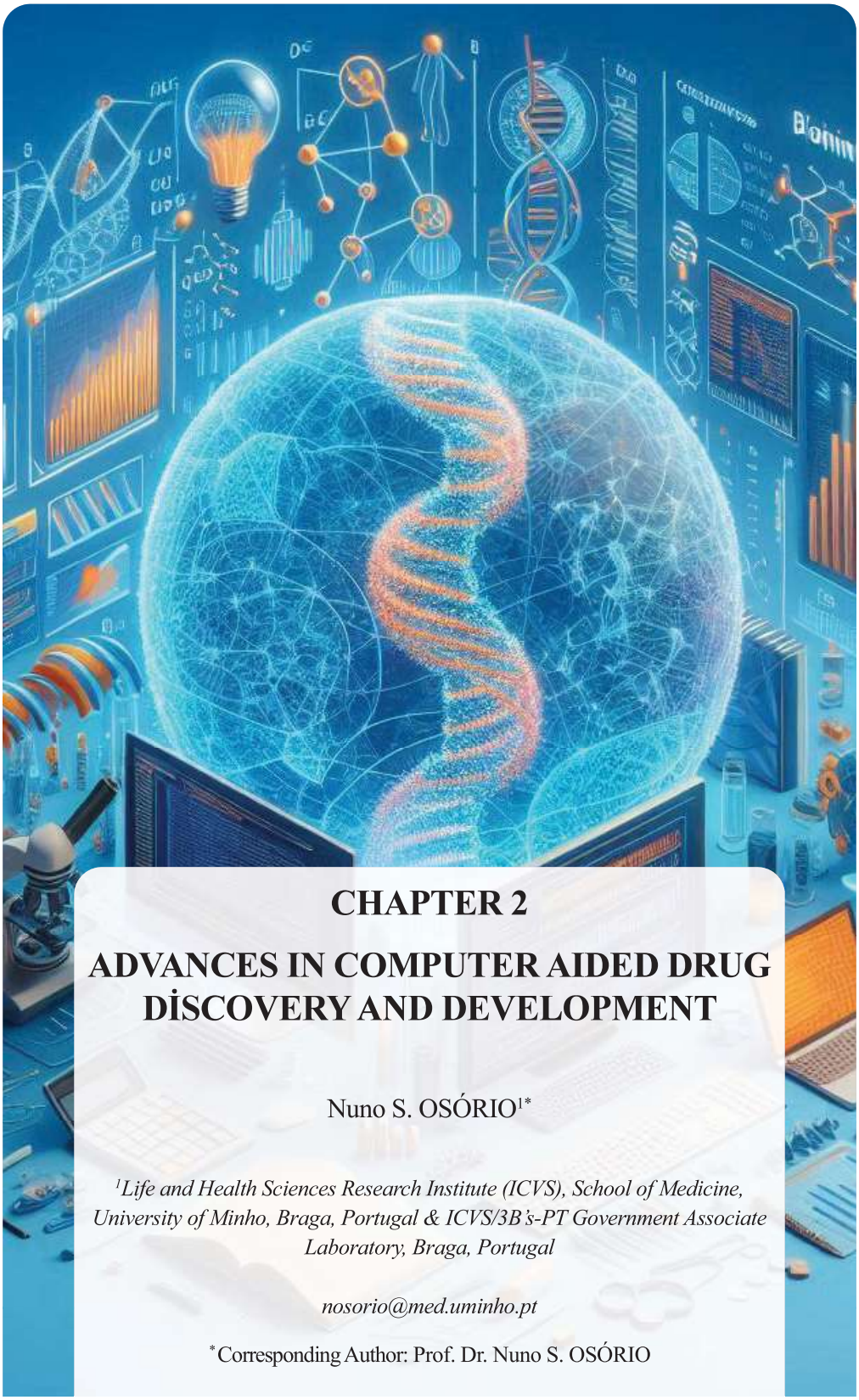


CHAPTER 2

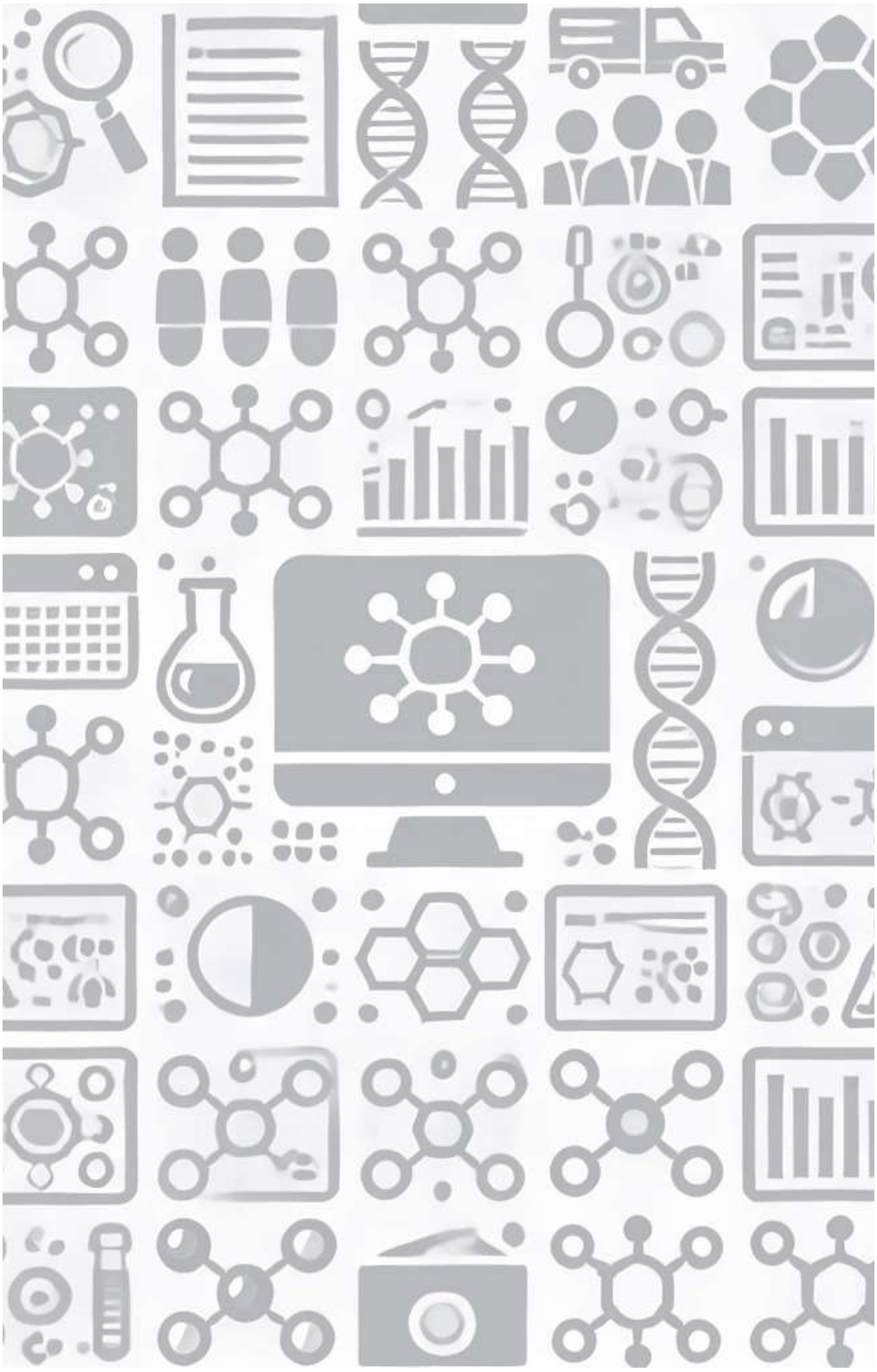
ADVANCES IN COMPUTER AIDED DRUG DİSCOVERY AND DEVELOPMENT

Nuno S. OSÓRIO[1*]

[1]*Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal & ICVS/3B's-PT Government Associate Laboratory, Braga, Portugal*

nosorio@med.uminho.pt

[*]Corresponding Author: Prof. Dr. Nuno S. OSÓRIO



For video Scan the QR code.

