Preprints.org

Not peer-reviewed version

Meta-Analysis and Review of *in silico* Methods in Drug Discovery – Part 1: Technological Evolution and Trends from Big Data to Chemical Space

Arife Uzundurukan , Mark Nelson , Christopher Teske , Md Shahidul Islam , Elzagheid Mohamed , John Victor Christy , Holli-Joi Martin , Eugene Muratov , Samantha Glover , DOMENICO FUOCO [*]

Review

Meta-Analysis and Review of In Silico Methods in Drug Discovery—Part 1: Technological Evolution and Trends from Big Data to Chemical Space

Arife Uzundurukan [1], **Mark Nelson** [2], **Christopher Teske** [3], **Mohamed Shahidul Islam** [4], **Elzagheid Mohamed** [5], **John Victor Christy** [6], **Holli-Joi Martin** [7], **Eugene Muratov** [8], **Samantha Glover** [9] and **Domenico Fuoco** [10],*

[1] Engineering Faculty, CRASH-UdS, Université de Sherbrooke, Sherbrooke, Québec, Canada. Arife.Uzundurukan@Usherbrooke.ca; https://orcid.org/0000-0003-1104-1644

[2] Piramal Pharma Solutions, Inc. 18655 Krause St., Riverview, MI 48193, Altoris, Inc., San Diego, CA. USA. Mark.Nelson@Piramal.com

[3] Oncology Research Faculty, Trinity Health, Ann Arbor, MI, USA. Christopher.teske@trinity-health.org; orcid.org/0000-0002-2914-306X

[4] Quality and Compliance Department, BIOVANTEK Global, 10149 chemin de la cote-de-liesse, Montréal, Québec, Canada. quality.one@biovantekglobal.com

[5] Royal Commission for Jubail and Yanbu, Jubail Industrial City, Kingdom of Saudi Arabia. elzagheid_m@rcjy.edu.sa; https://orcid.org/0000-0001-8945-9544

[6] R&D Department, Novacab, Ottawa, Ontario, Canada. johnvchristy@gmail.com https://orcid.org/0000-0002-3343-5532

[7] Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA holli27@email.unc.edu;-https://orcid.org/0000-0002-9123-0439

[7] UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA. murik@email.unc.edu; https://orcid.org/0000-0003-4616-7036

[8] Quantum Business Solution. Beverly Hills, Los Angeles. CA. USA. samantha@qbsbigdata.com

[9] Chemical Engineering, Polytechnique Montréal, Montréal, Québec, Canada. Domenico.Fuoco@polymtl.ca; https://orcid.org/0000-0002-8157-2291

* Correspondence: domenicofuoco@live.ca

Highlights:

- Advancements in computational methods in in silico drug discovery have become a viable option.

- Artificial intelligence and machine learning improve in silico drug discovery by swiftly analyzing data, predicting interactions, and optimizing candidates with precision.

- The recent technological evolution from 1980 to 2024 of in silico methods is discussed.

- We observed rising trends from big data to chemical space.

- An executive summary is structured according to the most cited articles.

- Milestones are assessed based on their respective timelines.

Abstract: Background: The present review summarizes the state-of-the-art of in silico methods and techniques that are the most useful in drug discovery, their relationship with data science, as well as the successful application of data science, machine learning (ML) and artificial intelligence (AI) applications. A meta-analysis of the various technologies available is furthermore proposed as a guideline for the non-expert, reader relative to the several subject areas is also discussed in this article. The scope of this meta-analysis is to rank the enlisted technologies by their field of applications and to depict the latter according to knowledge accessibility, from students to experts. **Method:** The search strategy utilized for this review first produced a general collection of 900 papers without duplications, which were subsequently streamlined and divided into two independent collections: the top 300 most-cited papers of all time (since 2000) and the papers with the highest interest for a systematic review analysis (high-impact exciting papers). **Results:** In Part 1, we discuss the most cited and quality 97 articles in these top 300 papers most relevant to the field of in silico drug discovery. The different disciplines are listed according to their industrial and economic incurred to society, independently from the "metric" results of how many new drug approvals (NDAs) each discipline has generated to date. **Conclusion:**

Big data, the ensemble of **known** items stored in publicly available databases, has improved our understanding of the many fates of a potential drug candidate during its development and even after its commercialization. Moreover, the combination of new screening techniques and "omics" with old drugs has led to a new paradigm in which the **unknown knowledge** of any biological molecule and its cellular structure, now plays an important role as a target for a series of yet-to-be-developed drugs: the chemical space. Furthermore, leveraging big data, data science, ML, and AI can revolutionize drug discovery by swiftly analyzing massive datasets, predicting efficacy and safety profiles, streamlining development, cutting costs, and boosting success rates for new drugs. AI also speeds up the search for promising drug candidates, advancing innovative therapies.

Keywords: data science; big data; data mining; bioinformatics; chemometric; medicinal chemistry; targets; knowledge discovery; artificial intelligence; machine learning; deep learning; data integration; metadata; database; QSAR; collaborative drug discovery; structure-based drug design; ligand-based drug discovery; clinical trials; product development

Graphical Abstract
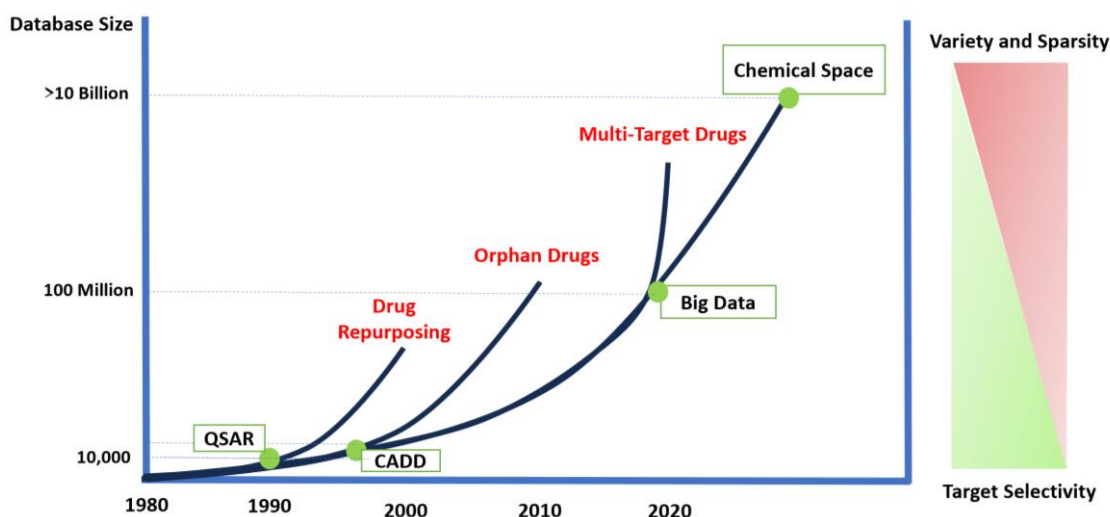
Legend: Timeline evolution and milestones of in silico methods in drug discovery. The main deviation from a typical pharmaceutical pipeline development is indicated in red.

Executive Summary

Definition	Drug discovery is a multidisciplinary science mainly devoted to bringing new chemical entities (NCEs) to the market for already-known medical applications. However, this process is very time-consuming (10-12 years) and highly cost-sensitive (in the range of US $ billions). As a result, scientists have developed computational strategies to maximize the results and yield better and larger product portfolios using less time and less money. These computational strategies are referred to in silico methods (as well as computer-assisted methods) in the literature. Integrating AI, data science, and machine learning into drug discovery streamlines processes, accelerating timelines and optimizing resource allocation. These technologies analyze vast datasets, predict drug efficacy and safety, and enhance accuracy by integrating diverse information sources. This synergy has the potential to revolutionize pharmaceuticals, making drug discovery faster, cost-effective, and more successful in delivering therapeutics to patients.
Needs	The origin of in silico methods dates back to the late '80s. At the time, medicinal chemists were required to produce evidence of structure-activity-relationship (SAR) from molecules within the same pharmacological class of therapeutical agents. This was, for all intents and purposes, essentially an academic exercise. In the '90s, thanks to the advent of the graphical processor unit (GPU), researchers moved from 2D- to 3D- up to multidimensional SAR. This evolution allowed the boom of in silico methods mainly via two techniques: ligand-based fitting and target-based docking. The growing availability of commercial software as well as the open-approach to developing computer-assisted visualization allowed generating an enormous body of publications in this field from a few hundred per year in the late '80s to 10,000 per year in the late '90s. This exponential growth of data prompted the launch of PubMed™ and DrugBank™ in the early 2000s. Following this trend, the largest and most prominent collection of crystallographic structures from Cambridge University has evolved to host several million chemical items and their respective chemical-physical information. Involuntarily, drug discovery entered the big data era.
Opportunity	During the study of new chemical entities (NCEs), medicinal chemists have several possibilities at their disposal to bring NCEs to the market as quickly as possible. Sometimes, the success story arrives via serendipity, other times by trial-and- error and, most recently, during clinical investigations via drug repurposing. Drug repurposing is one of those strategies that has improved mostly in the past ten years. A good example of this is the number of new approved drugs (NADs) granted by the USFDA during the COVID-19 pandemic crisis: 90% of these NADs have emerged as a result of drug repurposing.
Gap	One of the main forces driving the advance in drug discovery was the search for high selectivity of biological targets, following the principle: one molecule for one specific binding site. However, the inevitable overlap of data stemming from computational visualization, clinical trials, and medical reports has confirmed that this principle of one molecule for one binding site is an excellent theory in a perfect world, but in reality, highly difficult to achieve.

	The most recent trend in drug discovery is the multi-target drug approach (MTD). In this type of approach, the in silico methods are suitable to discriminate via ab initio the relationship between molecular activity, specific end-user population, and drug-drug (and drug-receptor) interaction upstream of clinical investigations. Moreover, the evolution of databases via web-based server applications has created a brand-new field in the search for novel drugs, called chemical space.
Recommendation	For more than a century, the R&D departments of the major pharmaceutical companies were known to be a highly secretive environment with tens of thousands of chemists working on individual and personalized clean benches, and not communicating outside of their close circle of team members. Today's setting is completely different. The loss of patent protections for the most important blockbusters, and the legal dispute surrounding patents relating to drug repurposing uses has reduced the necessity for this "special environment" in the pharmaceutical industry. Consequently, there has been a large layoff of highly skilled pharmaceutical professionals. Paradoxically, computer-assisted technologies have not produced the expected number of novel solutions, and the emergence of artificial intelligence (AI) now requires a new generation of skilled scientists in this new area going from data science to network pharmacology. Moreover, given the implicit difficulty in finding such multi-task medicinal chemists, the current R&D departments of the major pharmaceutical companies are increasingly becoming incubators/accelerators of start-ups and consortiums of public-private partnerships in which the various stakeholders are invited to interact and collaborate, especially in remote mode. *In silico* methods have not only brought out the digitalization of pharmaceutical science, but the entire manner in which scientists think about the definition of drug-biology interactions. The next step of in silico methods is the merging of chemical laboratory automation and synthetic tissues (organ-on-chip). At this juncture, even clinical trial investigations will become obsolete, and will run purely via computers: i. e. in silico medicine.

