	tanimoto	target_mol	query_mol	supplier
1	HTS1710-04812016	1.000			

2D	3D
	select 3D conformers: HTS1710-04812016-01
SDF file - HTS1710-04812016-01 [show] [download]	MOL2 file - HTS1710-04812016-01: [show] [download] - HTS1710-04812016-02: [show] [download]

Fig. (5). Graph model generated by the Kcombu program. The graph model of 4-phenyl-benzopyran-4-one is observed in (**A**). In a model of 4-phenyl-benzopyran-4-one (**B**), there are 2D and 3D representations of the compound. The compound was identified with a similarity coefficient of Tanimoto of 1.00 in the Nimiki710 database (**C**). The Bemi-Murcko skeleton and carbon are highlighted in green (**D**).

SMILES notation is the main type of linear representation used to describe the structures of compounds in 2D models in chemical entity libraries. It allows easy recognition of the groups of atoms, bonds, and connections established in a molecule. While SMILES and other linear notations are easily understood by machines, molecular graphs are the most amenable to chemists.

The main drawback of the design based on the graph model is related to the generation of the vectors that constitute the graph representation. Mercado *et al* [63] have established a model for molecular design based on graphs using neural networks and the canonical SMILES as the unit of departure. This robust work provides a perspective with immense potential for de *novo* design that can significantly impact drug discovery and development programs in pharmaceutical companies [64].

2.3. Progress on *de novo* design in chemical libraries

De novo drug design is a term used in medicinal chemistry that refers to the generation of chemical compounds using mathematical equations with the support of computational tools [66–68]. The inclusion of methodologies linked to AI has allowed considerable progress in drug design in the chemical and pharmaceutical industries. The ReLeaSE application for de novo design of new chemical compounds is based on the approaches of deep learning (DL) [66–69] and reinforcement learning (RL) [70]. For the design of virtual libraries, the ReLeaSe employs the RL algorithm, which takes into account physicochemical properties, specific biological activity, and chemical complexity [70]. DeepLigBuilder is a deep generative model for building 3D molecules that can be used in chemical library design. These generated structures of molecules and analogs are based on the physicochemical properties of interest in drugs discovery [70]. AutoGrow4 is a powerful computational hit discovery and lead optimization software with a greater application when the site of interaction on the biological target with ligands is unknown [71].

De novo drug design of multi-targeted chemical libraries based on AI [72] can be used in the process of drug discovery and development in industry and academia. This design based on multi-target allows the generation of new molecular entities with optimized drug-like properties and similarity-based constraints biasing specific biological targets. In addition, this *de novo* approach to chemical compound library design could be applied and strengthened in combination with other hybrid strategies based on conventional methods and AI [72].

Reaction-based *de novo* design refers to the in-silico generation of new chemical structures by combining reactants through structural transformations derived from known reactions. The implementation and validation of the model using a multi-marker reaction class recommender were developed *de novo* with a design based on a reaction vector combined with a reaction class recommender that considers the characteristics of the entire molecule as input molecule for suggest only those classes of synthetically accessible reaction that are most likely to occur in an environment chemical determined [73].

Diversity-oriented synthesis (DOS) is a synthetic strategy that aims to efficiently produce compound collections with elevated levels of structural diversity, and three-dimensionality and is therefore well-suited for the construction of novel fragment collections, demonstrating the utility of DOS within drug discovery efforts [74].

The ReFRAME is an example of a drug repurposing library that has contributed to drug discovery in academia [75].

The approach developed (ML), deep learning (DL), reinforcement learning (RL), and AI, in the generation of algorithms, scripts, programs, workflow schemes, and online web services to generate chemical structures in 2D or 3D, has had a significant impact and utility in the design of virtual chemical libraries. This has shortened the time in the drug discovery process in academia and pharmaceutical industries [76].

In summary, the incorporation of methodologies such as based-reaction, multi-marker reaction class recommender, diversity-oriented synthesis-based design, synthesis-based building block, and drug repositioning library has increased the use of computational tools. The use of these methodologies has led to an increase in the complexity of structural representation systems for which graph models have limitations in their generation, so that linear notation systems, such as SMILES, remain the means of choice for generating molecules in chemical libraries.

3. ANALYSIS OF CONTENTS AND DIVERSITY IN CHEMICAL LIBRARIES

Systematic analysis of the types of chemical structures contained in a compound library and their diversity is a basic practice for the rational design and use of such libraries. Indeed, the diversity of the analysis is a major criterion to identifying bioactive compounds and it can be applied for multiple purposes, such as design, acquisition, and selection of compounds for screening (virtual and experimental), and analysis of structure-activity relationships. Diversity analysis is also incorporated into *de novo* design strategies to evaluate the structural novelty of chemical libraries [77]. Furthermore, diversity analysis has applications not only in drug discovery but also in natural product research, food chemistry, organic chemistry, and material sciences, among other areas.

The contents and diversity analysis will depend on the compound representation, discussed in Section 2. In general, the representations can be grouped into two major categories: molecular scaffolds and structural fingerprints. The selection of the compound representation to study a chemical library gives rise to the definition of the “chemical space” and will depend on the goals of the research program. In this section, we briefly discuss the concepts of chemical space and methods used to study systematically the diversity, variety of chemical scaffolds, and diversity of fingerprints. Several reviews address the basics and early developments of the analysis of the contents and diversity analysis of chemical space of compound libraries [78]. As Dunn et al. [77, 79] recently pointed out, the rapid growth in the number and size of chemical libraries required the development of computational methods to analyze the heterogeneousness of ultra-large chemical libraries [80–81].

3.1 Diversity of chemical space

Chemical space is a core concept in chemoinformatics [82], which refers to all molecules as well as multi-dimensional conceptual space. In contrast, the universe includes all types of matter and energy; galaxies; solar systems; and all the contents of the space that could be called cosmic space, which means the entire universe. Unlike cosmic space, chemical space is dependent on the structural representation and the type of properties or descriptors used to represent the compounds of interest. For example, the types of descriptors employed to represent small organic molecules will be different than those describing organometallic molecules, and both constitute a fraction of the chemical space [83].

Progress on chemoinformatics methods to explore systematically the chemical space using quantitative and visual methods has been reviewed [84]. Despite the considerable progress in the methodologies to study the chemical space (and biological spaces as well), it remains one of the grand challenges in computer-aided drug design [85].

