Review

Exploring combinations in chemoinformatics: Toward a multidisciplinary view

José L. Medina-Franco,[1,*] Edgar López-López,[1,2] Johny R. Rodríguez-Pérez,[3,4] Héctor F. Cortés-Hernández,[3] Samuel Homberg,[5]

[1] DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, México City 04510, Mexico.

[2] Department of Chemistry and Graduate Program in Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute, Section 14-740, Mexico City 07000, Mexico.

[3] GIFAMol Research Group, School of Chemistry Technology, Universidad Tecnológica de Pereira, Pereira 660003, Colombia.

[4] GIEPRONAL Research Group, School of Basic Sciences, Technology and Engineering, Universidad Nacional Abierta y a Distancia, Dosquebradas 661001, Colombia.

[5] Institute of Pharmaceutical and Medicinal Chemistry, Westfälische Wilhelms-Universität Münster, 48149 Münster, Germany.

Abstract: In Chemoinformatics, as in many other computational-related disciplines, it is a common practice to identify the "single best" approach or methodology, for instance, identify the best fingerprint representation, the best single virtual screening approach or protocol, the optimal representation of the chemical space, the best predictive model, to name a few. In molecular modeling, a typical example is finding the best docking program. However, it is also known that each approach has its advantages and limitations. There are examples of benchmark studies comparing different approaches to find the most appropriate solution, and it is common to find that there are no single best programs in such studies. Yet, searching for the "best" methods is still common. The main goal of this work is to survey hybrid methodologies typically used in Chemoinformatics. The list of approaches is not exhaustive, but it aims to cover several representative applications. One of the major outcomes of the survey is that, for various purposes, individual methods do not perform as well as the combination of approaches because single methods have inherent limitations with advantages and disadvantages.

Keywords: art; consensus; chemical space; data fusion; education; ensemble; hybrid method; machine learning; multidisciplinary; open science.

Abbreviations: AI, artificial intelligence; ADMET, administration, distribution, metabolism, excretion, and toxicity; CADD, computer-aided drug design; ECFP, extended connectivity fingerprint; ML, machine learning; QSAR, quantitative structure-activity relationships; SAR, structure-activity relationships; SELFIES, SELF-referencIng Embedded Strings; SMARts, structure-multiple activity relationships; SP(A)R, structure-property (activity) relationships; SIR, structure-inactivity relationships; SPR, structure-property relationships.

1. Introduction

In drug discovery and many other complex endeavors, multidisciplinary approaches are essential. In science, it is fairly common to come across a combination of concepts, methodologies, and viewpoints that generate and develop novel research areas and pose novel ideas. Indeed, multidisciplinary research teams have been recognized as a key element to address health care problems.[1] Chemoinformatics[2] itself (also named in the literature cheminformatics, chemical informatics, etc.),[3] is a good example of such a merge of "traditional" disciplines.[4] Another good example is Bioinformatics.[5]

In drug discovery, abundant examples of the synergistic combination of compounds are well known to exhibit a better performance than the individual isolated compounds. The combinations can be very complex giving rise to study areas by themselves such as polypharmacy,[6] traditional medicine,[7] botanicals,[8] nutraceuticals, screening, and deconvolutions of mixture combinatorial libraries.[9] For example, polypharmacy refers to using multiple medications to treat one disease or condition, while traditional medicine often employs combinations of herbs or natural ingredients to treat various ailments. Botanicals can also contain a variety of compounds that act synergistically. Screening and deconvolution of combinatorial mixture libraries involves testing large numbers of compound combinations to identify those that show desired therapeutic activity. These are a few examples of research areas founded on the idea that individual and "best" single compounds, medicines, chemical libraries, methods, etc., are outperformed by their combination which, by itself, can be quite challenging.

Over the years, combinations of methodologies and approaches have been emerging and evolving in chemoinformatics for various practical applications including, but not limited to, molecular representation, chemical space analysis, similarity searching, property prediction, and structure-activity relationships (SAR), to name a few examples. Such combinations can be influenced by the numerous attempts to

identify the best single approach through benchmark or comparative studies. In many instances, the outcome is that the most appropriate approach depends on the study case or research system. This is frequent in molecular docking, one of the most widely used methods in computer-aided drug design (CADD). It serves as a fundamental technique for predicting the binding mode of bioactive compounds and conducting virtual screening.[10] Due to the requirement of enhancing its reliability in pose prediction and performance in virtual screening, new docking algorithms and scoring functions have been developed and optimized. However, it is unlikely to identify a single procedure that overcomes others in terms of reliability and precision or proves to be suitable for all types of molecular targets.[11]

The goal of this manuscript is to survey various types of combinations that are commonly done in Chemoinformatics. In light of the current rise of machine learning (ML), we also comment on emerging combinations that are being developed, paving the way for original and improved research areas. When the information, we provide the reference and or link to code, in particular when the tools are freely available. As discussed, the combination of research methods can be particularly interesting, as an alternative to single conventional strategies in which the research objectives are seen from a unique perspective. The manuscript is organized into four main sections. After this Introduction, section two discusses sub-disciplines that have emerged or are evolving as the combination of more traditional or long-established disciplines. Section three presents exemplary types of combinations commonly done involving chemoinformatics with different applications in molecular representation, property prediction, structure-property (activity) relationships -SP(A)R-, virtual screening, chemical space, ML, and other applications such as chemistry and art. This section does not include all hybrid methodologies but most common ones. Section four presents summary conclusions.

2. Combination of knowledge: Creating new disciplines

Science has evolved towards a more holistic and multi- and transdisciplinary perspective. Now, various disciplines are emerging, being the product of the combination of different perspectives that have emerged from multidisciplinary research groups. Thus, chemoinformatics, being a relatively young discipline, has given rise to the creation of new disciplines and subdisciplines that combine chemical, biological, and biomedical science data. **Figure 1** illustrates examples of data used in chemistry, biology, and biomedical

sciences, which lead to related disciplines. Now, disciplines related to the study of materials, polymers, and food chemicals have emerged, and other disciplines most related to biology and biomedical concepts have also benefited from methodologies, concepts, and protocols originally used in chemoinformatics focused on drug discovery. For example, molecular modeling, drug design, and toxicology-related informatics disciplines.