Introduction

   Drug discovery is a multidisciplinary science by which new chemical entities (NCE) are proposed for use as new therapeutic agents and to be administered during clinical trial phases until their final approval as new medications, also commonly known as active pharmaceutical ingredient (API) [1]. Drug design and drug discovery are long-established sciences, predating the emergence of informatics. These fields have developed autonomously and have been exhaustively analyzed and consolidated as comprehensively described in the historical overview by Pina in 2010 [2].

   Historically, new drugs were discovered by identifying the active ingredients from traditional remedies such as plants or by serendipitous discovery, e.g. when the identification of the NCE as an API occurred outside of the scope of the main investigation as in the case of penicillin and Viagra™ [3].

   Despite advances in technology and a better understanding of cellular pathophysiological signaling downstream, drug discovery remains both time-consuming (up to 10 years of the investigatory program) and capital-intensive (from $ US 800 million up to $ US 4 billion budget for

just one new API). It is moreover a highly competitive and regulated sector. Therefore, from a purely financial perspective, drug discovery is an inefficient industrial process [4].

More recently, chemical libraries of small synthetic molecules, natural products, or botanical extracts are being screened via computer-aided systems (also known as **virtual screening**), allowing new opportunities for understanding receptor docking/fitting and the mechanism of action [5]. One of the results of this approach is to elucidate alternative pharmacological models for new medical indications, classified under the label of investigational new drug (IND). This type of drug discovery process, based on the finding of a new medical indication and subsequently applied to an already-existent API, is commonly called **drug repositioning** (or repurposing) [6].

Traditional methods of discovering drugs have been effective, although they can take over 12 years and cost approximately USD 1.8 billion from identifying a lead candidate to conducting clinical trials [7]. Moreover, the attrition rate of candidate drugs can reach as high as 96% [7]. In recent times, in silico approaches have gained increasing consideration due to their ability to hasten the drug discovery process in terms of time, labor, and expenses. Owing to novel computational methods, several new drug compounds have been successfully developed.

Lately, an increasing number of scientists are delving into computer-assisted drug discovery (CADD) methods to address the challenges of conventional experimental approaches, such as scale, time, and cost [8].

In silico **methods** refer to the current practice of drug discovery via the use of specific software, artificial intelligence (AI), together with. IT infrastructures such as high-speed computer systems, web-based servers/databases, and cloud. These methods thus differ from traditional drug discovery approaches performed through biological assays defined as **in vitro** methods, versus pre-clinical and clinical trials or **in vivo** methods.

In general, scientists are well-equipped to handle large sets of data comprising several thousand molecules and their **properties**. Specifically, chemists are trained to re-elaborate large data sets of compounds (and information related to these compounds) for their classification purposes before and after applying statistical regressions, as long as there is an evident rationale: namely, **pattern recognition and molecular clustering** [9–11]. Both of these techniques are crucial in terms of drug delivery, as well as aiding in the design [8], optimization, and understanding of delivery systems [12]. The ensemble of data spanning all possible compounds pertaining to chemistry and its applications is called **chemical space** [13].

Nevertheless, the ever-increasing amount of data being collected in public repositories/databases, having reached such proportions in terms of volume and velocity, has led in the last decade to the advent of a new science called **big data** [14,15]. The main difference between large datasets and big data is the total lack in the latter of any apparent homogeneity among the data *per se* [14–16]. In 2018, the number of single chemical entities available for in silico screening overpassed the symbolic threshold of 1 billion compounds. Furthermore, contract research organizations (CROs) offering drug discovery services are already announcing the target of 8 billion chemical entities for the year 2024 [17]. Certain academic teams are even proposing virtual libraries in the range of 11 billion chemical entities [18,19]. The elements of big data are, by definition, a combination of structured, semi-structured, and unstructured datasets. In this review, we discuss how big data is affecting the productivity of chemists working in drug discovery and development, as well as offer an explanatory overview of the rationale and impact of the various selected topics in this field [20].

AI and machine learning (ML), deep learning big data have now become integral components in advancing drug delivery systems [21–23]. These technologies offer innovative solutions for designing, optimizing, and personalizing drug delivery, leading to more efficient and effective therapeutic outcomes [14].

This large number of data, generated from genomics, proteomics and metabolomic sciences, combinatorial chemistry, and automated high-throughput screening (HTS), has led to a new trend in drug discovery for **data integrations and data mining** and its related disciplines [14]. The extremely vast use of public databases from all categories of end-users, ranging from professionals to students,

has compelled hosting server providers to release several versions of software capable of data mining and graphical visualization simultaneously, to better serve their users for the many possible applications including virtual screening. The main areas of these applications are known as **bioinformatics** and **cheminformatics** [25,26].

In this review, we discuss how big data is affecting the productivity of chemists working in drug discovery and development, as well as offer an explanatory overview of the rationale and impact of the various selected topics in this field. Additionally, we provide an execution summary to facilitate understanding and discuss the large amount of data. To better understand the various trends and global technological evolution of this vast and complex field of study, a meta-analysis of the literature is first proposed, starting from the year 2000. This review contributes as a crucial tool in identifying and examining patterns, themes, and developments that have emerged over the past two decades by delving into the extensive body of literature considering mostly cited papers in the areas.

QSAR: Quantitative-Structure-Activity-Relationship

The main purpose of CADD until the early 2000s was to understand the QSAR of therapeutic agents used in therapy as a means to introduce novel generations of the most selective drugs into the therapeutical arsenal. The result of this gold rush" has translated into more than 30,000 APIs currently available in pharmacies.

CADD involves using computers to identify possible drug targets, screen large chemical libraries to find effective drug candidates, optimize candidate compounds, and assess their potential toxicity. Once these processes are digitally completed, the candidate compounds are then tested in vitro or in vivo to confirm their effectiveness. Consequently, CADD techniques can decrease the number of chemical compounds that need experimental evaluation while increasing the success rate by eliminating inefficient and toxic chemical compounds from consideration [27]. Pharmacophore modelling, QSARs, and AI have been the latest methods used in ligand-based virtual screening methodology [27].

Table 1 depicts the visual summary of the topics (disciplines and sub-disciples) discussed in this review. Each topic is listed as part of big data (pink color) or chemical space (grey color) according to the major relevance of the 300 most cited papers on in silico drug discovery chosen in this review. Key technology areas (blue) such as bioinformatics and cheminformatics are a common background for both big data and chemical space contents. The first result of the analysis of the evolution of big data assumed that one drug can be exclusively selective for one target. However, clinical evidence shows that this exclusivity and selectivity of one molecular target per drug is only theoretical.

Drug repositioning is a growing approach to drug discovery powered by the necessity to achieve greater "approved medical use" for the same API that is on the verge of losing its intellectual property protection status, or already out of the market because it is *de facto* considered an old drug. In the past, drug repositioning was driven by serendipity as well as by the off-label use of both marketed and generic drugs by physicians in their clinical practice. However, there are ongoing efforts to systematically conduct drug repurposing. The tetracyclines are a prime example of old drugs that are still of interest for new medical purposes. Indeed, for more than 60 years of their history, tetracyclines have been used as antibiotics while in the last 20 years, they have also been used in anticancer treatment as well as for neuroprotection [28,29].

Table 1. Interplay of current technology and their applications in drug discovery.

		Big Data				
Data Source	Data Collection	Data Integration	Data Validation	Data Application	Data Visualization	Data Real-World Evidence
in vivo	Diseases					

Disciplines and sub-disciplines by relevance					Active pharmaceutical ingredient (API) "One drug for one molecular target"	Pharmacophore and molecular target	Current use of API in clinical practice "One drug for many molecular targets, simultaneously"
		Clinical outcome	Translational Bioinformatics	Quantitative and Qualitative Structure Activity Relationship (QSAR)			
	ex vivo	Gene therapy	Bioinformatics				
		Metabolites					
	in vitro	OMICS					
		Biochemistry					
	in silico	Medicinal chemistry	Cheminformatics	Virtual screening and 3D models			
		Chemical physics					
		Predictive models					
Key technologies	DATABASE: • Web-based server • IT-cloud • Extremely fast computing systems	Data mining coupled with • Machine learning • Deep learning • Artificial intelligence	Graphical processing unit (GPU): • Dedicated CRO	Drug repurposing • IND	Network pharmacology	Multi-target drugs	
Chemical Space							

Methodology of the Meta-Analysis

The carried out initial search spanned from January 2023 to May 2023 to identify the most qualified and cited papers from the early 2000s to 2022 related to technological evolution and trends from big data to chemical space for in silico methods in drug discovery. A total of 1475 articles were identified across the three databases, PubMed, Web of Science, and Embase. A comprehensive literature search was conducted to identify relevant articles pertaining to drug discovery, big data in chemical space, QSAR, drug repurposing, orphan drugs, multi-target drugs, and computer-aided drug design (CADD). The search spanned three major databases: PubMed, Web of Science, and Embase. Key terms, including "drug discovery," "big data to chemical space," "QSAR," "drug repurposing," "orphan drugs," "multi-target drugs," and "CADD," were used to retrieve potential articles. Subsequently, duplicates were removed to streamline the dataset, resulting in 900 unique articles. From the second pool of articles, the top 300 most cited papers from the early 2000s to 2022, providing an overview of the most influential works in the field. The second sub-collection aimed at a systematic review analysis and included articles deemed of significant interest for this study. The inclusion criteria for the meta-analysis were meticulously defined to ensure the relevance and reliability of the chosen articles in Figure 1. A total of 97 articles were selected for inclusion in the meta-analysis, adhering to the eligibility criteria outlined in the decision-making strategies presented in the accompanying table.

The article selection process involved multiple stages of review and was conducted independently by the authors involved. Rigorous decision-making strategies were employed to ensure the precision and accuracy of the final selection of articles for inclusion in the meta-analysis.