Several published studies focus on the use of chemical space as a tool for assessing the diversity of data sets and exploring the relationships between compound collections. For instance, the chemical space of nearly 43,000 naturally occurring compounds from Latin America and other geographical regions has been analyzed. It was concluded that the natural products data sets occupy similar regions in the chemical space and are an attractive source for obtaining novel leaders able to become new pharmaceuticals [86]. These observations are consistent with previous studies that showed that natural products cover a broader region of the chemical space as compared to approved drugs and synthetic compounds, and they populate areas of the chemical space that are difficult to synthesize [87].

Another recent exploration the chemical space of large and ultra-large chemical libraries is represented by the study of Dunn *et al.* [77]. In that work, the authors introduced the Chemical Library Networks (CLNs) as a general and efficient approach to representing visually the chemical space of chemical libraries. Dunn *et al*, [77] exemplified the use of the CLNs to analyze the diversity of 19 compound data sets commonly used in natural product research and drug discovery, containing more than 18 million molecules. Notably, the code used to create the CLNs is freely available and was used in a separate study to visually represent the chemical space of 11 synthetic compound libraries focused on epigenetic targets containing over 50,000 molecules [88]. A survey of de novo virtual libraries [87, 89–91], peptides [92], and food chemicals [93, 94] is another representative diversity analysis of compound libraries in the chemical space.

3.2 Diversity based on molecular scaffolds

Another strategy to characterize the contents and diversity of compound databases is the use of molecular scaffolds or chemotypes i.e., the central or main core structure of a molecule [95]. Like physicochemical properties, molecular scaffolds can be interpreted straightforwardly and facilitate communication within research groups from different disciplines. The concept of scaffold is associated with "scaffold hopping" and "privileged structures". Scaffold content analysis is frequently used to compare compound databases, uncover novel scaffolds in a compound data set, analyze the SAR of sets of molecules with measured activity, and analyze the SAR of sets of molecules with measured activity [96].

Quantifying and comparing the scaffold diversity of compound libraries depends on many variables, such as the specific method used to generate the scaffolds, the number of compounds in the database, and the specific distribution or frequency of the molecules in those scaffold classes. Often, scaffold diversity is measured based on frequency counts. While these metrics are correct in the way they are defined, they do not provide enough information related to the specific distribution of the molecules across the different scaffolds, particularly the most populated ones. An entropy-based metric has been proposed to measure the distribution of the molecules across different scaffolds, particularly the most populated ones, as a complementary metric for the comprehensive scaffold diversity analysis of compound data sets [96].

Scaffold analysis of natural products databases from various sources has been published recently. For instance, Núñez MJ *et al.* [85] reported the scaffold diversity of 15 natural product databases. For that work, the authors used the scaffold definition of Bemis-Murko, and the scaffold diversity was measured using cyclic system recovery curves and scaffold counts. It was concluded that the collections of natural products from Brazil (in the NUBBE database) and the Panamanian flora (UPMA) are the most diverse. In contrast, the collections such as AfroDB, BIOFACQUIM, and other collections of natural products from commercial sources were the least diverse.

Bhurta and Bharate recently reported on the scaffold diversity of cyclin-dependent kinase (CDK) inhibitors, which are one of the major drug targets. The authors analyzed CDK inhibitors under preclinical and clinical development and found that the amino-pyrimidine framework is the most represented scaffold [97].

3.3 Diversity based on molecular fingerprints

Molecular fingerprints are an alternative approach to describing chemical structures systematically and quantitatively. One of the main advantages of molecular fingerprints is that they can be calculated extremely fast, being suitable for handling small -to- large and ultra-large compound databases containing millions of chemical structures.

The great versatility of molecular fingerprints in chemoinformatics has increased their usefulness in the computer-aided drug design process [98]. Fingerprints are particularly useful for quantifying molecular similarity, during database searching or clustering processes, and in the development of classification or predictive models [98]. Molecular diversity is usually quantified using a similarity coefficient, which is typically, although not necessarily the most adequate in all cases, the Tanimoto coefficient [99].

Like other molecular representations, the type of fingerprint used for a particular application will define the chemical space. To reduce the dependence of the similarity assessment on the specific type of molecular fingerprint used, it has been proposed to use the combination of multiple molecular fingerprints, merging the metrics into a consensus measure [100].

Although there are many well-established molecular fingerprints, the development of novel and improved fingerprints is an area of continued research [101]. In recent applications, molecular fingerprints have been employed to quantify the structural diversity of 19 compound libraries used in drug discovery and natural product research, containing more than 18 million molecules (*vide supra*) [101, 102]. In that work, a newly introduced similarity metric, extended by Tanimoto in combination with the RDKit fingerprints, showed the best performance to represent the chemical libraries. Molecular fingerprints have also been used broadly to quantify the molecular diversity of natural product databases such as natural product collections from Latin America and other geographical regions, propolis components, and other data sets [103].

4. APPLICATION OF COMPOUND LIBRARY

The use of chemical libraries in the computational analysis includes areas of chemoinformatics, bioinformatics, ADMET (absorption, distribution, metabolism, excretion, and toxicity properties), and PAINS alerts. These are developed in the following sections.

4.1 Cheminformatics

Cheminformatics is based on computational chemistry programs for the acquisition, analysis, and visualization of chemical structure data sets [80–81, 87, 89–92, 95, 96, 99, 102-103]. Section 3 described the most common metrics used in chemoinformatics studies. This discipline includes the management of biochemical or biological information from experimental data [104] and spectral data [105]. On the other hand, chemoinformatics is a fundamental tool in biochemical science research and biomedical areas related to biological systems, metabolomics, proteomics, and chemogenomics applied to drug discovery and development [106–108]

4.2. Bioinformatics

Bioinformatics involves using computational analysis of biological data, which is relevant in pharmacological, biological, and biomedical studies [109]. In drug discovery, bioinformatic helps to evaluate *in silico* the interaction of cellular components, tissues, peptides, proteins, ADN, ARN, antibodies with endogenous metabolites, chemical compounds, and tissular fragments (monoclonal antibody, antitoxin) to predict how these biological systems are affected because of these molecular interactions [110].

This research field is strongly leading to drug discovery and development (DDD) [111], genome analysis, with an approach to personalized medicine for the development of a drug, and the optimization of biological targets [112] using protein-protein simulation, homology models, and molecular docking and dynamic tools with chemical libraries [113].

Incorporating AI methodologies in the development of chemoinformatics and bioinformatics tools can increase their usefulness in biomedical, biological, and pharmaceutical sciences, thus strengthening drug discovery and development programs.