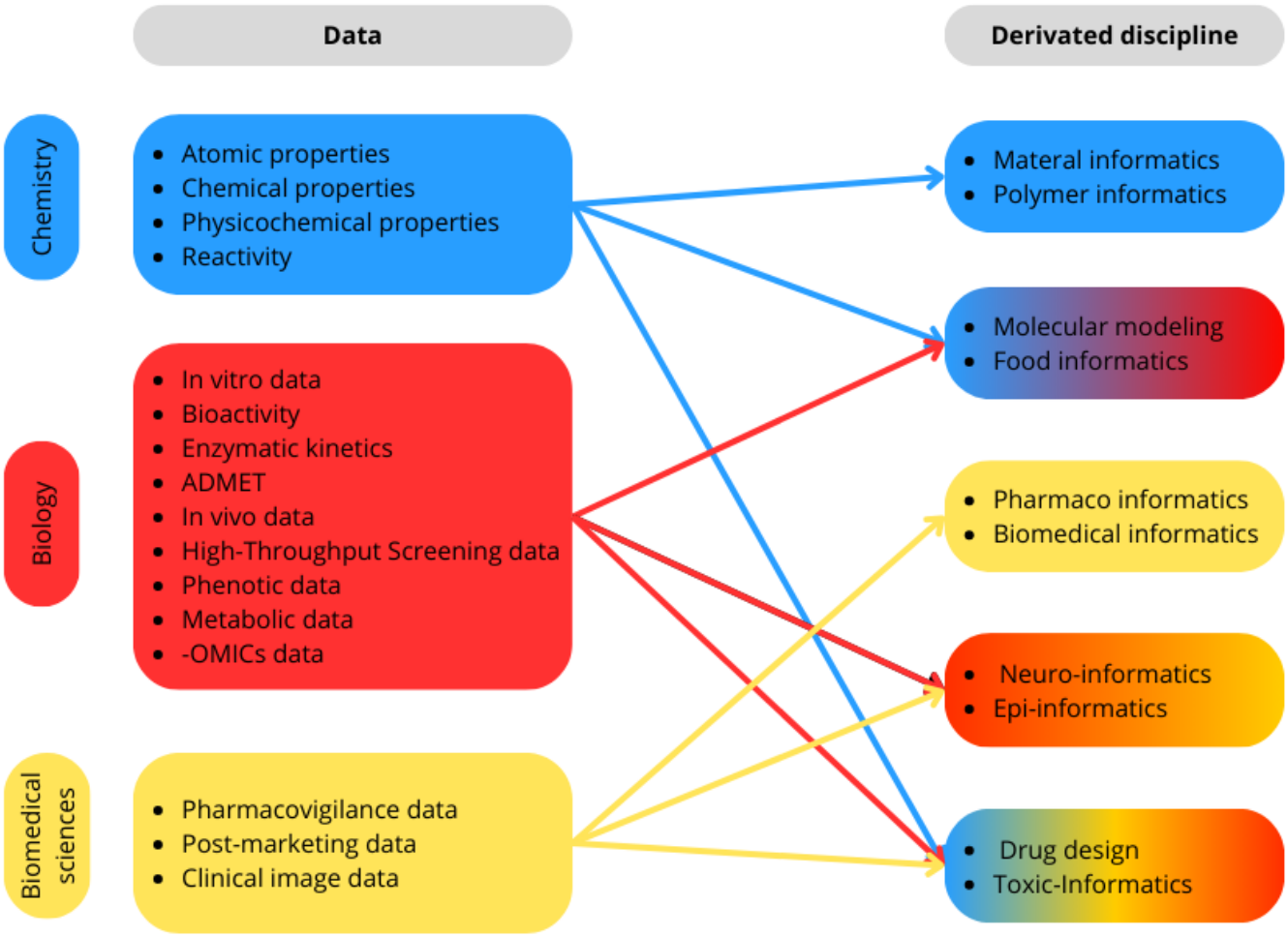

Figure 1. Exemplary data types used in chemistry, biology, and biomedical sciences and their combinations to generate new sub-disciplines. The list is not exhaustive but exemplary sub-disciplines are listed.

3. Combination of methods: the power and potential of consensus

There are numerous examples in which the integration of research strategies offers possibilities for evaluating situations beyond what a single technique or methodology would allow, for example, the combination of metabolomics (an emerging field of omics sciences that deals with quantitative and

qualitative analysis of molecules in a biological sample) and chemoinformatics, offers a powerful combination to annotate metabolites and identify biomarkers.[12]

Combined methodologies have led to the generation of new useful approaches for different kinds of disciplines, like pharmacology, food chemistry, toxicology, and molecular and material design. Table 1 presents examples of emerging combined approaches inspired by integrated methodologies and their utilities. The combinations in different areas, along with their representative references, are further discussed below and are organized by general application as outlined in Table 1.

Table 1. Emerging combined approaches inspired by methods used in chemoinformatics.

Approach	Utility	Combined methodologies
Molecular representation	Multiple applications	Molecular modeling, molecular similarity
Structure-property relationships (SPRs)	Molecular design	Molecular similarity, chemical spaces, data fusion
Structure-multiple activity relationships (SMARts)	Polypharmacology and multiparameter molecular design	Molecular similarity, chemical spaces, data fusion
ADMET prediction	Side and off-target prediction	AI methods, molecular similarity, chemical spaces,
De novo design	Exploration of chemical space, molecular design	AI methods, molecular similarity, ligand and structure-based methods, data fusion
Network-based	Biomarker prediction, SPR	-omics, molecular similarity, AI methods, chemical spaces
Virtual screening: similarly searching and consensus similarity	Molecular design, multiparameter optimization, SPR, side and off-target prediction	-omics, molecular similarity, chemical spaces, SPR, SMARts
Chemical space: Consensus and multiverse	Multiparameter molecular optimization	Data fusion, molecular similarity

Table 2 presents free tools applicable to combined approaches in various areas of application described throughout the section.

Table 2. Free tools to conduct combined methodologies involving chemoinformatics.