This methodology employed a systematic and comprehensive approach to identify, select, and categorize articles for a meta-analysis, ensuring the robustness and reliability of the findings.
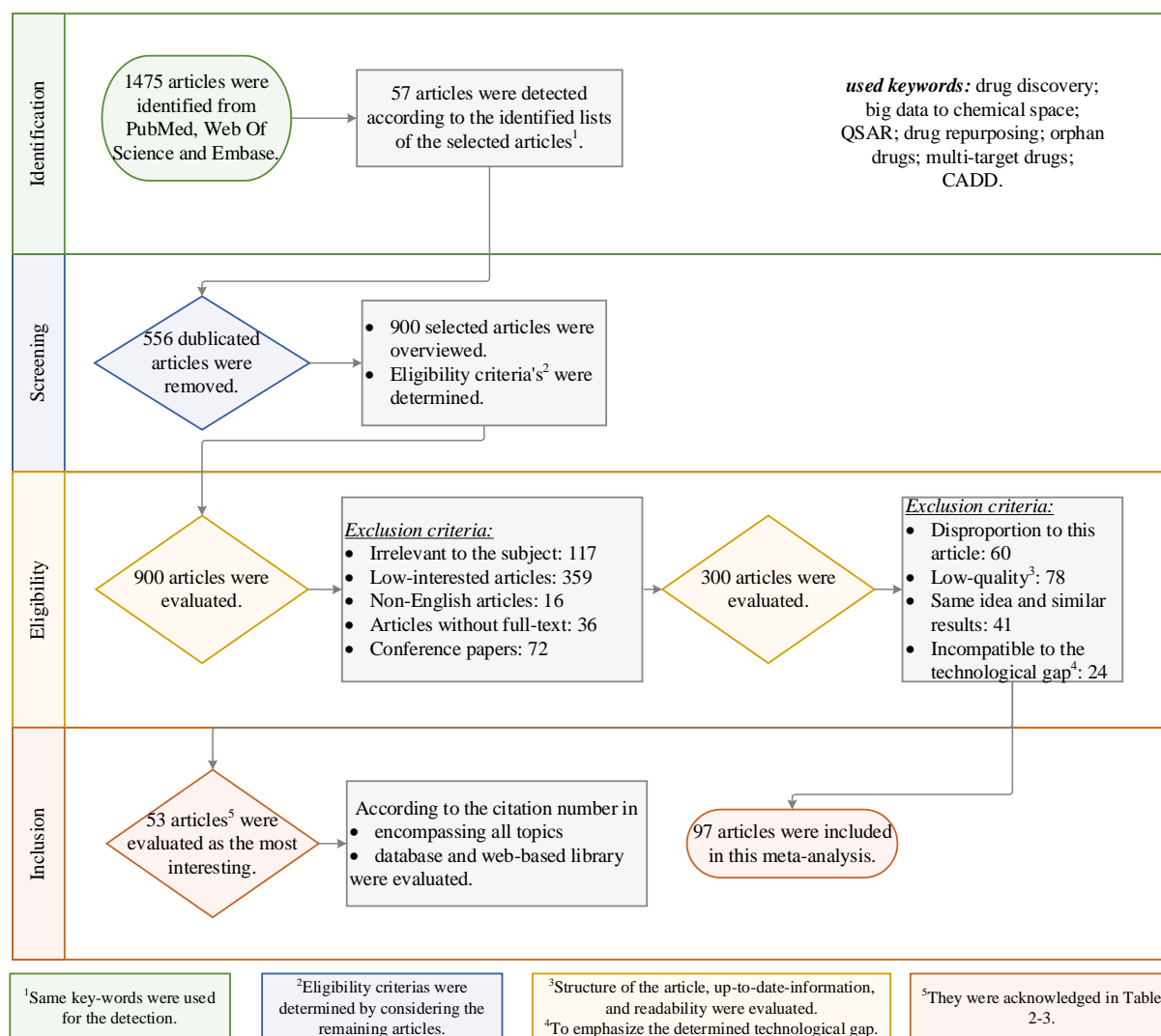

Figure 1. Article inclusion and exclusion criteria with the numbers.

Results

General Consideration: "Quality" of the Data

The definitions of **big data** and c**hemical space** are primarily based on the sole notion of "quantity" of the data, their storage, and their various fields of application. To date, there are no means to convey a measure of the "quality" of the data that are part of big data and chemical space.

Indeed, the interpretation of the validity (and reliability, accuracy, completeness, and consistency) of data is left to the researchers and the protocols established by their academic institutions [30,31].

While the goal of this review is not to delve into an extensive discussion regarding the "quality" of data, we nonetheless provide some details regarding the different sources of data and the overall relationship between the "quantity" of data and their impact on science and society [32], [33].

Given that the purpose of this review is to measure the "metric relevance" and the type of contribution of the different reported areas of interest, the absolute number of citations per article reported by Google Scholar is taken into consideration.

There is a large body of research as to whether or not Google Scholar can be qualified as a suitable source of scientific information that has been produced from 2017 to 2022 [31,34]. The result of the present search is in support of Google Scholar over other providers [34], particularly in a case

such as the current meta-analysis aimed at measuring the impact on the web of each sub-field that comprises the very vast subject of in silico drug discovery.

The mean number of published scientific papers per year in the field of drug discovery has dramatically grown at a pace of tenfold over the decades, passing from the range of hundreds/year in the late '80s to a few thousand/year in '90, and more than tens of thousands per year in the 2000s. AI, ML, and deep learning techniques greatly improve data quality assessment in big data and chemical space. They identify issues like inconsistencies and missing values, automate cleansing, and prioritize data for validation. Deep learning extracts complex features, enhancing accuracy. Leveraging these techniques ensures data reliability, improving scientific research impact in drug discovery. By leveraging these advanced techniques, researchers can ensure the reliability and integrity of data within big data and chemical space, ultimately enhancing the effectiveness and impact of scientific research in drug discovery.

For this review, the absolute number of citations per single article elaborated by Google Scholar is used to establish the impact of each selected article on the web. The dataset of the most cited papers reported herein has been limited to the first 300 papers in the field of drug discovery. The total amount of citations reported for these sole 300 papers corresponds to over 400,000 citations. The dataset chosen for this literature review is extremely significant in understanding the evolution and trends of in silico drug discovery over the last 22 years. It is reasonable to estimate that the chosen dataset is representative of more than 95% of all top-cited articles of the subject area.

Validation of the Proposed Meta-Analysis

The present meta-analysis involves the statistical synthesis of results from multiple studies to obtain a summary estimate of the effect size. In addition to the provided inclusion and exclusion criteria to maintain consistency and relevance in the studies selected, the latter are furthermore supported by other analyses such as VosViever as illustrated in Figure 2 which maps the analysis of the 100 most frequent keywords from over 1,000 articles published in peer-reviewed journals listed in Web of Science (2000-2022). This bibliometric analysis in the form of a network is very useful in gaining a better understanding of the interplay and degree of congruence among the various subject areas.
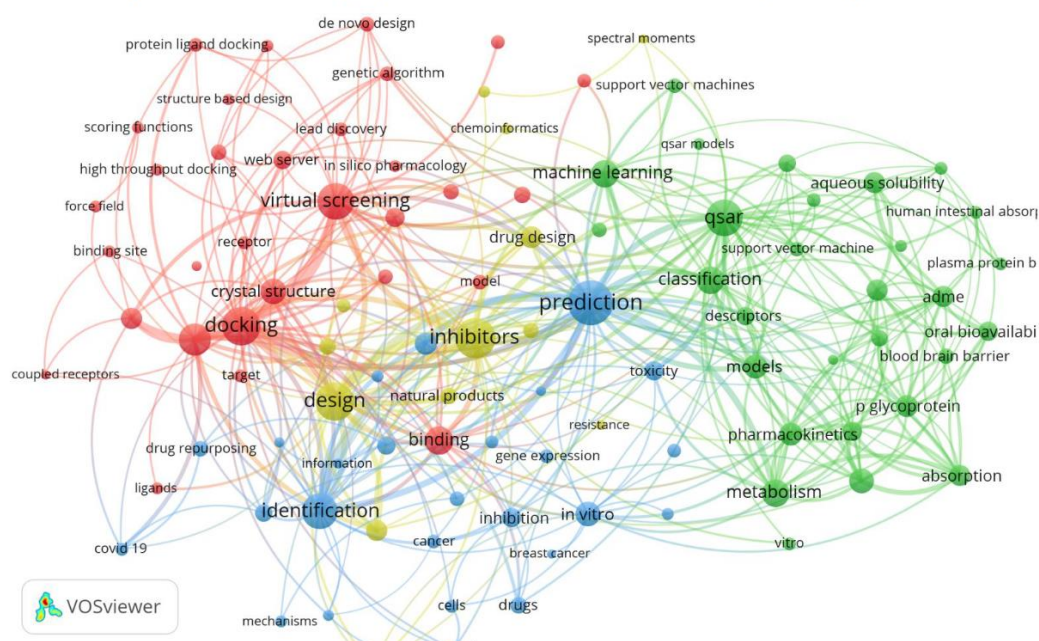

Figure 2. Bibliometric map derived via the publicly-available software VosViewer™ of the 100 most frequent keywords from over 1,000 articles published in peer-reviewed journals listed in Web of Science (2000-2022).

The distribution of the keywords generated 5 different clusters of articles, depicted in red, green, yellow, blue and (purple, mauve).

The keyword "prediction" emerged as the central point of this bibliometric map. Accordingly, the main purpose of in silico drug discovery is to predict a novel drug in terms of chemical structure and/or in terms of the mechanism of action.

The bibliometric analysis in Figure 3. in the form of a network is highly useful in illustrating the interplay and degree of connection among the different subject areas. This visualization proposes how the information and interest of the articles based on "virtual screening" have shifted from metabolic analysis and pharmacokinetics (2010) to molecular mechanisms, web servers and COVID-19, starting in 2016.

Network Map 2: Focusing on "virtual screening" as one of the main topics in drug discovery.

Figure 3. Bibliometric analysis illustrating the interplay and degree of congruence of the articles based on "virtual screening" that have migrated from metabolic analysis and pharmacokinetics (2010) to molecular mechanisms, web servers, and COVID-19, starting in2016.

The use of a heatmap is one of the first steps of the rationalization process (cause vs. effect) in which researchers are looking for similarities between very close fields of investigation within the same subject domain, or area.

The use of a color-shade gradient provides a visual interpretation of the degree of trends.

Primary Outcomes of this Meta-Analysis

The primary objective of this meta-analysis was to determine the qualitative and quantitative ranking/classification among the different disciplines of in silico drug discovery: for example, research centers and authors that are involved in the development of software for predictive measure of drugs involved in the development of predictive analytics software for drug monitoring/drug discovery are not necessarily involved in the elaboration of databases or chemical space concepts.

The second objective of this meta-analysis was to understand the relationship and the impact of data science methods in *'in silico'* drug discovery, if any.