4.3 ADMET and toxicological properties

In recent years, the growth of *in silico* studies to evaluate pharmacokinetic and toxicological properties has taken on a relevant role in the drug discovery and development process in the pharmaceutical industry. The incorporation of methodologies based on AI in the generation of *in silico* predictive models, for the determination of ADMET properties in real compound libraries, virtual chemical library, chemical libraries databases, and combinatorial chemistry libraries, has allowed the development of software standalone and online web platforms that promote the study of ADMET *in silico* [114–115].

The incorporation of quantitative structure-activity relationship (QSAR) methods in ADMET predictive studies has meant a great advance in the process of drug discovery and development in the pharmaceutical industry [116]. ADMET studies have been conducted on libraries designed with different approaches, for example: fragment-based drug discovery (FBDD) [117], natural products [118], small molecules [119], using the SMILES notation [117–118], SMARTS [119], and graph model [120].

The existence of numerous computational tools to predict ADMET properties in chemical compounds constitutes one of the main limitations of these *in silico* methods, as pointed out by Kar, S., and Leszczynski, J. [121], in their extensive work on *in silico* tools to predict the ADMET profile.

These authors conclude that depending on the metrics used (physicochemical properties (drug-like/lead-like), ADMET predictive properties, and the approach of the designed model) to generate these data, a great variation is observed in them due to the software, program, or web server used. This is in addition to the fact that predicting the properties of ADMET is an increasingly complex challenge for researchers, where obtaining reliable and reproducible values becomes a growing obstacle due to the exponential increase in the algorithms developed to predict these properties.

In general terms, it is recommended to use several *in silico* tools and to agree on the properties of ADMET with the available methodologies to minimize the statistical variations of the model used.

Some examples of web platforms are shown in **Table 2**.

ADMET predictor	Description	Source or Reference
PreADMET	is a web-based application for predicting ADMET data and building a drug-like library using *in silico* method	https://preadmet.webservice.bmdrc.org/
ADMET Predictor® 10.3	is the flagship ML platform for ADMET modeling and AI.	https://www.simulations-plus.com/
ADMETlab 2.0	an integrated online platform for accurate and comprehensive predictions of ADMET properties.	[122]
FPADMET	which is a repository of molecular fingerprint-based predictive models for ADMET properties [123].	https://gitlab.com/vishsoft/fpadmet
ProTox-II	a web server for the prediction of toxicity of chemicals.	[124]

Table 2. Example of web services to predict ADMET properties.

4.4. Structural alerts

PAINS (Pan-Assay INterference compoundS) are compounds that influence the interpretation of bioassay results [125] by interacting with single biological targets [126] or multiple biological targets [127], resulting in bioactive compounds with a high potential for optimization in the drug discovery and development process [127]. These interfering compounds have been called artifacts, promiscuous compounds, or false positives in cellular, biochemical, or pharmacological assays [128], which have frequently appeared when using libraries for high throughput screening (HTS), libraries based on fragments or small molecules (FBDD) [129], The Natural Product Library (NPL) [130], and many other chemical library databases or real compound libraries.

PAINS can be either synthetic, semi-synthetic, or natural compounds that have triggered biological alterations: oxidation by redox mechanisms, covalent interaction with proteins, metal chelation, and alteration of the lipid layer of cell membranes. These disturbances affect biological assays by interference with fluorescence and structural decomposition of the tested compounds [131], making the identification of these interfering compounds in chemical libraries a difficult problem [132–134].

To recognize the PAINS, present in compound libraries, several computational tools, and AI methodologies have been developed to filter, remove, or eliminate these artifacts to avoid their effects on pharmacological trials and molecular modeling studies [135]. However, the structural fragments present in these promiscuous compounds have shown variable biological activities in biological tests [136, 137]. Platforms based on substructure filters have been commonly used to identify and eliminate these interfering compounds [138]. Some platforms used are: AlphaScreen technology [139, 140], and PrePeP [141]. In addition, *in-house* methodologies have been developed for certain PAINS with the KNIME [142] software, the OpenEye [143] chemoinformatics tools, and the R and RStudio applications, which use the Java and R programming languages [16, 144].

Computational methods, including AI, are powerful tools for the detection of interfering compounds *in vitro* biological assays, as well as for virtual molecular modeling assays [145–147].

5. PERSPECTIVE AND FUTURE DIRECTION

The design of libraries of chemical compounds and chemical library databases has increased with the incorporation of AI in the chemical industry, while in the pharmaceutical industry it has led to great advances in the drug discovery and development stages. Medicines developed with the help of AI are expected to be on the market soon.

The development of chemical libraries in the Latin American region started a few years ago. They are focused on natural products, containing initially between 196 and 485 (El Salvador and Panama databases) [85] plant metabolites, but have been enriched with the inclusion of metabolites from fungi, bacteria, and marine organisms, highlighting the BIODIFACQUIM and NUBBE databases developed at universities in Mexico and Brazil. [83, 85, 89]. Because of the wide range of chemotypes in these compound libraries, they have a high structural diversity and molecular complexity. They have been evaluated through chemoinformatics studies and have a high potential for usefulness in drug discovery for emerging and re-emerging diseases affecting the population of Latin America.

The high biodiversity of these countries means that these databases may increase significantly, such as in the case of Panama, where research on natural products has expanded to include snakes and amphibians; land fungi; land microorganisms (bacteria, endophytic fungus); marine microorganisms; and marine microorganisms in symbiosis with corals and sponges, among other areas of increasing research in our country. Additionally, NAPROC-13 databases contain ^{13}C NMR spectroscopic data and collect most of the natural products isolated in Panama and El Salvador [38].

These initiatives are supported by the national agencies that provide funding.

CONSENT FOR PUBLICATION

D. A. O. A. Participated in the conceptualization, creation, execution, and methodological design of the review research and obtained funds. He also actively collaborated in writing, reviewing, and editing the manuscript. A. A. D. A. Collaborated in the writing, review, and edition of the manuscript. J. L. L. P. has worked in the review of the methodological design. He also participated in the writing, review, and edition of the manuscript. J. L. M. F. has worked in the conceptualization of research, reviewing the methodological design. Also, he participated in the writing, review, and edition of the manuscript.

FUNDING

The Vice-Rectory for Research and Postgraduate Studies of the University of Panama finances this work through the grant number VIP 14-01-2019-05.

CONFLICT INTEREST

The authors declare that there are no financial or commercial conflicts of interest.

ACKNOWLEDGMENT

This research is part of the International Collaborative Project: The Rise of Natural Product Libraries in Latin America and Computer-Aided Drug Discovery. D. A. O. A. and A. A. D. A. and J.L.L-P acknowledge the National System of Research of Panama (SNI) for supporting the development of scientific research.