Application area	Tool: combined approach	Reference	Link to the free tool
Prediction of reactivity, biological activity	Molecular representations: Advantages of combining different descriptors and fingerprints.	[13]	http://zivgitlab.uni-muenster.de/ag-glorius/published-paper/evompf
Diversity analysis	Consensus Diversity Plots: combine, in a single plot, diversity measures obtained with different representations.	[14]	https://consensusdiversityplots-difacquim-unam.shinyapps.io/RscriptsCDPlots/
Chemical space visualization	Constellation Plots: combine fingerprint and substructure-representations	[15,16]	https://github.com/navejaromero/analog-series
Chemical space	Multifusion Similarity Maps: combination of data fusion metrics to represent chemical spaces.	[17]	https://forum.knime.com/t/double-looping-to-create-multi-fusion-similarity-maps/1887
Chemart	Chemical Art Gallery: digital paintings based on actual visualizations of chemical space.	[18]	https://github.com/DIFACQUIM/Art-Driven-by-Visual-Representations-of-Chemical-Space-
Medicinal chemistry	Various tools focused on scaffolds and other representations of molecular structures.	[19–24]	https://peter-ertl.com/molecular/index.html

3.1. Molecular representation

One of the most fundamental questions in chemoinformatics and computational chemistry in general is how to represent molecules. Historically, molecular representations have evolved with the problems at hand. Indeed, the evolution of molecular representation is defined by combining methods; an example of which is the extended connectivity fingerprint (ECFP), nowadays a very common representation used for ML. The ECFPs, also called Morgan fingerprints, begin with the Morgan algorithm, which was originally designed to obtain a unique identifier for molecules. However, the molecules were still represented as a collection of tabular data.[25] Combining the idea of hashed fingerprints, which were originally designed for substructure searches[26] with the Morgan algorithm led to ECFPs.[27] These fingerprints, designed for structure-activity modeling, proved to be useful representations in ML tasks.[28]

While exact 3D representations, such as simple xyz-files with atom-coordinates are interesting for theoretical chemistry, chemoinformatics deals with large and ultra-large chemical spaces, where data has to be read and processed fast. Simple strings have proven efficient in handling large datasets. String-based representations (SMILES,[27] SMARTS,[29] InChI[30]) use the atom types and mathematical graphs to encode the molecule's connections (2D structure). With the advance of ML, string-based representations with fewer possibilities of generating invalid molecules were developed (DeepSMILES,[31] SELFIES[32]). SELFIES are interesting because they solve the need for a 100% valid string representation by combining the graph-connectivity with the concept of formal grammar from theoretical computer science. This allows to derive every string into a valid molecule. Despite that, it is not clear whether SELFIES are not generally superior string representations and it is unclear whether they improve generative models.[33]

Another approach to improving molecular representations moves away from *general* representations and instead focuses on *domain-specific* representations, mostly for ML or virtual screening. Here, ML is used to combine data (of a specific domain) into novel fingerprints. In contrast to standard, general-purpose fingerprints such as Morgan fingerprints, the fingerprints contain additional domain-specific information, for example about the activity on certain targets or the reactivity with certain catalysts.

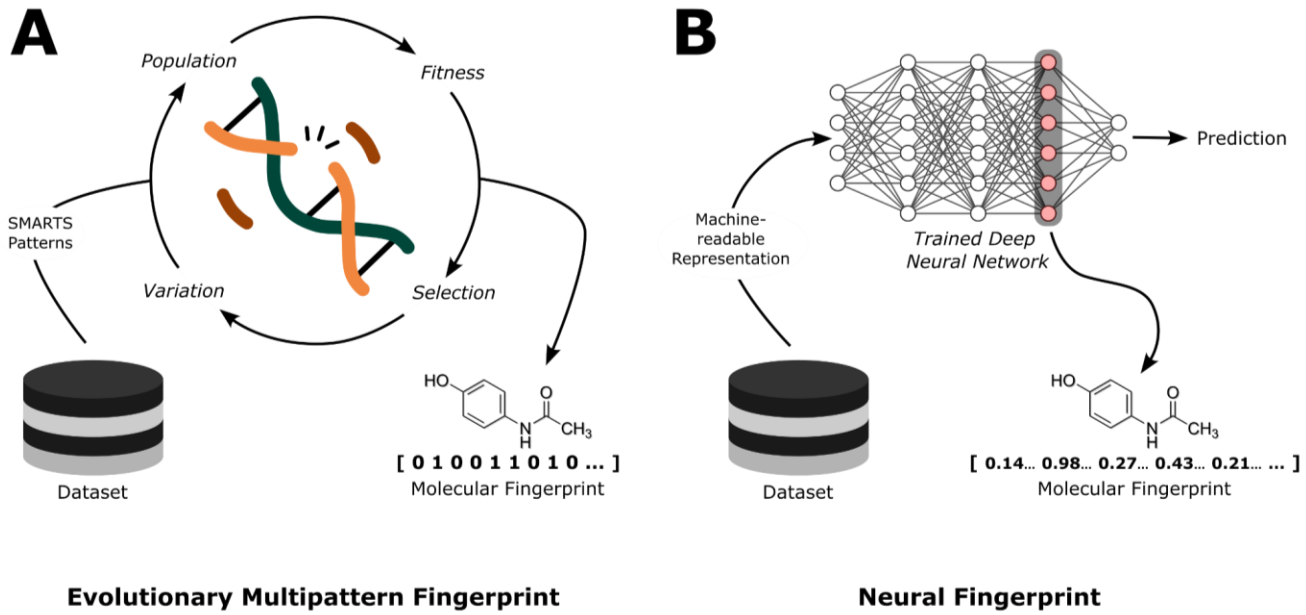

Figure 2. Domain-specific molecular fingerprints are generated by combining data using machine learning. **A** EvoMFP uses an evolutionary algorithm to combine SMARTS patterns into fingerprints for a specific dataset. **B** Neural Fingerprints generate dataset- and task-specific fingerprints by extracting the embeddings of the last layer of a neural network as a fingerprint.

The evolutionary multipattern fingerprint (EvoMFP, **Figure 2A**),[13] a recently developed molecular representation, uses combination as its core design to challenge classical fingerprints on ML tasks. It is generated with an evolutionary algorithm by combining substructures from a large set of SMARTS queries for an entire dataset. These fingerprints are easier to interpret and, while they are built specifically for each domain defined by the dataset, they can still be used for a variety of tasks, for example, prediction of reactivity and biological activity.

The neural fingerprint (**Figure 2B**)[34,35] uses a neural network to combine structural information with information from the task, e.g. activity on a target. The authors showed improvement in similarity-based virtual screening for a kinase and a natural product dataset. The fingerprint could also be used as a natural-product likeness score.