As shown in Figures 4 and 5 below, the findings of the present meta-analysis provide a measure of the impact of the different fields and approach methods within the same subject domain. The most

important discriminant is the total number of citations and the ponderal numbers per year, also commonly known as weighted value per year.
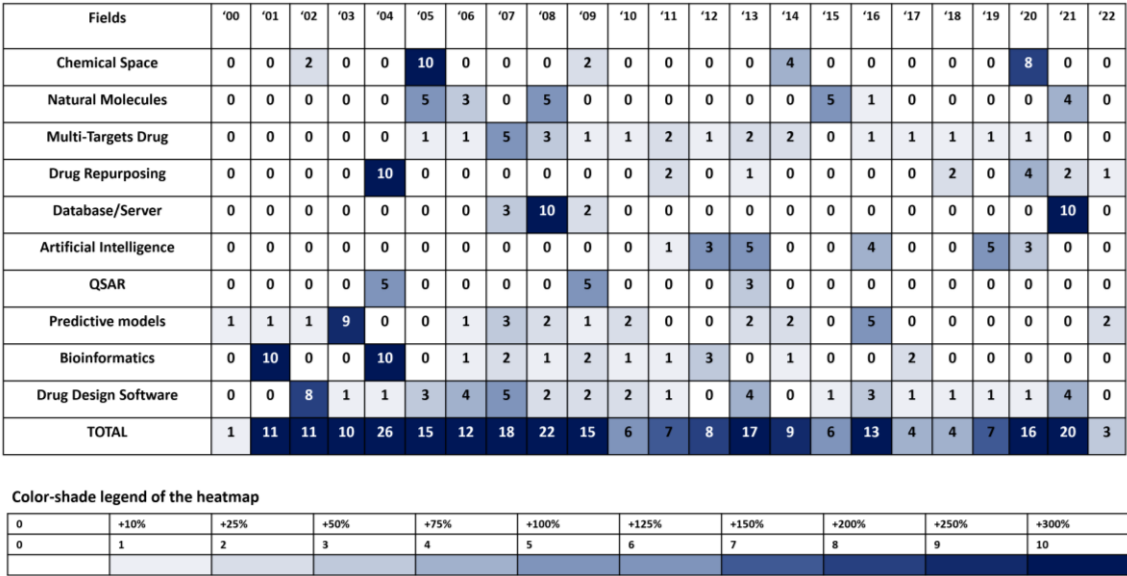
Fields	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	'17	'18	'19	'20	'21	'22
Chemical Space	0	0	2	0	0	10	0	0	0	2	0	0	0	0	4	0	0	0	0	0	8	0	0
Natural Molecules	0	0	0	0	0	5	3	0	5	0	0	0	0	0	0	5	1	0	0	0	0	4	0
Multi-Targets Drug	0	0	0	0	0	1	1	5	3	1	1	2	1	2	2	0	1	1	1	1	1	0	0
Drug Repurposing	0	0	0	0	10	0	0	0	0	0	0	2	0	1	0	0	0	0	2	0	4	2	1
Database/Server	0	0	0	0	0	0	0	3	10	2	0	0	0	0	0	0	0	0	0	0	0	10	0
Artificial Intelligence	0	0	0	0	0	0	0	0	0	0	0	1	3	5	0	0	4	0	0	5	3	0	0
QSAR	0	0	0	0	5	0	0	0	0	5	0	0	0	3	0	0	0	0	0	0	0	0	0
Predictive models	1	1	1	9	0	0	1	3	2	1	2	0	0	2	2	0	5	0	0	0	0	0	2
Bioinformatics	0	10	0	0	10	0	1	2	1	2	1	1	3	0	1	0	0	2	0	0	0	0	0
Drug Design Software	0	0	8	1	1	3	4	5	2	2	2	1	0	4	0	1	3	1	1	1	1	4	0
TOTAL	1	11	11	10	26	15	12	18	22	15	6	7	8	17	9	6	13	4	4	7	16	20	3

Color-shade legend of the heatmap

0	+10%	+25%	+50%	+75%	+100%	+125%	+150%	+200%	+250%	+300%
0	1	2	3	4	5	6	7	8	9	10

Figure 4. The impact of the different studies involved in in silico drug discovery. This heatmap reports the weighted values, which are extrapolated according to the 300 most cited papers in the respective fields. The full data is available in the Excel file in the supplementary section. Each cell represents the ponderal number per year of absolute citations (Google Scholar Number) of the 300 most cited papers on in silico drug discovery. Additional details regarding the use of the weighted values/year are discussed in the text and the table of reference is reported in an Excel file.
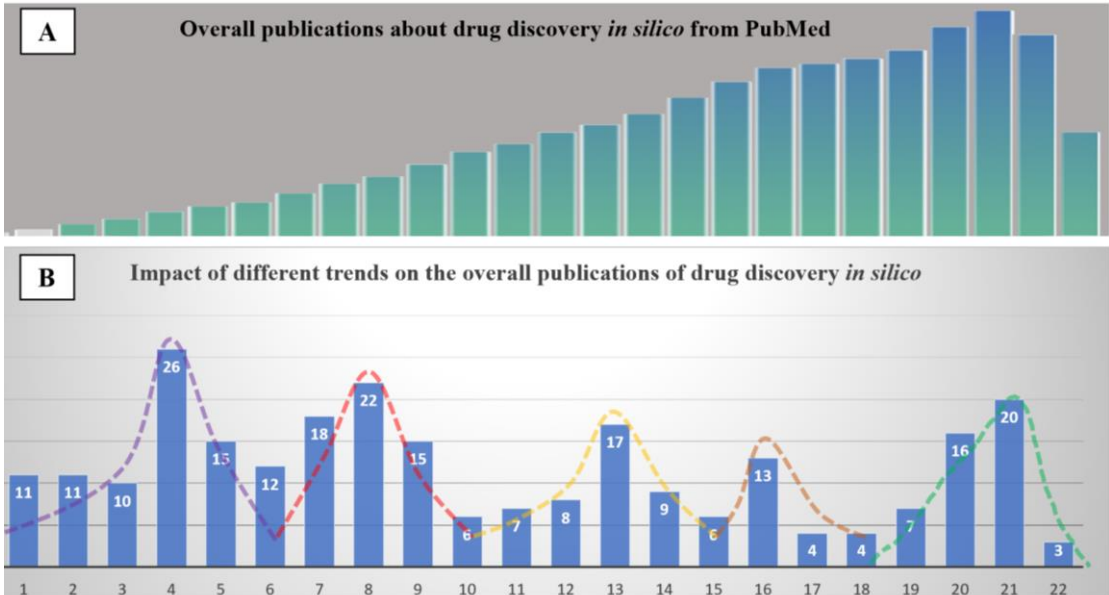

Figure 5. Visualization of (A) the sum of total publications per year and (B) their impact on the most cited articles over time. Total number of publications reported in PubMed about drug discovery (A). Results of the analysis of trends for the same time span (B). The different periods are depicted by dashed curves. Each color shows one of the five different periods that have emerged from the meta-analysis: purple, red, orange, brown, and green. The blue columns show the relative importance of the 300 selected papers, (on year per year basis). The numbers reported in this lower portion of panel B, are the same as those reported in the heatmap.

Although the total number of publications reported in PubMed has grown exponentially over the last two decades, with very minimal yearly reduction in pace, analysis of the data depicted in Figures 3 and 4B highlight five very distinct periods unequivocally related to the different trends and having specific impacts on scientific publications. This portrayal of the different trends is a result of the different keywords and topics of the 300 most cited publications as reported from Google Scholar, namely:

Period#1: 2000 – 2005, with a maximum in 2004: 59/300 articles for more than 145 K citations

Period#2: 2006 – 2010, with a maximum in 2008: 64/300 articles for more than 130 K citations

Period#3: 2011 – 2014, with a maximum in 2013: 50/300 articles for more than 100 K citations

Period#4: 2015 – 2017, with a maximum in 2016: 46/300 articles for more than 95 K citations

Period#5: 2018 – 2022, with a maximum in 2021: 79/300 articles for more than 160 K citations

Other extrapolated data from the several technological areas cited so far in this review:

Artificial Intelligence: 8 articles – ca 6,000 citations

Bioinformatics: 33 articles – ca 110,000 citations

Chemiometrics: 28 articles – ca 22,000 citations

Natural Molecule: 11 articles – ca 22,000 citations

Predictive Models : 84 articles – ca 85,000 citations

Drug Design Software: 15 articles – ca 16,000 citations

Chemical Space : 14 articles – ca 14,000 citations
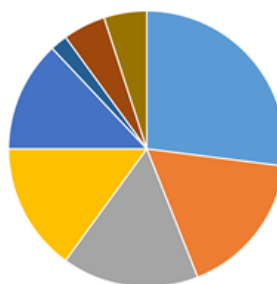
Database: 35 articles – 100,000 citations

Drug Repurposing: 16 articles – ca 33,000 citations

Muti-Target Drugs: 16 articles – ca 8,000 citations

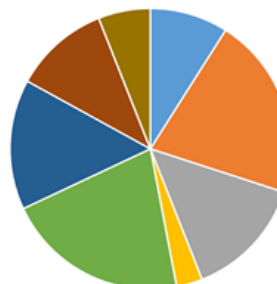QSAR: 40 articles – ca 35,000 citations

2000 – 2005
- Database (Cambridge, PubMed)
- Binding affinity
- QSAR (Chemical Similarity vs Diversity)
- HIV
- Bioinformatics
- Target-Based Discovery
-

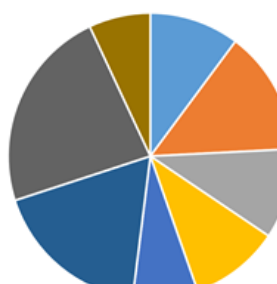59/300 articles were selected, cited more than 145k times

2006 – 2010
- Translational Bioinformatics
- Database (DrugBank,PubChem)
- QSAR (Drug-like Concept)
- Drug Repurposing
-

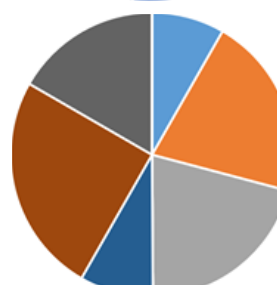70/300 articles were selected, cited more than 84k times

2011 – 2014
- Database (OpenBabel,
- QSAR (Orphan Drug,
- Personalized Medicine
- Drug Repurposing
- Network/System Pharmacology
- Big data and Chemical Space
-

48/300 articles were selected. cited more than 59k times

2015 – 2017
- Database (ChEMBL,
- IBM WATSON
- Virtual screening
-

47/300 articles were selected, cited more than 58k times

2018 – 2022
- Artificial Intelligence
- Data-Driven Conceptualization
- Antibodies
- COVID-19
-
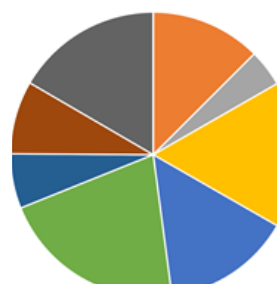
76/300 articles were selected, cited more than 48k times

■ Bioinformatics   ■ Drug Design Software   ■ Predictive Models   ■ Chemical Space   ■ Drug Repurposing
■ Database   ■ Multi-Targets Drug   ■ Natural Molecule   ■ Artificial Intelligence   ■ QSAR
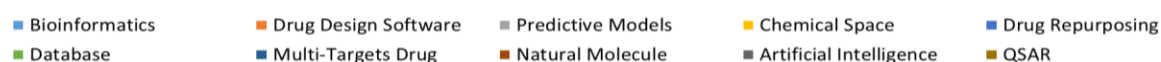
Figure 6. Variation in topics used in the field of in silico drug discovery methods over time. The relative impacts of each time period over the entirely different subject areas were taken into consideration.