REFERENCES

[1]. Sarker, S.D.; and Nahar, L. Application of Computation in Building Dereplicated Phytochemical Libraries In *Computational Phytochemistry*, Sarker, S.D.; Nahar, L., 1st Edition Eds.; Elsevier: Amsterdam, **2018**; pp 141–163.

[2]. Walters W.P. Virtual Chemical Libraries. *J. Med. Chem.* **2019**, *62*(3), 1116–1124.

[3]. Targetmol. **targetmol.com/all-compound-libraries.html** (targetmol.com/all-compound-libraries.html) (Accessed January 29, **2022**)

[4]. van Hilten, N.; Chevillard, F.; Kolb, P. Virtual Compound Libraries in Computer-Assisted Drug Discovery. *J. Chem. Inf. Model.*, **2019**, *59*(*2*)，644−651.

[5]. Ghani S.S.A comprehensive review of database resources in chemistry. *Eclética Química* **2020**, *45*(3), 57–68.

[6]. de la Vega de León, A.; Lounkine, E.; Vogt, M.; Bajorath, J. Design of diverse and focused compound libraries In *Tutorials in Chemoinformatics*, First Edition, Varnek A, Ed.; John Wiley & Sons Ltd: New Jersey, **2017**, pp 85–101.

[7]. Koutsoukas, A.; Paricharak, S.; Galloway, W.R., Spring, D.R.; Ijzerman, A.P.; Glen, R.C.; Marcus, D.; Bender, A. How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J. Chem. Inf. Model.*, **2014**, *27*, *54*(1), 230–242.

[8]. Petrone P.M.; Wassermann A.M.; Lounkine E., Kutchukian, P., Simms, B.; Jenkins, J.; Selzer, P.; Glick, M. Biodiversity of a small molecules-a new perspective in screening set selection. *Drug Discov. Today.*, **2013**, *18*(13-14), 674–680.

[9]. Shelat, A.A.; and Guy, R.K. Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.*, **2007**, *3*(8), 442–446.

[10]. Fitzgerald, S.H.; Sabat, M.; Geysen, H.M. Diversity space and its application to library selection and design. *J. Chem. Inf. Model.*, **2006**, *46*(4), 1588–1597.

[11]. Pascolutti, M.; and Quinn, R.J. Natural products as lead structures: chemical transformations to create lead-like libraries. *Drug Discov. Today*., **2014**, *19*(3), 215–221.

[12]. Camp, D.; Davis, R.A.; Campitelli, M.; Ebdon, J.; Quinn, R.J. Drug-like Properties: Guiding Principles for the Design of Natural Product Libraries. *J. Nat. Prod*., **2012**, *75*(1), 72-81.

[13]. Butler, M.S.; Fontaine, F.; Cooper, M.A. Natural product libraries: assembly, maintenance, and screening. *Planta Med.*, **2014**, *80*(14), 1161–1170.

[14]. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model*., **2010**, *24*, *50*(5), 742–754.

[15]. Díaz-Eufracio, B.I.; Palomino-Hernández, O.; Arredondo-Sánchez, A.; Medina-Franco, J.L. D-Peptide Builder: a web service to enumerate, analyze, and visualize the chemical space of combinatorial peptide libraries. *Mol. Inform.*, **2020**. 39, e 20200 0035

[16]. Saldívar-González, F.I.; Huerta-García, C.S.; Medina-Franco, J.L. Chemoinformatics-based enumeration of chemical libraries: a tutorial. *J. Cheminform.*, **2020**, *12*, 64.

[17]. Medina-Franco, J.L.; Martinez-Mayorga, K.; Meurice, N. Balancing novelty with confined chemical space in modern drug discovery. *Expert Opin. Drug Discov*., **2014**, *9*(2), 151–165.

[18]. Skalic, M.; Jiménez, J.; Sabbadin, D.; De Fabritiis, G. Shape-Based Generative Modeling for de Novo Drug Design. *J. Chem. Inf. Model*., **2019**, *59*(3), 1205−1214.

[19]. Kumar, A.; and Zhang, K. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front. Chem.*, **2018**, *6*, 315.

[20]. Kaserer, T.; Beck, K.R.; Akram, M.; Odermatt, A.; Schuster, D. Pharmacophore Models, and Pharmacophore-Based Virtual Screening: Concepts and Applications Exemplified on Hydroxysteroid Dehydrogenases. *Molecules*., **2015**, *20*(12), 22799–22832.

[21]. Naderi, M.; Alvin, C.; Ding, Y.; Mukhopadhyay, S.; Brylinski, M.A graph-based approach to construct target-focused libraries for virtual screening. *J. Cheminform.*, **2016**, *8*, 14.

[22]. Kidd, S. L.; Osberger, T. J.; Mateu, N.; Sore, H.F.; Spring, D.R. Recent Applications of Diversity-Oriented Synthesis Toward Novel, 3-Dimensional Fragment Collections. *Front. Chem.*, **2018**, *6*, 460.

[23]. Holth, T.; Walters, M.A.; Hutt, O.E.; Georg, G.I. Diversity-Oriented Library Synthesis from Steviol Isosteviol-Derived Scaffolds. *ACS Comb. Sci.*, **2020**, *22*(3), 150–155.

[24]. Arya, P.; Quevillon, S.; Joseph, R.; Wei, C.; Gan, Z.; Parisien, M.; Sesmilo, E.; Reddy, P.; Chen, Z.; Durieux, P.; Laforce, D.; Campeau, L.; Khadem, S.; Couve-Bonnaire, S.; Kumar, R.; Sharma, U.; Leek, D.; Daroszewska, M.; Barnes, M. Toward the library generation of natural product-like polycyclic derivatives by stereocontrolled diversity-oriented synthesis. *Pure Appl. Chem.*, **2005**, *77*(1), 163-178.

[25]. Bosc, N.; Muller, C.; Hoffer, L.; Lagorce, D.; Bourg, S.; Derviaux, C.; Gourdel, M.E.; Rain, J.C.; Miller, T.W.; Villoutreix, B.O.; Miteva, M.A.; Bonnet, P.; Morelli, X.; Sperandio, O.; Roche, P. Fr-PPIChem: An Academic Compound Library Dedicated to Protein-Protein Interactions. *ACS Chem. Biol.*, **2020**, *15*(6), 1566–1574.