With the advent of large and ultralarge chemical libraries has become the need to develop molecular representations that encode entire chemical libraries. One of such approaches are the database fingerprint[36] and its natural extension, the statistical-based database fingerprint.[37] Both approaches code in a single dimension the most significant bit present/absent in a compound database that can be of any size. The database fingerprints can be used with a variety of fingerprints, either general representations or domain-specific representations.

3.2. Property prediction

Property prediction is a common practice in many chemistry applications. In drug discovery, typical examples are the prediction of biological activities and binding modes of molecules with molecular targets.[38] To this aim, it has been recognized that consensus predictions and ensemble models usually perform better than single predictors. As such, ensemble models have become quite common. In the realm of predicting binding poses and simulating protein-ligand interactions, consensus docking[39] has been largely admitted, in general, outperforming single docking approaches.[11] Similarly, several studies have shown that there is not a single docking protocol that is “the best” across a broad range of molecular targets. It often happens that a given docking protocol works well for a given receptor family and, again, consensus docking is a more reliable alternative to a single docking software or protocol.

In drug discovery it is desirable to predict, whenever possible, compounds that interfere with the assays but do not show actual biological activity. Anticipating such compounds before experimental screening is a challenging task. It has been recently reviewed that the best practice is using different models as opposed to using a single approach.[40]

Nowadays, property predictions are rethinking, and novel methodologies based on multi-parametric and multi-disciplinary data have emerged. For example, novel binding affinity predictors use a combination of ligand- and target-based approaches, like molecular similarity and molecular dynamics techniques, fused with NMR spectral data to predict the putative activity of compounds.[41]Another key example is the use of chemical, *in vitro*, and *in vivo* data to anticipate future side effects of lead compounds.[42] And, recent advances, like network pharmacology, have opened the possibility of decoding complex natural product mixtures to identify the compounds associated with their reported bioactivity.[43] Namely, the paradigm to predict properties based on unique kinds of data (i.e. chemical, biological, or clinical) has been broken now.

3.2.1. ADMETox properties

Pharmacokinetics and pharmacodynamics approaches are key points in modern molecular design and development, especially for small and biotech drugs applied in medical, nutritional, agricultural, and industrial areas. Properties involved with the absorption, distribution, metabolism, and toxicity (ADMETox) of drugs could determine their success in clinical trials. There are some software and servers oriented to predict ADMETox properties for small molecules and peptides, however, their dataset is normally constructed based on the direct modulation of key targets, but their prediction in more complex systems (i.e., *in vivo* context) is limited.[44] This current methodological gap points out the need to fuse different kinds of datasets and approaches to improve the capacity of future models to predict ADMETox endpoints. There are novel approaches based on consensus algorithms and data fusion techniques that have demonstrated a dramatic improvement when different kinds of *in vitro*, *in vivo*, and clinical data were used to decode complex pharmacokinetics and pharmacodynamics issues.[45,46] In other words, the ADMETox properties predictions must be addressed by a multidisciplinary group of specialists that consider chemical, biological, and clinical implications related to these.

3.3. Structure-property (activity) relationships

Exploring structure-property (activity) relationships - SP(A)R - is a common practice in chemistry where the property can be of various kinds such as reactivity, resistance (e.g., in material science), aroma (e.g., in food chemistry), toxicity (e.g., in drug discovery, agrochemistry), to name a few.

In drug discovery, biological activity vs. one or multiple endpoints is one of the primary properties to be predicted and thus, SAR analysis is a cornerstone in medicinal chemistry.[47] Moreover, predicting properties of pharmaceutical relevance such as those related to ADMETox are also crucial, as commented in section 3.2.1. Recently it has been emphasized the need to explore systematically structure-inactivity relationships - SIR - as part of the generation of predictive models.[48]

As in many areas in computational chemistry, quantitative analysis of SP(A)Rs and SIRs strongly depends on the molecular representation. Similar to other areas, modeling of SP(A)Rs under the concept of property (activity) landscape modeling[49] is improved by considering multiple structure representations, for instance, various fingerprints of different designs. This includes the reliable detection of activity cliffs.[50]

3.3.1. Polypharmacology perspective (SMARTs)

Activity prediction with more than biological endpoints simultaneously (e.g., in multi-target drug discovery and design) and performing structure-multiple activity relationships (SMARTs) is, of course, harder than predicting the activity for single endpoints e.g., single-target drug design, but it is becoming more and more necessary to develop drugs that have clinical efficacy.[51] In the last ten years, polypharmacology has demonstrated its valuable contribution to drug design and development campaigns, offering new perspectives and methodologies to exploit available biological data e.g., *in vitro*, *in vivo*, clinical, and postmarketing data.[52] The combination of multiparametric datasets has contributed to developing an algorithm capable of predicting side and off-target effects and drug-drug interactions more efficiently.[53] Additionally, novel data fusion techniques have demonstrated their utility to decode complex diseases e.g., neurological, metabolic, and cardiovascular diseases, that allowed the identification of novel druggable targets, molecular pathways, and the design of polypharmacy treatments.[54,55]

3.4. Virtual screening

Nowadays, virtual screening is used routinely to filter compound databases and select compounds for experimental testing. To this aim, several computational methods, traditionally divided into structure- and ligand-based methods,[56,57] can be employed. In the current explosion of the size of chemical libraries, ML is being actively implemented to screen large and ultra-large libraries.[58–61]

Depending on the goals of the project, the experimental information, the resources at hand, and the size of the compound libraries, virtual screening is usually done with more than one computational approach, typically starting with fast-filtering methods, followed by more refined (albeit slower) procedures. Thus, most common virtual screening protocols are good examples of the combination of approaches as opposed to using a single computational methodology to select compounds for testing.[62]

3.4.1. Similarity searching

One of the fastest approaches to filter compound libraries (in particular, large and ultra-large libraries) is similarity searching which is based on the similarity principle. With the caveat of potential activity cliffs in compound data sets[63] similarity searching has been quite effective to rapidly filter libraries and identify active compounds that are further selected by more refined methods. Similarity searching depends on a suitable molecular or structural representation that is used as a basis to quantify the compound similarity (in conjunction with a similarity coefficient).