The Role of Database in Drug Discovery-Practical Training

1. INTRODUCTION

The development of new drugs is a complex, costly, and risky process that requires the integration of multiple disciplines and technologies. According to some estimates, the average cost of developing a new drug is around $4 billion, and only a small fraction of the drug candidates that enter development end up becoming a commercial product (Kolluri et al., 2022). The main reasons for such high drug attrition include efficacy, safety and toxicology, pharmacology, commercial, and cost of goods (COGs) (Winkler, 2021). Many of these issues are related to the design and molecular characteristics of drug candidates, in addition to manufacturing and delivery strategies utilized. Artificial intelligence (AI) and machine learning (ML) are computational methods and tools that can assist in the discovery or optimization of bioactive compounds using in silico and in vitro surrogate assays (Bleicher et al., 2022). AI and ML can be applied to various aspects of computer-aided drug design (CADD), such as target identification, hit identification, lead optimization, and preclinical evaluation (Patel et al., 2020). AI and ML can use different types of data and representations to generate or evaluate drug candidates, such as molecular structures, properties, activities, interactions, pathways, phenotypes, and clinical outcomes (Patel et al., 2020). AI and ML can also employ different types of algorithms and techniques to learn from data and make predictions, such as supervised learning, unsupervised learning, reinforcement learning, deep learning, and generative models (Patel et al., 2020). AI and ML can help to identify and address potential causes of attrition in preclinical and clinical stages related to product manufacturing, safety, delivery, and efficacy issues (Bleicher et al., 2022). However, AI and ML are not magic bullets that can guarantee success of drug development. AI and ML still face many challenges and limitations. For instance, the accuracy and reliability of computational models and predictions are still a concern (Deng et al., 2021). The availability and quality of experimental data is another challenge. The complexity and diversity of biological systems add another layer of difficulty. Furthermore, there are ethical and legal implications of using AI and ML in drug discovery (Deng et al., 2021). Therefore, AI and ML require a thorough validation and verification of their results using experimental methods and clinical trials. AI/ML techniques are appealing to the pharmaceutical industry due to their automated nature, predictive capabilities, and the consequent expected increase in efficiency (Kolluri et al., 2022). AI/ML approaches have been used in drug discovery over the past 15 to 20 years with increasing sophistication. The most recent aspect of drug development where positive disruption from AI/ML is starting to occur is in clinical trial design, conduct, analysis (Kolluri et al., 2022), as well as in the discovery of drugs for neglected tropical diseases (Winkler, 2021). As we move towards a world where there is a growing integration of AI/ML into R&D, it is critical to get past the related buzz-words. It is equally important to recognize that the scientific method is not obsolete when making inferences about data. Doing so will help in separating hope from hype leading to informed decision-making on the optimal use of AI/ML in drug development (Kolluri et al., 2022).

2. THE PIPELINE(S) TO A NEW DRUG

The process of discovering and developing new drugs is not a linear or standardized one, but rather a complex and dynamic one that involves multiple steps, methods, and disciplines (Deng et al., 2021). Different drugs may follow different paths and strategies, depending on the nature of the disease, the target, the compound, and the available resources. There is no single ”correct” pipeline to get to a new drug, but rather a large diversity of options and choices that can influence the outcome and success of drug development (Figure 1).

One of the main differences among drug pipelines is the starting point of the discovery phase. The discovery phase is the stage where potential drug candidates are identified and selected based on their biological activity, chemical properties, and safety profile. The discovery phase can be initiated by different approaches, such as target-based, phenotypic-based, or serendipity-based. Target-based drug discovery is an approach that starts with the identification and validation of a specific molecular target that is involved in the pathophysiology of a disease. A target can be any biomolecule in an enzyme, a receptor, a channel, that modulates a biological process or pathway. Target-based drug discovery aims to find compounds that can interact with the target and modulate its function in a desired way. This approach relies on the availability of structural and functional information about the target, as well as high-throughput screening techniques that can test large libraries of compounds against the target (Moffat et al., 2017). Phenotypic-based drug discovery is an approach that starts with the observation of a biological phenotype or effect that is relevant to a disease. A phenotype can be a cellular, tissue, organ, or organismal response that reflects a pathological condition or mechanism. Phenotypic-based drug discovery aims to find compounds that can induce or reverse the phenotype of interest, without prior knowledge of the molecular target or mechanism. This approach relies on the availability of robust and reliable assays that can measure the phenotype in vitro or *in vivo* (Moffat et al., 2017). Serendipity-based drug discovery is an unstructured approach that starts with an accidental or unexpected finding that leads to the discovery of a new drug. Serendipity-based drug discovery does not follow a rational or systematic plan but rather exploits chance events or observations that reveal novel biological activities or effects of compounds. This approach relies on the curiosity and creativity of researchers who can recognize and pursue serendipitous discoveries (Ban, 2006). All these approaches have advantages and disadvantages, and none of them can guarantee the success of drug development (Moffat et al., 2017; Ban, 2006). In fact, many drugs have been discovered by combining or switching between different approaches at different stages of development (Table 1). For example, some drugs have been discovered by phenotypic screening and then their targets have been identified later by reverse pharmacology (such as metformin and sildenafil). Some drugs have been discovered by target screening and then their phenotypes have been explored later by forward pharmacology (such as imatinib and rapamycin). Some drugs have been discovered by serendipity and then their targets have been elucidated later (such as penicillin and aspirin).