[26]. Zhang, X.; Betzi, S.; Morelli, X.; Roche, P. Focused chemical libraries--design and enrichment: an example of protein-protein interaction chemical space. *Future Med. Chem.*, **2014**, *6*(11), 1291–1307.

[27]. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform*., **2017**, *9*(1), 48.

[28]. Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.*, **2018**, *58*(6), 1194–1204.

[29]. Gómez-Bombarelli, R., Wei, J.N.; Duvenaud, D.; Hernández-Lobato, J.M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T.D.; Adams, R.P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.*, **2018**, *4(*2), 268–276.

[30]. Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-driven medicinal chemistry in the era of big data. *Drug Discov. Today*., **2014**, *19*(7), 859–868.

[31]. Paricharak, S.; Méndez-Lucio, O.; Chavan Ravindranath, A.; Bender, A.; IJzerman, A.P.; and van Westen, G. Data-driven approaches used for compound library design, hit triage, and bioactivity modeling in high-throughput screening. *Brief. Bioinformatics*., **2018**, *19*(2), 277–285.

[32]. Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci*., **2018**, *24*, 4(1),120–131.

[33]. Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inform*., **2018**, *37*(1-2), 1700153.

[34]. Liu, J.; Wang, S.; Balius, T.E.; Singh, I.; Levit, A.; Moroz, Y.S.; O'Meara, M.J.; Che, T.; Alga, E.; Tolmachova, K.; Tolmachev, A.A.; Shoichet, B.K.; Roth, B.L.; Irwin, J.J. Ultra-large library docking for discovering new chemotypes. *Nature*., **2019**, *566*(7743), 224–229.

[35]. NMRShiftDB2. NMR database. Web database. (https://nmrshiftdb.nmr.uni-koeln.de/html) (Accessed Feb 09, **2022**)

[36]. The Spectral Database for Organic Compounds (SDBS). Interactive databases. (https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi.html) (Accessed Feb 09, **2022**)

[37]. The NP-MRD (Natural Products Magnetic Resonance Database) is a freely available cloud-based, FAIR electronic database. (https://np-mrd.org/html) (Accessed Feb 09, **2022**)

[38]. Natural Products 13C NMR Database. Interactive Analysis Tool. (https://c13.materia-medica.net/html) (Accessed Feb 09, **2022**)

[39]. Perez-Riverol, Y.; Wang, R.; Hermjakob, H.; Müller, M.; Vesada, V.; Vizcaíno, J. A. Open-source libraries and frameworks for mass spectrometry-based proteomics: a developer's perspective. *Biochim. Biophys. Acta*., **2014**, *1844*(1 Pt A), 63–76.

[40]. Gabriel, J.; Höfner, G.; Wanner K.T. A Library Screening Strategy Combining the Concepts of MS Binding Assays and Affinity Selection Mass Spectrometry. *Front. Chem.,* **2019**, DOI: 10.3389/fchem.2019.00665

[41]. McLaren, D.G.; Shah, V.; Wisniewski, T.; Ghislain, L.; Liu, C.; Zhang, H.; Saldanha, S.A. High-Throughput Mass Spectrometry for Hit Identification: Current Landscape and Future Perspectives. *SLAS Discovery: Advancing life Sciences R & D*, **2021**, *26*(2), 168–191.

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

[43]. Weininger, D.; Weininger, A.; Weininger J.L. SMILES 2 Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, **1989**, *29*(2), 97–101.

[44]. Weininger, D. SMILES a language for molecules and reactions. In: *Handbook of chemoinformatics*. Gasteiger, J. ed.; Wiley-VCH Verlag GmbH & Co. KgaA, Germany, **2003**, 80–102.

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

[46]. Hanson, R.M. Jmol SMILES and Jmol SMARTS: Specifications and applications. *J. Cheminform*., **2016**, *8*:50

[47]. Winter, R.; Montanari, F.; Noé, F.; Clevert, D.A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.,* 2018, *10*(6), 1692–1701.

[48]. Arús-Pous, J.; Patronov, A.; Bjerrum, E.J.; Tyrchan, C.; Reymond, J. L.; Chen, H.; Engkvist, O. SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminform.,* **2020**, *12*(1), 38.

[49]. Shampa Raghunathan, S.; and Priyakumar, U.D. Molecular representations for machine learning applications in chemistry. *Int. J. Quantum Chem*., ***2021***, e26870. DOI: /10.1002/qua.26870

[50]. Meyers, J.; Fabian, B.; Brown, N. De novo molecular design and generative models. *Drug Discov. Today*. **2021**, *26*(11), 2707–2715.

[51]. Weininger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *J. Chem. Inf. Model*., **1990**, *30*(3), 237–243.

[52]. Schmidt, R.; Ehmki, E.; Ohm, F.; Ehrlich, H.C.; Mashychev, A.; Rarey, M. Comparing Molecular Patterns Using the Example of SMARTS: Theory and Algorithms. *J. Chem. Inf. Model.*, **2019**, *59*(6), 2560–2571.

[53]. McNaught, A. The IUPAC International Chemical Identifier. *Chem. Int.*, **2006**, *28* (*6*), 12−15.

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

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

[56]. InChIKey. (https://www.inchi-trust.org/html) (Accessed Feb 15, **2022).**

[57]. Ullmann, J.R. An Algorithm for Subgraph Isomorphism. *Journal of the ACM (JACM)* **1976**, *23*, 31−42.

[58]. Mahmood, O.; Mansimov, E.; Bonneau, R.; Cho, K. Masked graph modeling for molecule generation. *Nat. Commun.,* **2021**, *12*(1), 3156.

[59]. Yirik M.A.; Steinbeck.; C. Chemical graph generators. *PLoS Comput. Biol.*, **2021**, *17*(1): e1008504.

[60]. Butina, D. Unsupervised DataBase Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *Chem. Inf. Comput. Sci*., **1999**, *39* (4), 747–750.

[61]. Lim, J.; Hwang, S.; Kim, S.; Moon, S.; Kim, W.Y. Scaffold-based molecular design with a graph generative model. *Chem. Sci.,* **2019**, *11*, 1153 - 1164.

[62]. Shampa Raghunathan, S.; Priyakumar, U.D. Molecular representations for machine learning applications in chemistry. *Int. J. Quantum Chem.*, **2021**, e26870. DOI: /10.1002/qua.26870.

[63]. Mercado, R.; Rastemo, T.; Lindelöf, E.; Klambauer, G.; Engkvist, O.; Chen, H.; Bjerrum, E.J. Graph Networks for Molecular Design. *Mach. Learn.: Sci. Technol.*, 2, **2020**, 025023. ChemRxiv, Cambridge Open Engage

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

[65]. Kawabata T. Build-up algorithm for atomic correspondence between chemical structures. *J. Chem. Info. Model*., **2011**,*5*1, 1775-1787.