2D fingerprints are basic and simplified molecular representations that are still in use to filter compound databases.[64] Thus, 2D-fingerprint similarity searching is a classical approach that is still in use today. However, several other types of molecular representations are also used to quantify similarity giving rise to different types of similarity searching such as pharmacophore similarity; phenotypic similarity; -omic similarity, etc. depending on the criteria or focus of the study.

Since there is not a single molecular representation that captures “all” relevant structural features for a given problem (see section 3.1.), multiple representations can be used. Furthermore, the results of the similarity searches can be combined giving rise to the so-called data fusion (and related concepts such as similarity fusion and group fusion) that, over the years, has proven to outperform single similarity searches.[65–67]

As discussed in section 3.4, similarity searching is integrated with other CADD methodologies to identify active compounds. For example, recently, consensus docking was combined with similarity searching, and *de novo* design to identify novel inhibitors of the epigenetic enzyme DNA methyltransferase 1.[68]

3.5. Chemical space

Chemical space is a cornerstone concept in chemoinformatics.[4] It has been defined as the ensemble of molecular descriptors that define the position of a given set of molecules.[69] As such, the concept of chemical space strongly depends on molecular representation and the descriptors used to define the space. For many practical applications, chemical space is frequently used in the context of visualization. Since it is common to use multiple descriptors to define chemical space, the visualization involves dimensionality reduction techniques.[70]

Identifying the "ideal" set of descriptors and chemical space representations has been a common subject of study. For instance, in an excellent study Casciuc et al. demonstrated that for virtual screening using multiple generative topographic maps, each encoding different descriptors, is better than using a unique topographic map.[71]

In an independent study, Naveja et al. showed that constellation plots are a suitable option to visualize in low-dimensions the chemical space of compound data sets through the combination of molecular fingerprints with molecular scaffolds to represent chemical compounds and analog series. The plots can be the basis to map and explore SP(A)Rs.[15] Constellation plots have been used to explore the SAR of tubulin inhibitors using cell-based inhibition data.[72]

The notion of using multiple structure representations to study the chemical space of compound data sets has been recently called "chemical multiverse" defined as "*a group of multiple chemical spaces, each one defined by a given set of descriptors e. g., a group of "descriptor universes*".[73] In this sense, a number studies that use multiple descriptors to represent the chemical space of a compound data sets can be regarded as "chemical multiverses," as reviewed elsewhere.[73] In a different approach, Chemical Library Networks have been recently proposed to combine and represent the chemical space and study the diversity of large and ultra-large chemical libraries.[74]

Figure 3 shows examples of visualizations of chemical space using various structural representations for multiple purposes. For example, Figure 3A illustrates the use of constellation plots to explore the SAR of compounds with analgesic and sedative effects from independent datasets, and Figure 3B illustrates a chemical space based on ECFP4 fingerprints. Both cases remark on their utility to fuse concepts applied in chemoinformatics (like chemical space) to explore biological/clinical data and to generate intuitive SAR representations for informaticians, chemists, pharmacologists, or clinicians.

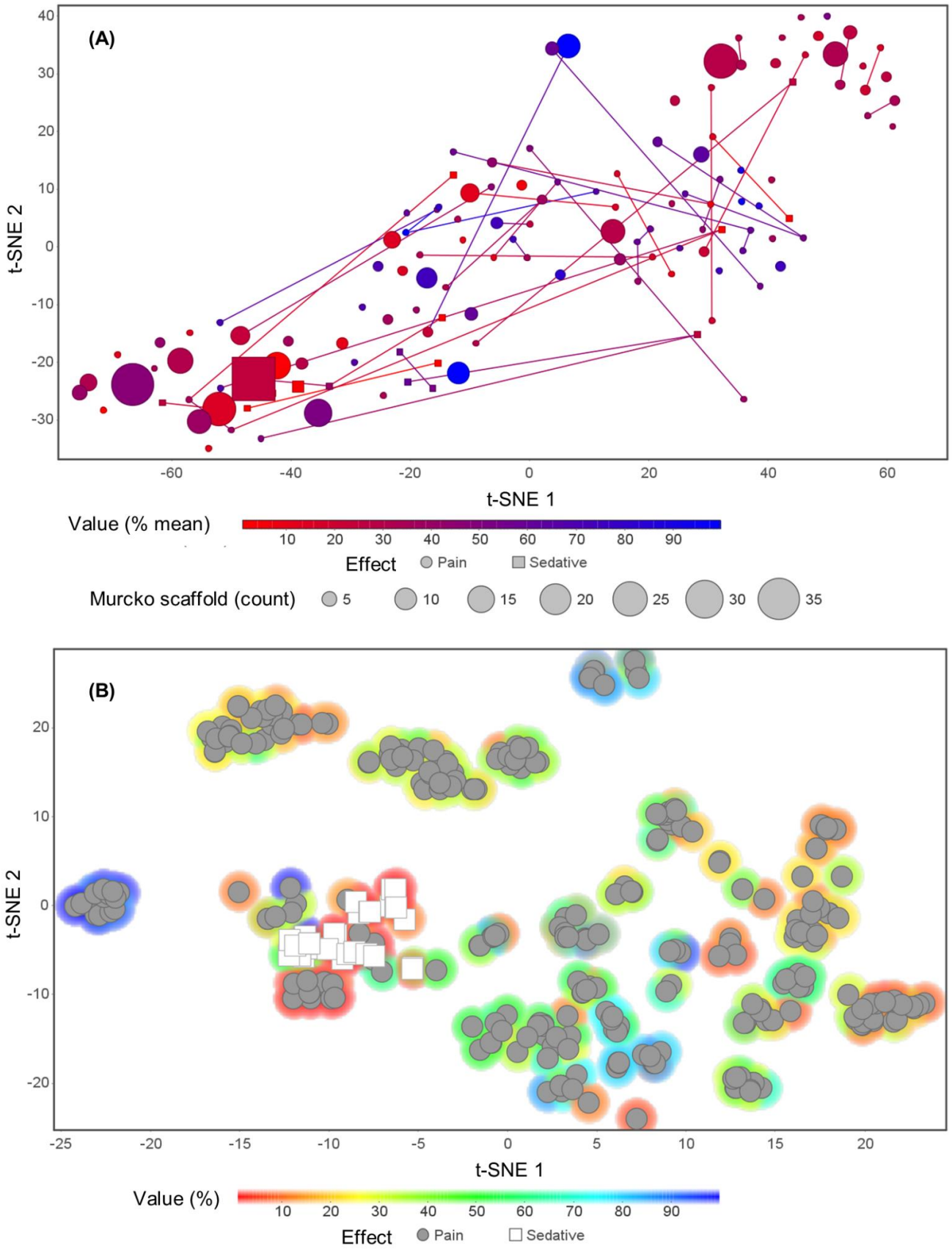

Figure 3. Multiple chemical space representations of compounds related to the treatment of pain and that generate sedative side effects. **A)** Constellation plot. The plot shows 130 data points, each one representing a scaffold. The data point's size indicates the relative number of compounds in each analog series, and the color is the average activity percentage of the compound in the series. Circle points represent scaffolds related to