The average number of drugs approved (NDA) by the USFDA per year since 1980 was 43 (series 1) while the average number of new chemical entities (NCE) per year was 14 (series 2) over the same time period (Figure 7, panel A). The difference in terminology NDA vs. NCE lies in the fact that an NDA can be a second-generation drug or any other modification (and/or bioisotere) of a previously

approved NCE structure, although NCE is referred to as a chemical moiety of the NDA. The USFDA assesses candidate drugs under final review with two different designations from a panel of experts. Of note, until 2012, the near totality of the NCEs was designated under the category of "Small Molecules", while after 2012 antibodies, large peptides, and proteins were introduced as NCEs. However, as can be seen in Figure 5B, the very large availability of data did not necessarily lead to an increase in the number of approved drugs nor the variety of chemical structures of the novel approved therapeutics.
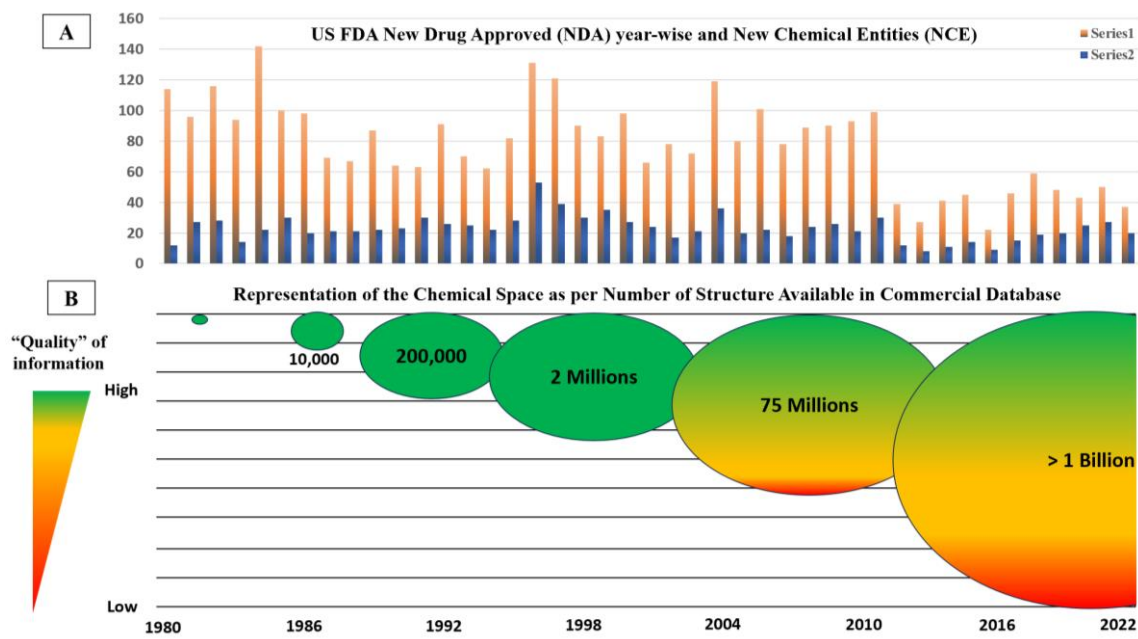
Figure 7. Visualization of (A) the average number of drugs approved (NDA) by the USFDA per year, and (B) big data.

Table 2. List of the most cited papers encompassing all topics.

Number	Topic	Title	First author	Year	Citations	Ref
1	Bioinformatics	Initial sequencing and analysis of the human genome	HG Consortium	2001	28696	[35]
2	Database	The Cambridge Structural Database: a quarter of a million crystal structures and rising	FH Allen	2002	14390	[36]
3	Bioinformatics	Bioconductor: open software development for computational biology and bioinformatics	RC Gentleman	2004	13462	[37]
4	Database	The human genome browser at UCSC	WJ Kent	2002	10583	[38]
5	Database	The Cambridge structural database	CR Groom	2016	7366	[39]
6	Database	OpenBabel: An open chemical toolbox	NM O Boyle	2011	6735	[40]
7	Bioinformatics	A review of feature selection techniques in bioinformatics	Y Saeys	2007	5556	[41]

8	Software	CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields	K Vanommeslaeghr	2010	5130	[42]
9	Bioinformatics	The Perseus computational platform for comprehensive analysis of omics data	S Tyanova	2016	5068	[43]
10	Database	DrugBank 5.0: a major update to the DrugBank database for 2018	DS Wishart	2018	4951	[44]
11	Bioinformatics	Accurate whole human genome sequencing using reversible terminator chemistry	DR Bentley	2001	4691	[45]
12	Database	Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening	TA Halgren	2004	4234	[46]
13	Model Predict	Lead- and drug-like compounds: the rule-of-five revolution	CA Lipinski	2004	4213	[47]
14	Model Predict	Drug-like properties and the causes of poor solubility and poor permeability	CA Lipinski	2000	3912	[48]
15	Bioinformatics	Biopython: freely available Python tools for computational molecular biology and bioinformatics	PJA Cock	2009	3885	[49]
16	Database	PubChem substance and compound databases	S Kim	2016	3853	[50]
17	Bioinformatics	The druggable genome	AL Hopkins	2002	3851	[51]
18	Model Predict	Drug discovery: a historical perspective	J Drews	2000	3582	[52]
19	Database	New software for searching the Cambridge Structural Database and visualizing crystal structures	IJ Bruno	2002	3581	[53]
20	Database	DrugBank: resource in silico drug discovery and exploration	DS Wishart	2006	3544	[54]
21	Bioinformatics	From genomics to chemical genomics: new developments in KEGG	M Kaneshisa	2006	3473	[55]
22	Chemical Space	Network pharmacology: the next paradigm in drug discovery	AL Hopkins	2008	3438	[56]
23	Database	ChEMBL: a large-scale bioactivity database for drug discovery	A Gaulton	2012	3390	[57]
24	Database	HMDB 3.0--The Human Metabolome Database in 2013	DS Wishart	2012	3061	[58]
25	QSAR	Random forest: a classification and regression tool for compound classification and QSAR modelingD	V Svetnik	2003	3052	[59]
26	Database	HMDB 4.0: the human metabolome database for 2018	DS Wishart	2018	2928	[60]

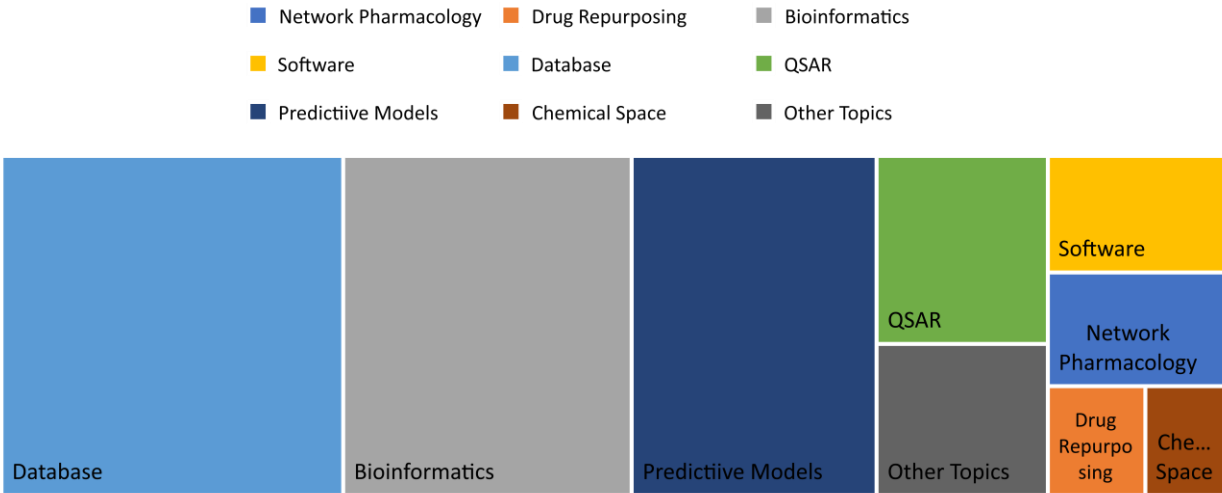| 27 | **Drug Repositioning** | Drug repositioning: identifying and developing new uses for existing drugs | TT Ashburn | 2004 | 2914 | [61] |
| 28 | **Database** | DrugBank: a knowledgebase of drugs, drug actions, and drug targets | DS Wishart | 2008 | 2785 | [62] |

Figure 8. Tree map of the most cited papers reported by topics.

Table 3. List of the most cited papers in Database and Web-based Library.

Number	Title	First Author	Year	Citation	Reference
1	The Cambridge Database: a quarter of a million structure and rising	FH Allen	2002	14390	[36]
2	The Human Genome Browser at UCSC	WJ Kent	2002	10583	[38]
3	New software for searching the Cambridge Structural Database and visualizing crystal structures	IJ Bruno	2002	3581	[53]
4	Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening	TA Halgren	2004	4234	[46]
5	DrugBank: a comprehensive resource for in silico drug discovery and exploration	DS Wishart	2006	3544	[54]
6	BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities	T Liu	2007	1663	[63]
7	ChEBI: a database and ontology for chemical entities of biological interest	K Deglyarenko	2007	1155	[64]
8	PubChem: Integrated platform of small molecules and biological actives	EE Bolton	2008	1537	[65]
9	DrugBank: a knowledgebase for drugs, drug actions, and drug targets	DS Wishart	2008	2785	[62]
10	PubChem: a public information system for analyzing bioactivities of small molecules	Y Wang	2009	1317	[66]