[66]. Schneider, P.; Schneider, G. De Novo Design at the Edge of Chaos. *J. Med. Chem.*, **2016**, *59*(9), 4077–4086.

[67]. Schneider, G.; Clark, D.E. Automated De Novo Drug Design: Are We Nearly There Yet? *Angewandte International edition Chemie* **2019**, *58*(32), 10792-10803.

[68]. Mouchlis, V.D.; Afantitis, A.; Serra, A.; Fratello, M.; Papadiamantis, A. G.; Aidinis, V.; Lynch, Greco, D.; Melagraki, G. Advances in de Novo Drug Design: From Conventional to Machine Learning Methods. *Int. J. Mol. Sci*., **2021**, *22*(4), 1676.

[69]. Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.*, **2018**, *4(*7), eaap7885. https://doi.org/10.1126/sciadv.aap7885.

[70]. Li, Y.; Pei, J.; Lai L. Structure-based de novo drug design using 3D deep generative models. *Chem. Sci*., **2021**, *12*, 13664–13675.

[71]. Spiegel, J.O.; Durrant, J.D. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *J. Cheminform*., **2020**, *12*, 25.

[72]. Domenico A.; Gambacorta N.; Trisciuzzi D.; Ciriaco F.; Amoroso N.; Nicolotti O. De Novo Drug Design of Targeted Chemical Libraries Based on Artificial Intelligence and Pair-Based Multiobjective Optimization. *J. Chem. Inf. Model*., **2020**, *60*(10), 4582–4593.

[73.]. Ghiandoni, G.M.; Bodkin, M.J.; Chen, B.; Hristozov, D.; Wallace, J. E.A.; Webster, J.; Gillet, V. J. Enhancing reaction-based de novo design using a multi-label reaction class recommender. *J. Comput. Aided Mol. Des.*, **2020**, *34*, 783–803.

[74]. Kidd, S.L.; Osberger, T.J.; Mateu, N.; Sore, H.F.; Spring, D.R. Recent Applications of Diversity-Oriented Synthesis Toward Novel, 3-Dimensional Fragment Collections. *Front. Chem.,* **2018**, *6*, 460.

[75]. Janes, J.; Young, M.E.; Chen, E.; Rogers, N.H.; Burgstaller-Muehlbacher, S.; Hughes, L.D.; Love, M.S.; Hull, M.V.; Kuhen, K.L.; Woods, A. K.; Joseph, S. B.; Petrassi, H. M.; McNamara, C. W.; Tremblay, M.S.; Su, A.I.; Schultz, P.G.; Chatterjee, A.K. The ReFRAME library is a comprehensive drug repurposing library and its application to the treatment of cryptosporidiosis. *Proceedings of the National Academy of Sciences of the United States of America*, **2018**, *115*(42), 10750–10755.

[76]. Meyers J.; Fabian B.; Brown N. De novo molecular design and generative models. *Drug Discov. Today.*, **2021**, *26*(11), 2707-2715.

[77]. Dunn, T.B.; Seabra, G. S.; Kim, T.K.; Juárez-Mercado, K.E.; Li, C.; Medina-Franco, J.L.; Miranda-Quintana, R.A. Diversity and Chemical Library Networks of Large Data Sets. *J. Che. Inf. Model.*, **2022,** 62, 2186-2201.

[78]. Medina-Franco, J.L.; Sánchez-Cruz, N.; López-López, E.; Díaz-Eufracio, B.I. Progress on open chemoinformatic tools for expanding and exploring the chemical space. *J. Comput-Aided Drug Design.*, **2022**, in press. DOI: 10.1007/s10822-021-00399-1

[79]. Quartararo, A.J.; Gates, Z.P.; Somsen, B.A.; Hartrampf, N.; Ye, X.; Arisa Shimada, A.; Kajihara, Y.; Ottmann, C.; Pentelute, B.L. Ultra-large chemical libraries for the discovery of high-affinity peptide binders. *Nat. Commun.*, **2020**, *11*, 3183.

[80]. Grygorenko, O.O.; Radchenko, D.S.; Dziuba, I.; Chuprina, A.; Gubina, K.E.; Moroz, Y.S. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23*(11), 101681.

[81]. Varnek, A.; and Baskin, I. I. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol. Inf.,* **2011**, *30*(*1*), 20–32. DOI: 10.1002/minf.201000100

[82]. Meggers E. Exploring biologically relevant chemical space with metal complexes. *Cur. Opin. Chem. Biol.,* **2007,** *11*(*3*), 287–292.

[83]. Saldívar-González, F.I.; Medina-Franco, J.L. Chemoinformatics approaches to assess chemical diversity and complexity of small molecules. In: *Small molecule drug discovery. Methods, molecules, and applications*. (eds) Trabocchi, A.; Lenci, E., Ed.; **2020**, 83-102.

[84]. Medina-Franco J.L. Grand Challenges of Computer-Aided Drug Design: The Road Ahead. *Front. Drug Discov.;* **2021**, *1*, 728551.

[85]. Núñez, M.J.; Díaz-Eufracio, B.I.; Medina-Franco, J.L.; Olmedo, D.A. Latin American databases of natural products: Biodiversity and drug discovery against SARS-CoV-2. *RSC Adv.*, **2021**, *11*(26), 16051–16064.

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

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

[88]. Olmedo, D.A.; González-Medina, M.; Gupta, M.P.; Medina-Franco, J.L. Cheminformatic characterization of natural products from Panama. *Mol. Divers.*, **2017**, *21*(4), 779–789.

[89]. Saldívar-González, F.I.; Pilón-Jiménez, B.A.; Medina-Franco, J.L. Chemical space of naturally occurring compounds. *Phys. Sci. Rev.* **2018**, *4*.

[90]. Rodrígues T. Harnessing the potential of natural products in drug discovery from a cheminformatics vantage point. *Org. Biomol. Chem.,* **2017**, *15*(44), 9275–9282.

[91]. Osolodkin, D.I.; Radchenko, E.V.; Orlov, A.A.; Voronkov, A.E.; Palyulin, V.A.; and Zefirov, N.S. Progress in visual representations of chemical space. *Expert Opin Drug Discov*., **2015**, *10*(9), 959–973.