anti-pain activity, and triangle points illustrate compounds related to sedative side effects. The t-SNE coordinates were constructed from druglike properties. **B)** Classical chemical space representation based on ECFP4 fingerprint. The plot shows 390 data points, each one representing a unique molecule. The data points' shape represents their associated activity, and the color around each data point represents their activity expressed in percentage; The dataset of compounds only contains activities expressed in percentage, from the 100 queries with the higher number of reported compounds. All data was collected automatically from ChEMBL v.34 database,[75] and the KNIME protocols used to construct the constellation plot are available in the supplementary material sección. The interactive visualizations were generated with DataWarrior software.[76,77]

3.6. Machine learning in chemoinformatics

Currently, in drug discovery, the combination of chemoinformatics methods and quantitative structure-activity relationship (QSAR) modeling presents itself as a highly favorable duo, allowing researchers to incorporate a new player into the equation: the use of ML techniques for predictive molecular design and analysis.[78] Chemoinformatics is understood as the interface between chemistry and informatics, where inductive learning is employed to predict chemical phenomena.[79] With increasing accessibility to chemical data, the application of ML in chemoinformatics emerges as a significant tool for exploring, analyzing, and predicting the properties and activities of molecules.[80] ML models undertake prediction tasks based on training data provided in the form of mathematical equations or numerical representations,[78] with many of these data available in chemical datasets or databases.

3.7. Other combinations: beyond chemistry

Art and science have been intimately related, for example, in the so-called "bioart"[81] defined as a "*creative practice that adapts scientific methods and draws inspiration from the philosophical, societal, and environmental implications of recombinant genetics, molecular biology, and biotechnology.*" By analogy, it can be proposed the area of "chemart." As discussed elsewhere, "chemart" can be merged or related with realistic views of scientific developments attracting the general public to science.[81] Exemplary combinations of chemoinformatics with artistic approaches such as music and painting, are commented briefly in this section.

3.7.1. Chemoinformatics and music

Another combination that seems unusual at first glance is between chemistry and music. Through sonification, non-musical information can be encoded in the high-dimensional space of music, which

humans have accessed with intuition and creativity for thousands of years. A recently developed, novel molecular representation encodes molecules as melodies.[82] The Sonic Architecture for Molecule Production and Live-Input Encoding Software, SAMPLES, combines SELFIES with chemical descriptors into music. This opens the door to applying ML techniques developed for music in the chemical sciences and might help blind chemists to interact with molecules. The authors showed that using SAMPLES, a new molecule can be generated by interpolating the melodies of two SAMPLES representations, that original molecules can be created from a musical piece played on the piano (although this still requires some knowledge in chemistry and musical theory) and perceive the use of music generation methods as. Most importantly, similar molecules encoded with SAMPLES sound similar and can be differentiated from distinct molecules, showing that their representation obeys the similar property principle, the foundation for similarity-based virtual screening. Thus, combining music and chemistry enables the creative application of previously unusable techniques from different fields of research.

3.7.2. Chemoinformatics and artistic painting

Recently the combination between visual representations of the chemical space as a means to communicate art has been proposed.[18] The underlying idea is to generate visual representations of the chemical space using well-known and standard methodologies but looking at the pictures from an artistic point of view. The visualizations of chemical spaces through artistic interpretations can serve to communicate emotions (like in art) and, at the same time, attract the general audience to chemistry in general and to chemoinformatics, in particular. Building upon this concept the first version of a *Chemical Art Gallery* has been developed and it is freely available at https://www.difacquim.com/chemical-art-gallery/. The code to generate chemart visualizations is also freely available on GitHub (see link in **Table 2**). **Figure 4** shows a further example of a digital painting - art piece - based on a visualization of chemical space. Specifically, the figure exhibits an artistic use of the constellation plot illustrated in Figure 3A.

Figure 4. Chemical art example. “Dreaming without pain”. Method: Constellation plots; Dataset: anti-pain and sedative compounds.

4. Conclusions

Chemoinformatics is founded on the concept of combining ideas from chemistry with computational techniques. Advances in the field are often achieved by merging previously ignored ideas and can be easily published and distributed via the internet. Yet, among the many techniques available, there is often no clear best method. Often the best results can be achieved by combining different methods, at the cost of increased computational resources. While many individual methods can be used without programming knowledge, the automatic combination of different methods is difficult even with expert domain knowledge. As part of the training of students and newcomers to the fields, we encourage them to do not necessarily seek the single-best approach to solve a specific problem or try to identify the “one-size-fits-all” (the methodology that will solve all problems as these might usually be complex, as typically happens in drug discovery). Throughout the survey, we have seen that combining methods in different areas such as molecular representation, virtual screening, SP(A)Rs, and chemical space analysis usually offers superior

results as opposed to using only one methodology. A challenge and potential shortcoming is identifying the appropriate combination(s). Lastly, combinations of Chemoinformatics with other non-chemistry areas such as art, open up new avenues to generate novel disciplines. As such, we want to encourage combining chemistry with unusual and artistic fields out of simple curiosity and perhaps discover beauty and novel techniques along the way.

Acknowledgments

We thank fruitful discussions with Prof. Alfonso Mieres Hermosillo, with members and former members of the DIFACQUIM research group. E.L.-L. thanks CONAHCYT, Mexico, for Ph.D. scholarship number 894234.

Funding

We thank the support of DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), Grant No. IG200124. This work was supported by the DFG through the SPP 2363 “Utilization and Development of Machine Learning for Molecular Applications – Molecular Machine Learning”.

Declaration of interest

None of the authors have any conflict of interest related to this paper.