| 11	DrugBank 3.0: a comprehensive resource for 'OMICS' research on drugs	C Knox	2010	2034	[67]
12	Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database	PCD Hawkins	2010	1367	[68]
13	Open Babel: An open chemical toolbox	NM O Boyle	2011	6735	[40]
14	ChEMBL: a large-scale bioactivity database for drug discovery	A Gaulton	2012	3390	[57]
15	HMDB 3.0—The Human Metabolome Database in 2013	DS Wishart	2012	3061	[58]
16	ZINC: a free tool to discover chemistry for biology	JJ Irwin	2012	2481	[69]
17	ChEMBL: a large-scale bioactivity database for drug discovery	A Gaulton	2012	1782	[57]
18	DrugBank 4.0: Shedding new light on drug metabolism	V Law	2014	2103	[70]
19	The Cambridge structural database	CR Groom	2016	7366	[39]
20	PubChem substance and compound databases	S Kim	2016	3853	[50]
21	The ChEMBL database in 2017	A Gaulton	2017	1731	[71]
22	DrugBank 5.0: a major update to the DrugBank database for 2018	DS Wishart	2018	4951	[44]
23	HMDB 4.0: the human metabolome database for 2018	DS Wishart	2018	2928	[60]
24	PubChem 2019 update improved access to chemical data	S Kim	2019	2380	[72]
25	PubChem in 2021: new data content and improved web interfaces	S Kim	2021	1657	[73]
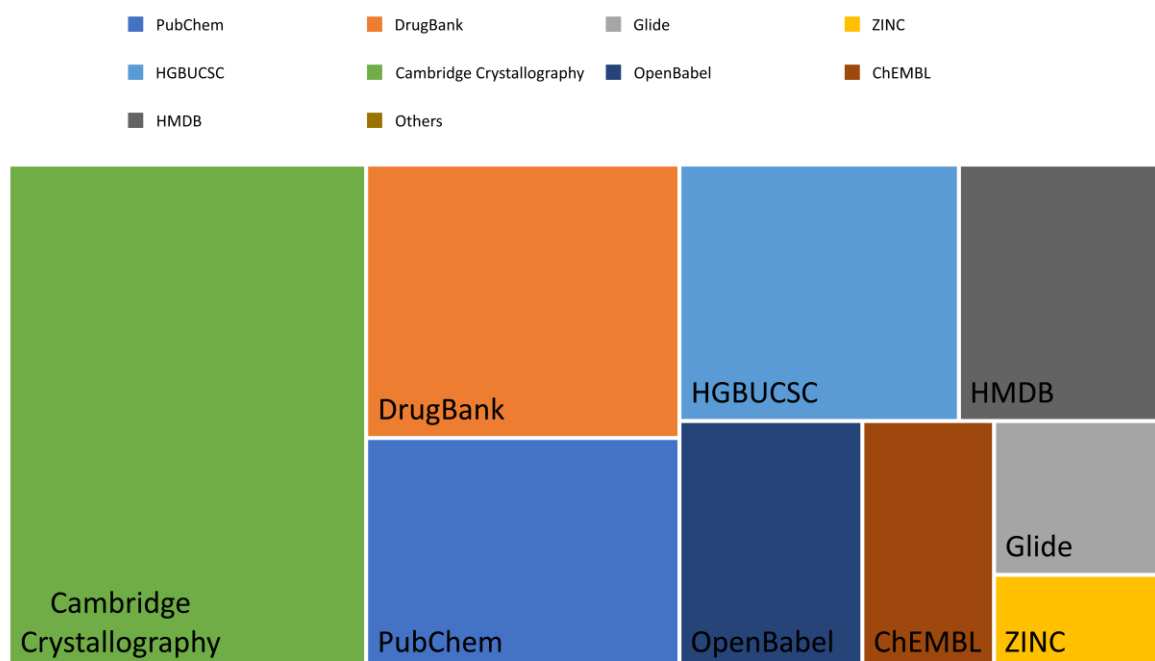
Figure 9. Tree map of the sub-categories of the main topic: DATABASE/Web-based Library. The list of the most cited papers encompassing all topics for Cambridge Crystallography, DrugBank, PubChem, HGBUCSC, HMDB, OpenBabel, ChEMBL, Glide, and ZINC are shown.

Discussion, Limitations/Uncertainty

The availability of the many compounds documented to date (accounting for active fragments of small and large molecules) has allowed exposure to a myriad of hypotheticals. The most significant result of the "OMICS age" is the growing amount of information introduced by genomics and proteomics in real-time. This wealth of data has provided innumerable targets that could theoretically combine with as many hits for an equal number of active pharmaceutical ingredients [74].

In 2002, Oprea proposed a dataset of one million compounds tested via HTS. However, it would appear that the probability of a compound reaching market status is one in a million [169] .

A mere 10 years later, PubChem had attained 89 million chemical entities in its database such that 89 new drugs could theoretically be obtained as a maximum for this dataset, or one could search for different targets in the same dataset and obtain 89 new drugs every time [75,76].

Millions of scientists from very different domains are familiar with the theory known as the "***Lock-Key Model***", the father of which is the 1908 Nobel Prize for Medicine laureate, Sir Paul Ehrlich. The concept that one key can open a lock is at the forefront of molecular pharmacology and, in the last ten years, has evolved into a much more complex paradigm in which a specific key (drug) can unlock a series of locks not necessarily interrelated, exactly as in the case of a master key.

Sir Paul Ehrlich was aware that there is no dogma in science; in fact, he was the first to introduce the concept of ***Zauberkugel*** (*"magic bullet"*). The technology that has been available in the last ten years has shown the full applicability of this paradigm, more than a hundred years after its enunciation, under the terminology of **multi-target drugs**.

Period #1 – "Testing the reliability of database vs. the selectivity concepts."

Until the development of web-based library resources, the only force leading the global field of in silico drug discovery was the analysis, via homemade software, of Quantitative-Structure-Activity-Relationship (QSAR) studies, albeit via a limited private library of just a few thousand structures. The principle behind the QSAR analysis was based on the chemical similarity of proven therapeutic agents for the same diseases.

The availability of the first large web-based database, such as PubMed, compelled researchers to introduce new methods in drug discovery, namely data mining and data integration.

In 2004, Pearsons and others [77] designed a good example of data integration among various libraries, from genes vs. a series of bioactive compounds, to link these results to very specific cellular targets or at least to understand whether any correlated pathways existed.

In 2005, Csermely and coll., [78] had already noted that despite the availability of new digital tools in drug discovery, the additional amount of data was not contributing to generating more therapeutical options. On the other hand, they began confirming the new paradigm of multi-target drugs which was the real novelty of this new era of computer-assisted drug design. As stated by the authors:

"At present, the goal is to:

(i)   find a target of suitable function.

(ii)   identify the 'best binder' by high-throughput screening of large combinatorial libraries and/or by rational drug design based on the three-dimensional structure of the target.

(iii)   provide a set of proof-of-principle experiments.

(iv)   develop a technology platform that predicts potential clinical applications.

However, despite all the careful studies and the considerable drug-development efforts undertaken, the number of successful drugs and novel targets did not increase appreciably during the past decade."

Period #2 – "Looking for applications of novel drugs inside an old pharmacological space."

In 2006, a multi-disciplinary team from Pfizer, [79] published a first-ever map of the known pharmacological space at the time. The purpose of this colossal work was to plot the direction or best fitting of more than 275,000 biologically active compounds vs. a library of 1,000 molecular targets. The result of this experience was the introduction of the concepts of chemical promiscuity of therapeutic agents for the same disease and degrees of draggability of cellular targets.

Another interesting novelty of this type of approach is the maturity of bioinformatics and translational bioinformatics.

On another hand, the ensemble of this new evidence of non-pre-determinability of the biological effect, or the obvious lack of exclusive selectivity for the "target", has induced the purists of the quantitative structure activity-relationship (QSAR) realm to introduce additional dimensions of interactions. The QSAR hence shifted from a mere proprieties of the molecular scaffold (2D) up to induced conformation of the target via environmental forces (3D up to 6D), Lill 2007 [80] and Furches 2010 [81].

Period #3 – "Data-driven drug discovery."

The most successful result of the early 2010s in the field of drug discovery has been in admitting that one approved drug can have multiple medical indications. This new paradigm based on the concept of drug repurposing produced a record number of 37 designated and marketed orphan drugs in only two years, from 2009-2010, Muthyala [82] .

The easy access to large libraries and real-time collaborative platforms with the major pharmaceutical companies, together with the possibility of exchanging clinical information for new medical indications without incurring legal issues, has fueled the exponential growth of publications and thematic journals. The results of this enormous amount of data unintentionally catapulted drug discovery into the big data era (Lusher 2014 [83]).

Paradoxically, these same digital tools, which were introduced to facilitate in silico drug discovery, also became somewhat of an obstacle requiring the introduction of a new class of researcher to be added to the medicinal chemists' team, namely the data scientist dedicated to managing and streamlining big data [84].

Period #4 – "Druggability: one drug and many targets"

The growing consensus for muti-target drugs and the search for specific genetic targets dramatically changed the perspective and the scope of the medicinal chemist [85]. The total lack of target selectivity is not due to the chemical moiety per se but rather to the dominant concept of

network pharmacology and phenotypic selection [86]. While medicinal chemists usually rely on some version of the "lock and key" paradigm to design novel therapeutics, modern pharmacology recognizes that the mid- and long-term effects of a given drug on a biological system may depend not only on specific ligand-target recognition events but also on the influence of the repeated administration of a drug on cell-gene signature (Talevi 2015 [87], Lavecchia, 2016 [88]).

Period #5 – "Re-discovering the Use of Natural Molecules for Complex Diseases."

Since 2018, drug repositioning methods have been leading the success of the entire sector of in silico drug discovery, including the search for remedies for COVID-19 treatment. Indeed, COVID-19 is unequivocally a complex disease, consisting not only of a simple viral infection but a complicated framework of symptoms and unpredictable transmissibility (as a result of the variants) with varying outcomes [89] according to the different patient populations (de Oliviera et al. 2018 [90], Ramsay et al. 2018 [91]).

Other fields of interest in this period include neurology [92], with Alzheimer's disease leading the topics for drug repurposing (Mesiti 2019 et al. (2019) [93]), as well as two other areas related to namely seizures (Geets et al. (2019) [94]) and depression Buenz et al. (2018) [92].

In totally different pathways, cannabinoids (and Cannabis extracts) have resulted in a very interesting natural remedy owing to the success of Epidolex™ in the treatment of seizures in children Namdar et al. (2020) [95] [96].

Finally, following the re-discovering of natural therapy, the latest success in time is the drug repurposing of classical abuse/recreational psychedelics such as psilocybin (and dried "magic mushrooms" powder) for depression and PTDS (Mullowney et al. (2023) [97]).

Conclusion

The significant role of in silico tools has increased exponentially in the last ten years and has markedly revolutionized the field of CADD and brought about technological evolution and recent trends that have galvanized the search for new drugs. All in all, the search approach employed in this review initially yielded a comprehensive set of 900 non-duplicated papers. In Part 1, our focus shifts to the examination of the top 300 papers, identifying the most cited and high-quality 97 articles pertinent to the realm of in silico drug discovery.

A final consideration should also be given as to how these technological trends are affecting the socio-economic aspects of pharmaceutical companies, including:

- *Big Data: Data Science, Data Integration and Data Mining*

The result of these disciplines is the continuing improvement in IT infrastructure and software. Incorporating advanced AI and ML techniques can enhance in silico drug discovery by rapidly analyzing vast amounts of data, predicting molecular interactions, and optimizing drug candidates with higher precision and efficiency.

- *Cheminformatics: classification, pattern recognition and clustering*

The result of these disciplines is the improvement in current knowledge regarding the mechanisms of actions of drugs owing to a better understanding of their QSAR.

- *Bioinformatics and Translational Bioinformatics*

The result of these disciplines is the seeking of new molecular targets and alternative physio/pathological mechanisms downstream. In particular, these disciplines are pivotal for the understanding of epigenetics and meta-genomics phenomena.

- *Drug Repurposing*

The primary scope of this discipline is the life extension of expired patent applications. In practice, the controversy related to patent issues and the ensuing transfer of Drug Master Files (DMF) have accelerated the need for collaborative models among pharmaceutical stakeholders. The result is the shrinking of the required internal R&D workforce.

- *Chemical Space*

The chemical space is the ensemble of all possible chemical structures, which is believed to contain up to several billion molecules of potential interest for drug discovery as mentioned in this review. One proposed means to explore chemical space is based on the selection of virtual libraries of common scaffold-tree algorithms (grouping) that are overlapped to other "organized maps" of chemical-physical information (pharmacokinetics) and/or chemical interactions (pharmacodynamics). The result of this "spatial analysis" has been used in the last 20 years to generate a new discipline known as Network Medicine or Network Pharmacology. Conversely to drug repurposing strategies in which one molecule ("old API") is investigated for a new medical indication following the principle of one drug for one targe the results in Network Pharmacology approaches to drug discovery are completely different. Indeed, in terms of the definition of chemical space, there are billions of potential drugs that could virtually match billions of targets. The best matching combinations are hence known as multi-target drugs (MTD), including in Food, Aroma, and other fields.