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

[93]. Naveja, J.J.; Rico-Hidalgo, M.P.; and Medina-Franco, J.L. Analysis of a large food chemical database: chemical space, diversity, and complexity. *F1000Research* **2018**, *7*, Chem. Inf. Sci-993.

[94]. Bayer, S.; Mayer, A.I.; Borgonovo, G.; Morini, G.; Di Pizio, A.; Bassoli, A. Chemoinformatics View on Bitter Taste Receptor Agonists in Food. *J. Agric. Food Chem.*, **2021**, *69*(46), 13916–13924.

[95]. Schuffenhauer, A.; Varin, T. Rule-Based Classification of Chemical Structures by *Scaffold. Mol. Inf.*, **2011**, *30*(8), 646–664.

[96]. Medina-Franco, J.L.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR Comb. Sci.* **2009**, *28*, 1551-1560.

[97]. Bhurta, D.; Bharate, S. B. Analyzing the scaffold diversity of cyclin-dependent kinase inhibitors and revisiting the clinical and preclinical pipeline. *Med. Res. Rev.*, **2022**, *42*, 654–709.

[98]. Maldonado, A.G.; Doucet, J.P.; Petitjean, M.; Fan, B.T. Molecular similarity and diversity in chemoinformatics: from theory to applications. *Mol. Divers.* **2006**, *10*(1), 39–79.

[99]. Rácz, A.; Bajusz, D.; Héberger, K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *J. Cheminform.*, **2018**, *10*, 48.

[100]. Yongye, A.B.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G.M.; Medina-Franco, J.L. Consensus models of activity landscapes with multiple chemicals, conformer, and property representations. *J. Chem. Inf. Model.,* **2011**, *51*(6), 1259–1270.

[101]. Medina-Franco, J.L.; Saldívar-González, F.I. Cheminformatics to Characterize Pharmacologically Active Natural Products. *Biomolecules*., **2020**, *10*(11), 1566.

[102]. Zagidullin, B.; Wang, Z.; Guan, Y.; Pitkänen, E.; Tang, T. Comparative analysis of molecular fingerprints in the prediction of drug combination effects, *Briefings in Bioinformatics*., **2021**, *22*(*6*), 1–15.

[103]. Tran, T.D.; Ogbourne, S.M.; Brooks, P.R.; Sánchez-Cruz, N.; Medina-Franco, J.L.; Quinn, R.J. Lessons from Exploring Chemical Space and Chemical Diversity of Propolis Components. *Int. J. Mol. Sci.*, **2020**, *21*(14), 4988.

[104]. Olmedo, D. A.; and Medina-Franco, J. L. Chemoinformatic Approach: The Case of Natural Products of Panama. In: *Cheminformatics and its Applications.* Stefaniu, A.; Rasul, A.; Hussain, G. Eds.; Intechopen Ltd. The United Kingdom. **2020**, 83-106.

[105]. Amberg, A.; Riefke, B.; Schlotterbeck, G.; Ross, A.; Senn, H.; Dieterle, F.; Keck, M. NMR and MS Methods for Metabolomics. In: *Drug Safety Evaluation. Methods in Molecular Biology*, (eds) Gautier J. C.; Humana Press, New.Jersey, **2017**, vol.1641, 229–258.

[106]. Schlotterbeck, G.; Ross, A.; Dieterle, F.; Senn, H. Metabolic profiling technologies for biomarker discovery in biomedicine and drug development. *Pharmacogenomics*., **2006**, *7*(7), 1055–1075.

[107]. Yang, X.; Parker, D.; Whitehead, L.; Ryder, N. S.; Weidmann, B.; Stabile-Harris, M.; Kizer, D.; McKinnon, M.; Smellie, A.; Powers, D. A collaborative hit-to-lead investigation leveraging medicinal chemistry expertise with high throughput library design, synthesis, and purification capabilities. *Comb. Chem. High Throughput Screen.*, **2006**, *9*(2), 123–130

[108]. Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res*., ***2006***, *34*(Database issue), D668–D672.

[109]. Xia X. Bioinformatics and Drug Discovery. *Curr. Top. Med. Chem*., **2017**, *17*(15), 1709–1726.

[110]. Romano, J.D.; Tatonetti, N.P. Informatics and Computational Methods in Natural Product Drug Discovery: A Review and Perspectives. *Front. Genet.*, **2019**, *10*.

[111]. Behl, T.; Kaur, I.; Sehgal, A.; Singh, S.; Bhatia, S.; Al-Harrasi, A.; Zengin, G.; Babes, E.E.; Brisc, C.; Stoicescu, M.; Toma, M.M.; Sava, C.; Bungau, S.G. Bioinformatics Accelerates the Major Tetrad: A Real Boost for the Pharmaceutical Industry. *Int. J. Mol. Sci.*, **2021**, *22*, 6184.

[112]. Wooller, S.K.; Benstead-Hume, G.; Chen, X.; Ali, Y.; Pearl, F. (2017). Bioinformatics in translational drug discovery. *Biosci. Rep*., **2017**, *37*(4), BSR20160180.

[113]. Yan, Q. *Translational Bioinformatics and Systems Biology Methods for Personalized Medicine*; 1st edition ed.; Academic Press Elsevier: Massachusetts, **2017.**

[114]. Wu, F.; Zhou, Y.; Li, L.; Shen, X.; Chen G.; Wang, X.; Liang, X.; Tan, M.; Huang, Z. Computational Approaches in Preclinical Studies on Drug Discovery and Development. *Front. Chem.*, **2020**, 8, 726.

[115]. Pérez-Santín, E.; Rodríguez-Solana, R.; González-García, M.; García-Suárez, M.; Blanco-Díaz, G.D.; Cima-Cabal1, M.D.; Moreno-Rojas, J.M.; López Sánchez, J.I. Toxicity prediction based on artificial intelligence: A multidisciplinary overview. *WIREs Comput Mol.Sci*., **2021**, e1516. DOI: 10.1002/wcms.1516

[116]. Jia, L.; Gao, H. Machine Learning for *In silico* ADMET Prediction. *Methods Mol. Biol*., **2022**; *2390*, 447-460.

[117]. de Souza Neto, L.R.; Moreira-Filho, J.T.; Neves, B.Jr.; Maidana-Riveros, R.L.B.; Guimarães-Ramos, A.C.; Furnham, N.; Andrade, C.H.; and Silva, F.P. *In silico* Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Front. Chem*., 2020, *8*.

[118]. Durán-Iturbide, N.A.; Díaz-Eufracio, B.I.; Medina-Franco, J.L. *In silico* ADME/Tox. Profiling of Natural Products: A Focus on BIOFACQUIM. *ACS Omega*., **2020**, *5*(26), 16076–16084.