References

1 M. L. Disis and J. T. Slattery, *Sci. Transl. Med.* 2010, **2**, 22cm9.

2 J. Gasteiger, *Chemphyschem*, 2020, **21**, 2233–2242.

3 J. Miranda-Salas, C. Peña-Varas, I. Valenzuela Martínez, D. A. Olmedo, W. J. Zamora, M. A. Chávez-Fumagalli, D. Q. Azevedo, R. O. Castilho, V. G. Maltarollo, D. Ramírez and J. L. Medina-Franco, *Artif. Intell. Life Sci.*, 2023, **3**, 100077.

4 A. Varnek and I. I. Baskin, *Mol. Inform.*, 2011, **30**, 20–32.

5 E. López-López, J. Bajorath and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2021, **61**, 26–35.

6 J. Guillot, S. Maumus-Robert and J. Bezin, *Therapie*, 2020, **75**, 407–416.

7 G. A. Cordell and M. D. Colvard, *J. Nat. Prod.*, 2012, **75**, 514–525.

8 C. Wu, S.-L. Lee, C. Taylor, J. Li, Y.-M. Chan, R. Agarwal, R. Temple, D. Throckmorton and K. Tyner, *J. Nat. Prod.*, 2020, **83**, 552–562.

9 R. A. Houghten, C. Pinilla, M. A. Giulianotti, J. R. Appel, C. T. Dooley, A. Nefzi, J. M. Ostresh, Y. Yu,

G. M. Maggiora, J. L. Medina-Franco, D. Brunner and J. Schneider, *J. Comb. Chem.*, 2008, **10**, 3–19.
10 G. Poli and T. Tuccinardi, *Curr. Bioact. Compd.*, 2020, **16**, 182–190.
11 C. Blanes-Mira, P. Fernández-Aguado, J. de Andrés-López, A. Fernández-Carvajal, A. Ferrer-Montiel and G. Fernández-Ballester, *Molecules*, 2023, **28**, 175.
12 M. D. Mashabela, P. Masamba and A. P. Kappo, *Biology*, 2022, **11**, 1156.
13 P. M. Pflüger, M. Kühnemund, F. Katzenburg, H. Kuchen and F. Glorius, *Chem*, 2024, in press. DOI:10.1016/j.chempr.2024.02.004.
14 M. González-Medina, F. D. Prieto-Martínez, J. R. Owen and J. L. Medina-Franco, *J. Cheminform.*, 2016, **8**, 63.
15 J. J. Naveja and J. L. Medina-Franco, *Front Chem*, 2019, **7**, 510.
16 J. J. Naveja, F. I. Saldívar-González, D. L. Prado-Romero, A. J. Ruiz-Moreno, M. Velasco-Velázquez, R. A. Miranda-Quintana and J. L. Medina-Franco, Visualization, Exploration and Screening of Chemical Space in Drug Discovery. In *Computational Drug Discovery*. Nathan Poongavanam and Vijayan Ramaswamy, ed. Wiley-VCH. 2024. Pp. 365–393.
17 J. L. Medina-Franco, G. M. Maggiora, M. A. Giulianotti, C. Pinilla and R. A. Houghten, *Chem. Biol. Drug Des.*, 2007, **70**, 393–412.
18 D. Gaytán-Hernández, A. L. Chávez-Hernández, E. López-López, J. Miranda-Salas, F. I. Saldívar-González and J. L. Medina-Franco, *J. Cheminform.*, 2023, **15**, 100.
19 P. Ertl, *J. Cheminform.*, 2020, **12**, 8.
20 P. Ertl, *Chemistry Methods*,2022, **2**, e202200041.
21 P. Ertl, *J. Chem. Inf. Model.*, 2022, **62**, 2164–2170.
22 P. Ertl, E. Altmann, S. Racine and R. Lewis, *Eur. J. Med. Chem.*, 2022, **238**, 114483.
23 P. Ertl and B. Rohde, *J. Cheminform.*, 2012, **4**, 12.
24 P. Ertl, E. Altmann, S. Racine and O. Decoret, *Bioorg. Med. Chem.*, 2023, **91**, 117405.
25 H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–113.
26 R. D. Brown, *Perspect. Drug Discov. Des.*, 1996, **7**, 31–49.
27 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
28 X. Xia, E. G. Maliski, P. Gallant and D. Rogers, *J. Med. Chem.*, 2004, **47**, 4463–4470.
29 Daylight theory: SMARTS - A language for describing molecular patterns, https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, (accessed July 16, 2024).
30 S. R. Heller, A. McNaught, I. Pletnev, S. Stein and D. Tchekhovskoi, *J. Cheminform.*, 2015, **7**, 23.
31 N. O'Boyle and A. Dalke, *ChemRxiv*, 2018, DOI:10.26434/chemrxiv.7097960.v1.
32 M. Krenn, F. HÃ¤se, A. Nigam, P. Friederich and A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045024.
33 M. A. Skinnider, *Nat. Mach. Intell.* 2024, **6**, 437–448.
34 J. Menke and O. Koch, *J. Chem. Inf. Model.*, 2021, **61**, 664–675.
35 J. Menke, J. Massa and O. Koch, *Comput. Struct. Biotechnol. J.*, 2021, **19**, 4593–4602.