References

1. Stanley H. Nusim, *Active Pharmaceutical Ingredients*, vol. 205. 2005.
2. A. S. Pina, A. Hussain, and A. C. A. Roque, "An Historical Overview of Drug Discovery," in *Ligand-Macromolecular Interactions in Drug Discovery: Methods and Protocols*, A. C. A. Roque, Ed., Totowa, NJ: Humana Press, 2010, pp. 3–12. doi: 10.1007/978-1-60761-244-5_1.
3. T. A. Ban, "The role of serendipity in drug discovery," *Clinical research*, vol. 8, no. 3, 2006.
4. S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, and D. Greyson, "The cost of drug development: a systematic review," *Health Policy*, vol. 100, no. 1, pp. 4–17, Apr. 2011, doi: 10.1016/j.healthpol.2010.12.002.
5. C. Xu *et al.*, "Systemic In Silico Screening in Drug Discovery for Coronavirus Disease (COVID-19) with an Online Interactive Web Server," *J. Chem. Inf. Model.*, vol. 60, no. 12, pp. 5735–5745, Dec. 2020, doi: 10.1021/acs.jcim.0c00821.
6. S. Pushpakom *et al.*, "Drug repurposing: progress, challenges and recommendations," *Nat Rev Drug Discov*, vol. 18, no. 1, pp. 41–58, Jan. 2019, doi: 10.1038/nrd.2018.168.
7. "How to improve R&D productivity: the pharmaceutical industry's grand challenge | Nature Reviews Drug Discovery." Accessed: Nov. 28, 2023. [Online]. Available: https://www.nature.com/articles/nrd3078
8. "CADD and Informatics in Drug Discovery | SpringerLink." Accessed: Nov. 28, 2023. [Online]. Available: https://link.springer.com/book/10.1007/978-981-99-1316-9
9. S. Nag *et al.*, "Deep learning tools for advancing drug discovery and development," *3 Biotech*, vol. 12, no. 5, p. 110, May 2022, doi: 10.1007/s13205-022-03165-8.

10.   C. Wu, R. C. Gudivada, B. J. Aronow, and A. G. Jegga, "Computational drug repositioning through heterogeneous network clustering," *BMC Systems Biology*, vol. 7, no. Suppl 5, p. S6, 2013, doi: 10.1186/1752-0509-7-S5-S6.

11.   Y. Chen and J. Kirchmair, "Cheminformatics in Natural Product-based Drug Discovery," *Mol Inform*, vol. 39, no. 12, p. e2000171, Dec. 2020, doi: 10.1002/minf.202000171.

12.   W. Chen, X. Liu, S. Zhang, and S. Chen, "Artificial intelligence for drug discovery: Resources, methods, and applications," *Molecular Therapy - Nucleic Acids*, vol. 31, pp. 691–702, Mar. 2023, doi: 10.1016/j.omtn.2023.02.019.

13.   A. Cherkasov, "The 'Big Bang' of the chemical universe," *Nat Chem Biol*, vol. 19, no. 6, pp. 667–668, Jun. 2023, doi: 10.1038/s41589-022-01233-x.

14.   H. Zhu, "Big Data and Artificial Intelligence Modeling for Drug Discovery," *Annu Rev Pharmacol Toxicol*, vol. 60, pp. 573–589, Jan. 2020, doi: 10.1146/annurev-pharmtox-010919-023324.

15.   S. M. Schüssler-Fiorenza Rose *et al.*, "A longitudinal big data approach for precision health," *Nat Med*, vol. 25, no. 5, pp. 792–804, May 2019, doi: 10.1038/s41591-019-0414-6.

16.   B. Chen and A. Butte, "Leveraging big data to transform target selection and drug discovery," *Clin Pharma and Therapeutics*, vol. 99, no. 3, pp. 285–297, Mar. 2016, doi: 10.1002/cpt.318.

17.   D. Neri and R. A. Lerner, "DNA-Encoded Chemical Libraries: A Selection System Based on Endowing Organic Compounds with Amplifiable Information," *Annu. Rev. Biochem.*, vol. 87, no. 1, pp. 479–502, Jun. 2018, doi: 10.1146/annurev-biochem-062917-012550.

18.   A. A. Sadybekov *et al.*, "Synthon-based ligand discovery in virtual libraries of over 11 billion compounds," *Nature*, vol. 601, no. 7893, pp. 452–459, Jan. 2022, doi: 10.1038/s41586-021-04220-9.

19.   J. Lyu, J. J. Irwin, and B. K. Shoichet, "Modeling the expansion of virtual screening libraries," *Nat Chem Biol*, vol. 19, no. 6, Art. no. 6, Jun. 2023, doi: 10.1038/s41589-022-01234-w.

20.   T. Hart and L. Xie, "Providing data science support for systems pharmacology and its implications to drug discovery," *Expert Opin Drug Discov*, vol. 11, no. 3, pp. 241–256, 2016, doi: 10.1517/17460441.2016.1135126.

21.   L. David, A. Thakkar, R. Mercado, and O. Engkvist, "Molecular representations in AI-driven drug discovery: a review and practical guide," *J Cheminform*, vol. 12, no. 1, p. 56, Sep. 2020, doi: 10.1186/s13321-020-00460-5.

22.   S. Vatansever *et al.*, "Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: State-of-the-arts and future directions," *Med Res Rev*, vol. 41, no. 3, pp. 1427–1473, May 2021, doi: 10.1002/med.21764.

23.   "Drug discovery with explainable artificial intelligence | Nature Machine Intelligence." Accessed: Nov. 28, 2023. [Online]. Available: https://www.nature.com/articles/s42256-020-00236-4

24.   X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery," *Chem. Rev.*, vol. 119, no. 18, pp. 10520–10594, Sep. 2019, doi: 10.1021/acs.chemrev.8b00728.

25.   S. K. Wooller, G. Benstead-Hume, X. Chen, Y. Ali, and F. M. G. Pearl, "Bioinformatics in translational drug discovery," *Biosci Rep*, vol. 37, no. 4, p. BSR20160180, Jul. 2017, doi: 10.1042/BSR20160180.

26.   P. Carracedo-Reboredo *et al.*, "A review on machine learning approaches and trends in drug discovery," *Comput Struct Biotechnol J*, vol. 19, pp. 4538–4558, 2021, doi: 10.1016/j.csbj.2021.08.011.

27.   M. D. Segall and C. Barber, "Addressing toxicity risk when designing and selecting compounds in early drug discovery," *Drug Discov Today*, vol. 19, no. 5, pp. 688–693, May 2014, doi: 10.1016/j.drudis.2014.01.006.

28.   D. Fuoco, "Classification Framework and Chemical Biology of Tetracycline-Structure-Based Drugs," *Antibiotics*, vol. 1, no. 1, pp. 1–13, Jun. 2012, doi: 10.3390/antibiotics1010001.

29.   G. Subramanian, A. M. M. Mjalli, and M. E. Kutz, "Integrated approaches to perform in silico drug discovery," *Curr Drug Discov Technol*, vol. 3, no. 3, pp. 189–197, Sep. 2006, doi: 10.2174/157016306780136790.

30.   G. Halevi, H. Moed, and J. Bar-Ilan, "Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature," *Journal of Informetrics*, vol. 11, no. 3, pp. 823–834, Aug. 2017, doi: 10.1016/j.joi.2017.06.005.

31.   F. Radicchi, A. Weissman, and J. Bollen, "Quantifying perceived impact of scientific publications," *Journal of Informetrics*, vol. 11, no. 3, pp. 704–712, 2017.

32.   R. Sauvayre, "Types of Errors Hiding in Google Scholar Data," *J Med Internet Res*, vol. 24, no. 5, p. e28354, May 2022, doi: 10.2196/28354.

33.   E. Mayo-Wilson, T. Li, N. Fusco, K. Dickersin, and MUDS investigators, "Practical guidance for using multiple data sources in systematic reviews and meta-analyses (with examples from the MUDS study)," *Res Synth Methods*, vol. 9, no. 1, pp. 2–12, Mar. 2018, doi: 10.1002/jrsm.1277.

34.   M. Gusenbauer and N. R. Haddaway, "Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources," *Res Synth Methods*, vol. 11, no. 2, pp. 181–217, Mar. 2020, doi: 10.1002/jrsm.1378.

35.   E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, Art. no. 6822, Feb. 2001, doi: 10.1038/35057062.

36.   F. H. Allen, "The Cambridge Structural Database: a quarter of a million crystal structures and rising," *Acta Cryst B*, vol. 58, no. 3, pp. 380–388, Jun. 2002, doi: 10.1107/S0108768102003890.

37.   R. C. Gentleman *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, p. R80, Sep. 2004, doi: 10.1186/gb-2004-5-10-r80.

38.   W. J. Kent *et al.*, "The Human Genome Browser at UCSC," *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jan. 2002, doi: 10.1101/gr.229102.

39.   C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The Cambridge Structural Database," *Acta Cryst B*, vol. 72, no. 2, pp. 171–179, Apr. 2016, doi: 10.1107/S2052520616003954.

40.   N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *J Cheminform*, vol. 3, no. 1, p. 33, Dec. 2011, doi: 10.1186/1758-2946-3-33.

41.   Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007, doi: 10.1093/bioinformatics/btm344.

42.   K. Vanommeslaeghe *et al.*, "CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," *J Comput Chem*, vol. 31, no. 4, pp. 671–690, Mar. 2010, doi: 10.1002/jcc.21367.

43.   S. Tyanova *et al.*, "The Perseus computational platform for comprehensive analysis of (prote)omics data," *Nat Methods*, vol. 13, no. 9, pp. 731–740, Sep. 2016, doi: 10.1038/nmeth.3901.

44.   D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018, doi: 10.1093/nar/gkx1037.

45.   D. R. Bentley *et al.*, "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, no. 7218, Art. no. 7218, Nov. 2008, doi: 10.1038/nature07517.

46.   T. A. Halgren *et al.*, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening," *J. Med. Chem.*, vol. 47, no. 7, pp. 1750–1759, Mar. 2004, doi: 10.1021/jm030644s.

47.   C. A. Lipinski, "Lead- and drug-like compounds: the rule-of-five revolution," *Drug Discovery Today: Technologies*, vol. 1, no. 4, pp. 337–341, Dec. 2004, doi: 10.1016/j.ddtec.2004.11.007.

48.   C. A. Lipinski, "Drug-like properties and the causes of poor solubility and poor permeability," *Journal of Pharmacological and Toxicological Methods*, vol. 44, no. 1, pp. 235–249, Jul. 2000, doi: 10.1016/S1056-8719(00)00107-6.