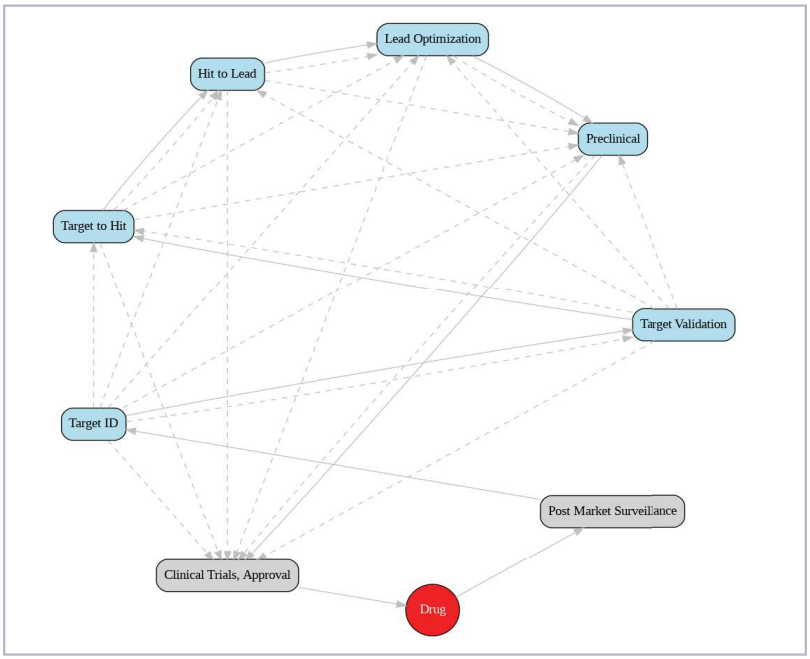


Figure 1. Drug discovery and development pipelines. The typical stages are: target ID, where a biological target with therapeutic potential is identified; target validation, where the target's relevance and druggability are confirmed; target-to-hit, where a molecule (the "hit") that affects the target is discovered; hit-to-Lead, where the "hit" is optimized into a "lead" compound with better properties; lead optimization, where the lead compound is further refined; preclinical, where the drug candidate is tested in non-human models for safety and efficacy; and clinical trials, Approval, where the drug is tested in humans and approved by regulators if safe and effective. The continuous line shows a common path from discovery to market, but different drugs may follow different paths and strategies depending on various factors. There is no single "correct" pipeline, but rather a large diversity of options and choices. The red circle is the approved Drug and the other grey box is Post Market Surveillance, which monitors the drug after it is marketed. The dashed lines show possible alternative paths that might be taken in the development process, highlighting that drug development is not always linear and can involve revisiting or skipping stages based on new information or strategic considerations. The Figure 1 was created using the Digraph module from the graphviz library in Python. ©2023

Table 1. A summary of successful drugs discovered by combining or switching between different approaches.

Drug	Description
Metformin	A drug that lowers blood glucose by inhibit- ing complex I of the mitochondrial respira- tory chain. Its target was identified later by reverse pharmacology (Owen, Doran, and Halestrap, 2000).
Sildenafil (Viagra)	A drug that treats erectile dysfunction by inhibiting phosphodiesterase type 5 (PDE5). It was originally developed for cardiovascular problems and discovered by phenotypic screening (Boolell et al., 1996).
Imatinib (Gleevec)	A drug that treats chronic myeloid leukemia by inhibiting the BCR-ABL tyrosine kinase. It was designed by target screening and its phe- notypes were explored later by forward phar- macology (Druker, 2004).
Rapamisin	A drug that inhibits the mTOR pathway and has various clinical applications, such as im- munosuppression and anti-aging. It was dis- covered in a soil sample and found to have antifungal properties (Ve ̄zina, Kudelski, and Sehgal, 1975; Seto, 2012).
Penisilin	An antibiotic that kills bacteria by interfering with their cell wall synthesis. It was discov- ered by serendipity when Alexander Fleming noticed that a mold contaminated his bacterial cultures (Fleming, 1955).
Aspirin	A drug that relieves pain and fever by inhibit- ing cyclooxygenase (COX) enzymes. The ac- tive ingredient, salicylic acid, was known for its medicinal properties, and aspirin was syn- thesized later by adding an acetyl group (Mc- Kee, Sane, and Deliargyris, 2002).