36 E. Fernández-de Gortari, C. R. García-Jacas, K. Martinez-Mayorga and J. L. Medina-Franco, *J. Cheminform.*, 2017, **9**, 9.
37 N. Sánchez-Cruz and J. L. Medina-Franco, *J. Cheminform.*, 2018, **10**, 55.
38 S. Singh, N. Kaur and A. Gehlot, *Comput. Biol. Med.*, 2024, **179**, 108810.
39 P. S. Charifson, J. J. Corkery, M. A. Murcko and W. P. Walters, *J. Med. Chem.*, 1999, **42**, 5100–5109.
40 L. Tan, S. Hirte, V. Palmacci, C. Stork and J. Kirchmair, *Nat. Rev. Chem.*, 2024, **8**, 319–339.
41 R. Nepravishta, J. Ramírez-Cárdenas, G. Rocha, S. Walpole, T. Hicks, S. Monaco, J. C. Muñoz-García and J. Angulo, *J. Med. Chem.*, 2024, **67**, 10025–10034.
42 A. Liu, S. Seal, H. Yang and A. Bender, *SLAS Discov*, 2023, **28**, 53–64.
43 P. Zhang, D. Zhang, W. Zhou, L. Wang, B. Wang, T. Zhang and S. Li, *Brief. Bioinform.*, 2024, **25**, bbad518.
44 S. Q. Pantaleão, P. O. Fernandes, J. E. Gonçalves, V. G. Maltarollo and K. M. Honorio, *ChemMedChem*, 2022, **17**, e202100542.
45 E. López-López and J. L. Medina-Franco, *Biomolecules*, 2023, **13**, 176.
46 S. P. Collins, B. Mailloux, S. Kulkarni, M. Gagné, A. S. Long and T. S. Barton-Maclaren, *Front. Pharmacol.*, 2024, **15**, 1307905.
47 J. Jiménez-Luna, F. Grisoni, N. Weskamp and G. Schneider, *Expert Opin. Drug Discov.*, 2021, **16**, 949–959.
48 E. López-López, E. Fernández-de Gortari and J. L. Medina-Franco, *Drug Discov. Today*, 2022, **27**, 2353–2362.
49 G. Maggiora, J. L. Medina-Franco, J. Iqbal, M. Vogt and J. Bajorath, *J. Chem. Inf. Model.*, 2020, **60**, 5873–5880.
50 J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2012, **52**, 2485–2493.
51 B. Mouysset, M. Le Grand, L. Camoin and E. Pasquier, *Cancer Lett.*, 2024, **588**, 216800.
52 E. López-López and J. L. Medina-Franco, *Drug Discov. Today*, 2024, **29**, 104046.
53 A. Cichońska, B. Ravikumar and R. Rahman, *Curr. Opin. Struct. Biol.*, 2024, **84**, 102771.
54 M. Rafehi, M. Möller, W. Ismail Al-Khalil and S. M. Stefan, *Pharm. Res.*, 2024, **41**, 411–417.
55 S. Wang, S. Zhu, X. Li and Z. Yang, *Curr. Comput. Aided Drug Des.*, 2024, in press. DOI:10.2174/0115734099282620240521102006.
56 J. Vázquez, R. García, P. Llinares, F. J. Luque and E. Herrero, *J. Comput. Aided Mol. Des.*, 2024, **38**, 18.
57 J. Carlsson and A. Luttens, *Curr. Opin. Struct. Biol.*, 2024, **87**, 102829.
58 A.-J. Banegas-Luna, J. P. Cerón-Carrasco and H. Pérez-Sánchez, *Future Med. Chem.*, 2018, **10**, 2641–2658.
59 C. Bedart, C. V. Simoben and M. Schapira, *Curr. Opin. Struct. Biol.*, 2024, **86**, 102812.
60 N. Kumar and V. Acharya, *Med. Res. Rev.*, 2024, **44**, 939–974.
61 F. Gentile, J. C. Yaacoub, J. Gleave, M. Fernandez, A.-T. Ton, F. Ban, A. Stern and A. Cherkasov,

Nat. Protoc., 2022, **17**, 672–697.

62 D. R. Houston and M. D. Walkinshaw, *J. Chem. Inf. Model.*, 2013, **53**, 384–390.

63 G. M. Maggiora, *J. Chem. Inf. Model.*, 2006, **46**, 1535.

64 I. Muegge and Y. Hu, *Expert Opin. Drug Discov.*, 2022, **17**, 1173–1176.

65 V. J. Gillet, J. D. Holliday and P. Willett, *Mol. Inform.*, 2015, **34**, 598–607.

66 J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1177–1185.

67 P. Willett, *Drug Discov. Today*, 2006, **11**, 1046–1053.

68 D. L. Prado-Romero, A. Gómez-García, R. Cedillo-González, H. Villegas-Quintero, J. F. Avellaneda-Tamayo, E. López-López, F. I. Saldívar-González, A. L. Chávez-Hernández and J. L. Medina-Franco, *Front. Drug Discov.*, 2023, **3**, 1261094.

69 A. M. Virshup, J. Contreras-García, P. Wipf, W. Yang and D. N. Beratan, *J. Am. Chem. Soc.*, 2013, **135**, 7296–7303.

70 D. I. Osolodkin, E. V. Radchenko, A. A. Orlov, A. E. Voronkov, V. A. Palyulin and N. S. Zefirov, *Expert Opin. Drug Discov.*, 2015, **10**, 959–973.

71 I. Casciuc, Y. Zabolotna, D. Horvath, G. Marcou, J. Bajorath and A. Varnek, *J. Chem. Inf. Model.*, 2019, **59**, 564–572.

72 E. López-López, C. M. Cerda-García-Rojas and J. L. Medina-Franco, *Molecules*, 2021, **26**, 2483.

73 J. L. Medina-Franco, A. L. Chávez-Hernández, E. López-López and F. I. Saldívar-González, *Mol. Inform.*, 2022, **41**, e2200116.

74 T. B. Dunn, G. M. Seabra, T. D. Kim, K. E. Juárez-Mercado, C. Li, J. L. Medina-Franco and R. A. Miranda-Quintana, *J. Chem. Inf. Model.*, 2022, **62**, 2186–2201.

75 B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum and A. R. Leach, *Nucleic Acids Res.*, 2024, **52**, D1180–D1192.

76 T. Sander, J. Freyss, M. von Korff and C. Rufener, *J. Chem. Inf. Model.*, 2015, **55**, 460–473.

77 E. López-López, J. J. Naveja and J. L. Medina-Franco, *Expert Opin. Drug Discov.*, 2019, **14**, 335–341.

78 S. K. Niazi and Z. Mariam, *Int. J. Mol. Sci.*, 2023, **24**, 11488.

79 J. Gasteiger, *Molecules*, 2016, **21**, 151.

80 S. Moshawih, H. P. Goh, N. Kifli, A. C. Idris, H. Yassin, V. Kotra, K. W. Goh, K. B. Liew and L. C. Ming, *Chem. Biol. Drug Des.*, 2022, **100**, 185–217.

81 A. K. Yetisen, J. Davis, A. F. Coskun, G. M. Church and S. H. Yun, *Trends Biotechnol.*, 2015, **33**, 724–734.

82 B. Mahjour, J. Bench, R. Zhang, J. Frazier and T. Cernak, *Digital Discovery*, 2023, **2**, 520–530.