49.   P. J. A. Cock *et al.*, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009, doi: 10.1093/bioinformatics/btp163.

50.   S. Kim *et al.*, "PubChem Substance and Compound databases," *Nucleic Acids Res*, vol. 44, no. D1, pp. D1202–D1213, Jan. 2016, doi: 10.1093/nar/gkv951.

51.   "The druggable genome | Nature Reviews Drug Discovery." Accessed: Dec. 09, 2023. [Online]. Available: https://www.nature.com/articles/nrd892

52.   J. Drews, "Drug Discovery: A Historical Perspective," *Science*, vol. 287, no. 5460, pp. 1960–1964, Mar. 2000, doi: 10.1126/science.287.5460.1960.

53.   I. J. Bruno *et al.*, "New software for searching the Cambridge Structural Database and visualizing crystal structures," *Acta Crystallogr B*, vol. 58, no. Pt 3 Pt 1, pp. 389–397, Jun. 2002, doi: 10.1107/s0108768102003324.

54.   D. S. Wishart, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, no. 90001, pp. D668–D672, Jan. 2006, doi: 10.1093/nar/gkj067.

55.   M. Kanehisa *et al.*, "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D354-357, Jan. 2006, doi: 10.1093/nar/gkj102.

56.   A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nat Chem Biol*, vol. 4, no. 11, Art. no. 11, Nov. 2008, doi: 10.1038/nchembio.118.

57.   A. Gaulton *et al.*, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, no. D1, pp. D1100–D1107, Jan. 2012, doi: 10.1093/nar/gkr777.

58.   D. S. Wishart *et al.*, "HMDB 3.0--The Human Metabolome Database in 2013," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D801-807, Jan. 2013, doi: 10.1093/nar/gks1065.

59.   V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *J Chem Inf Comput Sci*, vol. 43, no. 6, pp. 1947–1958, 2003, doi: 10.1021/ci034160g.

60.   D. S. Wishart *et al.*, "HMDB 4.0: the human metabolome database for 2018," *Nucleic Acids Res*, vol. 46, no. D1, pp. D608–D617, Jan. 2018, doi: 10.1093/nar/gkx1089.

61.   "Drug repositioning: identifying and developing new uses for existing drugs | Nature Reviews Drug Discovery." Accessed: Dec. 09, 2023. [Online]. Available: https://www.nature.com/articles/nrd1468

62.   D. S. Wishart *et al.*, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, no. suppl_1, pp. D901–D906, Jan. 2008, doi: 10.1093/nar/gkm958.

63.   T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities," *Nucleic Acids Res*, vol. 35, no. Database issue, pp. D198-201, Jan. 2007, doi: 10.1093/nar/gkl999.

64. K. Degtyarenko *et al.*, "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D344-350, Jan. 2008, doi: 10.1093/nar/gkm791.

65. E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, "Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities," in *Annual Reports in Computational Chemistry*, vol. 4, R. A. Wheeler and D. C. Spellmeyer, Eds., Elsevier, 2008, pp. 217–241. doi: 10.1016/S1574-1400(08)00012-1.

66. Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "PubChem: a public information system for analyzing bioactivities of small molecules," *Nucleic Acids Research*, vol. 37, no. Web Server, pp. W623–W633, Jul. 2009, doi: 10.1093/nar/gkp456.

67. C. Knox *et al.*, "DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs," *Nucleic Acids Research*, vol. 39, no. suppl_1, pp. D1035–D1041, Jan. 2011, doi: 10.1093/nar/gkq1126.

68. P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, and M. T. Stahl, "Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database," *J. Chem. Inf. Model.*, vol. 50, no. 4, pp. 572–584, Apr. 2010, doi: 10.1021/ci100031x.

69. J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "ZINC: A Free Tool to Discover Chemistry for Biology," *J. Chem. Inf. Model.*, vol. 52, no. 7, pp. 1757–1768, Jul. 2012, doi: 10.1021/ci3001277.

70. V. Law *et al.*, "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic Acids Research*, vol. 42, no. D1, pp. D1091–D1097, Jan. 2014, doi: 10.1093/nar/gkt1068.

71. A. Gaulton *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Research*, vol. 45, no. D1, pp. D945–D954, Jan. 2017, doi: 10.1093/nar/gkw1074.

72. S. Kim *et al.*, "PubChem 2019 update: improved access to chemical data," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1102–D1109, Jan. 2019, doi: 10.1093/nar/gky1033.

73. S. Kim *et al.*, "PubChem in 2021: new data content and improved web interfaces," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1388–D1395, Jan. 2021, doi: 10.1093/nar/gkaa971.

74. H. Matthews, J. Hanison, and N. Nirmalan, "'Omics'-Informed Drug and Biomarker Discovery: Opportunities, Challenges and Future Perspectives," *Proteomes*, vol. 4, no. 3, Sep. 2016, doi: 10.3390/proteomes4030028.

75. S. Kim, "Getting the most out of PubChem for virtual screening," *Expert Opinion on Drug Discovery*, vol. 11, no. 9, pp. 843–855, Sep. 2016, doi: 10.1080/17460441.2016.1216967.

76. D. Fuoco, "Hypothesis for changing models: current pharmaceutical paradigms, trends and approaches in drug discovery," *PeerJ*, vol. 3, Feb. 2015, doi: 10.7287/peerj.preprints.813v1.

77. A. B. Parsons *et al.*, "Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways," *Nat Biotechnol*, vol. 22, no. 1, pp. 62–69, Jan. 2004, doi: 10.1038/nbt919.

78. P. Csermely, V. Agoston, and S. Pongor, "The efficiency of multi-target drugs: the network approach might help drug design," *Trends in Pharmacological Sciences*, vol. 26, no. 4, pp. 178–182, Apr. 2005, doi: 10.1016/j.tips.2005.02.007.

79. G. V. Paolini, R. H. B. Shapland, W. P. Van Hoorn, J. S. Mason, and A. L. Hopkins, "Global mapping of pharmacological space," *Nat Biotechnol*, vol. 24, no. 7, pp. 805–815, Jul. 2006, doi: 10.1038/nbt1228.

80. M. A. Lill, "Multi-dimensional QSAR in drug discovery," *Drug Discovery Today*, vol. 12, no. 23–24, pp. 1013–1017, Dec. 2007, doi: 10.1016/j.drudis.2007.08.004.

81. D. Fourches, E. Muratov, and A. Tropsha, "Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research," *J. Chem. Inf. Model.*, vol. 50, no. 7, pp. 1189–1204, Jul. 2010, doi: 10.1021/ci100176x.

82. R. Muthyala, "Orphan/rare drug discovery through drug repositioning," *Drug Discovery Today: Therapeutic Strategies*, vol. 8, no. 3–4, pp. 71–76, Dec. 2011, doi: 10.1016/j.ddstr.2011.10.003.

83. S. J. Lusher, R. McGuire, R. C. Van Schaik, C. D. Nicholson, and J. De Vlieg, "Data-driven medicinal chemistry in the era of big data," *Drug Discovery Today*, vol. 19, no. 7, pp. 859–868, Jul. 2014, doi: 10.1016/j.drudis.2013.12.004.

84. D. Fuoco, "A new method for characterization of natural zeolites and organic nanostructure using atomic force microscopy," *Nanomaterials*, vol. 2, no. 1, pp. 79–91, 2012.

85. D. Fuoco *et al.*, "Identifying nutritional, functional, and quality of life correlates with male hypogonadism in advanced cancer patients," *ecancermedicalscience*, vol. 9, 2015, Accessed: Feb. 04, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4544574/

86. D. Fuoco, "Cytotoxicity induced by tetracyclines via protein photooxidation," *Advances in Toxicology*, vol. 2015, 2015, Accessed: Feb. 04, 2024. [Online]. Available: https://downloads.hindawi.com/archive/2015/787129.pdf

87. A. Talevi, "Multi-target pharmacology: possibilities and limitations of the 'skeleton key approach' from a medicinal chemist perspective," *Front. Pharmacol.*, vol. 6, Sep. 2015, doi: 10.3389/fphar.2015.00205.

88. A. Lavecchia and C. Cerchia, "In silico methods to address polypharmacology: current status, applications and future perspectives," *Drug Discovery Today*, vol. 21, no. 2, pp. 288–298, Feb. 2016, doi: 10.1016/j.drudis.2015.12.007.

89.   F. Cheng *et al.*, "Network-based approach to prediction and population-based validation of in silico drug repurposing," *Nat Commun*, vol. 9, no. 1, p. 2691, Jul. 2018, doi: 10.1038/s41467-018-05116-5.

90.   J. D. O. Viana, M. B. Félix, M. D. S. Maia, V. D. L. Serafim, L. Scotti, and M. T. Scotti, "Drug discovery and computational strategies in the multitarget drugs era," *Braz. J. Pharm. Sci.*, vol. 54, no. spe, Nov. 2018, doi: 10.1590/s2175-97902018000001010.

91.   R. R. Ramsay, M. R. Popovic-Nikolic, K. Nikolic, E. Uliassi, and M. L. Bolognesi, "A perspective on multi-target drug discovery and design for complex diseases," *Clinical & Translational Med*, vol. 7, no. 1, p. e3, Dec. 2018, doi: 10.1186/s40169-017-0181-2.

92.   E. J. Buenz, R. Verpoorte, and B. A. Bauer, "The Ethnopharmacologic Contribution to Bioprospecting Natural Products," *Annu. Rev. Pharmacol. Toxicol.*, vol. 58, no. 1, pp. 509–530, Jan. 2018, doi: 10.1146/annurev-pharmtox-010617-052703.

93.   F. Mesiti, D. Chavarria, A. Gaspar, S. Alcaro, and F. Borges, "The chemistry toolbox of multitarget-directed ligands for Alzheimer's disease," *European Journal of Medicinal Chemistry*, vol. 181, p. 111572, Nov. 2019, doi: 10.1016/j.ejmech.2019.111572.

94.   H. Geerts *et al.*, "Quantitative Systems Pharmacology for Neuroscience Drug Discovery and Development: Current Status, Opportunities, and Challenges," *CPT Pharmacom & Syst Pharma*, vol. 9, no. 1, pp. 5–20, Jan. 2020, doi: 10.1002/psp4.12478.

95.   D. Namdar, O. Anis, P. Poulin, and H. Koltai, "Chronological Review and Rational and Future Prospects of Cannabis-Based Drug Development," *Molecules*, vol. 25, no. 20, p. 4821, Oct. 2020, doi: 10.3390/molecules25204821.

96.   Y. Song, A. Furtos, D. Fuoco, Y. Boumghar, and G. S. Patience, "Meta-analysis and review of cannabinoids extraction and purification techniques," *Can J Chem Eng*, vol. 101, no. 6, pp. 3108–3131, Jun. 2023, doi: 10.1002/cjce.24786.

97.   M. W. Mullowney *et al.*, "Artificial intelligence for natural product drug discovery," *Nat Rev Drug Discov*, vol. 22, no. 11, pp. 895–916, Nov. 2023, doi: 10.1038/s41573-023-00774-7.