[119]. Lagorce, D.; Bouslama, L.; Becot, J.; Miteva, M.A.; Villoutreix, B.O. FAF-Drugs4: free ADME-tox filtering computations for chemical biology and early stages drug discovery, *Bioinformatics*., **2017**, *33*(22), 3658–3660.

[120]. Pires D.E.V, Blundell T.L, Ascher D.B. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *J. Med. Chem., 2015*, *58*, 4066–4072.

[121]. Kar, S.; Leszczynski, J. Open access *in silico* tools to predict the ADMET profiling of drug candidates. *Expert Opin. Drug Discov*., **2020**, *15*(12), 1473-1487.

[122]. Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res.*, **2021**, *49*(W1), W5-W14.

[123]. Venkatraman, V. FP-ADMET: a compendium of fingerprint-based ADMET prediction models. *J. Cheminform., 2021*, *13*(1):75.

[124]. Banerjee, P.; Eckert, A. O.; Schrey, A. K.; Preissner, R. ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res.*, **2018**, *46*(W1), W257–W263.

[125]. Baell, J.B.; Holloway, G.A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and their exclusion in bioassays. *J. Med. Chem.*, **2010**, *53*(7), 2719–2740.

[126]. Baell, J.; Walters M.A. Chemistry: Chemical con artists foil drug discovery. *Nature*., **2014**, *513,* 481–483.

[127]. Gilberg, E.; Jasial, S.; Stumpfe, D.; Dimova, D.; Bajorath, J. Highly Promiscuous Small Molecules from Biological Screening Assays Include Many Pan-Assay Interference Compounds but Also Candidates for Polypharmacology. *J. Med. Chem.*, **2016**, *59*(22), 10285–10290.

[128]. Jasial, S.; Hu, Y.; Bajorath, J. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PloS one*., **2016**, *11*(4), e0153873.

[129]. Sun, J.; Zhong, H.; Wang, K.; Li, N.; Chen, L. Gains from no real PAINS: Where 'Fair Trial Strategy' stands in the development of multi-target ligands. *Acta Pharm. Sin. B.*, **2021**, *11*(11), 3417–3432.

[130]. Baell, J.B.; Ferrins, L.; Falk, H.; Nikolakopoulos, G. PAINS: Relevance to tool compound discovery and fragment-based screening. Australian *J. Chem.*, **2014**, *66*(12), 1483–1494.

[131]. dos Santos, J. L.; and Chin, C. M. Pan-Assay Interference Compounds (PA1INS): Warning Signs in Biochemical-Pharmacological Evaluations. *Biochem. Pharmacol*., (Los Angel) *2015*, *4*, e173.

[132]. Baell, J.B. Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *J. Nat. Prod*., **2016**, *79*(3), 616–628.

[133]. Lagorce, D.; Oliveira, N.; Miteva, M. A.; Villoutreix, B. O. Pan-assay interference compounds (PAINS) that may not be too painful for chemical biology projects. *Drug Discov. Today*., **2017**, *22*(8), 1131–1133.

[134]. Baell, J.B.; Nissink, J. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017-Utility and Limitations. *ACS Chem. Biol*., **2018**, *13*(1), 36–44.

[135]. Gilberg, E.; Stumpfe, D.; Bajorath, J. Towards a systematic assessment of assay interference: Identification of extensively tested compounds with high assay promiscuity. *F1000Research* **2017**, *6*, Chem. Inf. Sci-1505.

[136]. Vidler, L.R.; Watson, I.A.; Margolis, B.J.; Cummins, D.J.; Brunavs, M. Investigating the Behavior of Published PAINS Alerts Using a Pharmaceutical Company Data Set. *ACS Med. Chem. Lett*., **2018**, *9*(8), 792–796.

[137]. Gilberg, E.; Stumpfe, D.; Bajorath, J. Activity profiles of analog series containing pan assay interference compounds. *RSC Adv*., **2017**, *7*, 35638–35647.

[138]. Swarit, J.; Ye, H.; Bajorath, J. How Frequently Are Pan-Assay Interference Compounds Active? Large-Scale Analysis of Screening Data Reveals Diverse Activity Profiles, Low Global Hit Frequency, and Many Consistently Inactive Compounds. *J. Med. Chem.*, **2017**, *60* (9), 3879–3886.

[139]. Capuzzi, S.J.; Muratov, E.N.; and Tropsha, A. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay INterference CompoundS. *J. Chemi. Inf. Model*., **2017**, *57*(3), 417–427.

[140]. Chakravorty, S.J.; Chan, J.; Greenwood, M.N.; Popa-Burke, I.; Remlinger, K.S.; Pickett, S.D.; Green, D.; Fillmore, M.C.; Dean, T.W.; Luengo, J.I.; Macarrón, R. Nuisance Compounds, PAINS Filters, and Dark Chemical Matter in the GSK HTS Collection. *SLAS Discovery: Advancing life Sciences R & D*., **2018**, *23*(6), 532–545.

[141]. Koptelov, M.; Zimmermann, A.; Bonnet. P.; Bureau, R.; Crémilleux, B. *PrePeP: A Tool for the Identification and Characterization of Pan Assay Interference Compounds*. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Aug **2018**, Londres, United Kingdom. pp.462-471.

[142]. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In: *Data Analysis, Machine Learning and Applications*: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e. V., Albert-Ludwigs-Universität Freiburg, March 7 9, 2007. Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R. Eds.; Springer, Berlin, Germany, 2008, pp. 319–326.

[143]. OEChem TK, OpenEye Scientific Software, Inc.; Santa Fe, NM, U.S.; 2012

[144]. RStudio: Integrated Development Environment for R, RStudio, Inc.; Boston, MA, 2016.

[145]. Bajorath J. Evolution of assay interference concepts in drug discovery. *Expert Opin. Drug Discov.*, **2021**, *16*(7), 719–721.

[146]. Magalhães, P.R.; Reis, P.; Vila-Viçosa, D.; Machuqueiro, M.; Victor, B.L. Identification of Pan-Assay INterference compoundS (PAINS) Using an MD-Based Protocol. *Methods Mol. Biol.*, **2021**, *2315*, 263–271.

[147]. Matlock, M.K.; Hughes, T.B.; Dahlin, J.L.; and Swamidass, S.J. Modeling Small-Molecule Reactivity Identifies Promiscuous Bioactive Compounds. *J. Chemi. Inf. Model.*, **2018**, *58*(8), 1483–1500.