In recent years, artificial intelligence (AI) and machine learning (ML) have become increasingly important tools in drug discovery.

Table 2. A summary of challenges in drug discovery.

Stage	Method	Challenge
Target Identification	AAI/ML, omics data, literature	Complexity and diversity of biological systems
Target Validation	AI/ML, omics data, literature	HValidation of target rele- vance and druggability
Target-to-Hit	AI/ML, high-throughput screening, virtual screening	Exploration and navigation of vast chemical space
Hit-to-Lead	AI/ML, structure-based or ligand-based drug design	ptimization of hit potency, selectivity, and ADME properties
İlaead Optimization	AI/ML, structure-based or ligand-based drug design	Optimization of lead efficacy, safety, and developability
Preclinical	AI/ML, in vitro and *in vivo* assays	Prediction and evaluation of pharmacokinetics, pharmacodynamics, and toxicity

AI/ML techniques are appealing to the pharmaceutical industry due to their automated nature, predictive capabilities, and consequent expected increase in efficiency (Kolluri et al., 2022). They are being used in various stages of drug development including target identification, hit identification, lead optimization, preclinical evaluation, clinical trial design, conduct analysis (Kolluri et al., 2022), as well as in discovering drugs for neglected tropical diseases (FDA, 2023). The development phase is the stage where selected drug candidates are tested for their efficacy, safety, pharmacokinetics, pharmacodynamics, and manufacturability in preclinical and clinical trials (Scannell et al., 2012). The development phase can be subdivided into different phases according to the regulatory requirements and milestones that need to be achieved before marketing approval. Preceding the development phase, a typical drug discovery pipeline can be divided into five phases: target identification, target validation, target-to-hit, hit-to-lead and lead optimization (Scannell et al., 2012). However, this division between drug discovery and development and its composing steps of the drug discovery and development pipeline is not fixed or universal, and in different drugs may have different duration, costs, risks, and outcomes. The diversity of drug pipelines reflects the complexity and uncertainty of drug discovery and development (Table 2). There is no simple formula or recipe to produce new drugs, but rather a variety of scientific, technological, managerial, regulatory, ethical, and economic factors that influence the process and outcome of drug development (Scannell et al., 2012). Therefore, it is important to understand the strengths and limitations of each approach, method, and phase, and to adopt a flexible and adaptive strategy that can optimize the efficiency and effectiveness of drug development (Scannell et al., 2012). The entire process from discovery to approval can take 10-15 years on average (Brown et al., 2021). Despite advances in understanding biological systems and the development of cutting-edge technologies, the process is still long, costly with a high attrition rate (Brown et al., 2021).

3. AI ENHANCED TARGET IDENTIFICATION

Target identification is the initial step of most modern drug discovery processes, in which a specific biological target is selected for which a drug can be designed. A target can be any biomolecule that interacts with and modulates a biological process or pathway that is involved in the pathophysiology of a disease. Finding the best possible target for a disease, disease subset, or subset of patients is nowadays considered critical to maximize the success of drug development as it can reduce the risk of failure and increase the efficacy and safety of the drug candidates

(Kim et al., 2020). However, target identification is not an easy task as it requires the integration and analysis of large and diverse datasets of multimodal data such as literature (research articles, disease classification documents, symptoms descriptions, regulations, patents), experimental data (omics data, pathways data, systems biology data, compound properties) (Kim et al., 2020). These datasets are often scattered across different sources and formats and contain various types of noise and bias. Therefore, it is challenging to extract relevant and reliable information from these datasets and to identify novel targets (Kim et al., 2020). Artificial Intelligence (AI) techniques are appealing to pharmaceutical industry due to their automated nature, predictive capabilities, and consequent expected increase in efficiency (Savage, 2021). They are being used in various stages of drug development including target identification. AI can computationally sort through and compare various properties of millions of potential small molecules looking for 10 or 20 to synthesize test and optimize in lab experiments before selecting the eventual drug candidate for clinical trials (Savage, 2021). Identifying druggable targets that can address unmet medical needs is a significant challenge in drug discovery (Rasul et al., 2022). The complexity of biological systems, the diversity of diseases, and the scarcity of validated targets make it difficult to find effective and safe drug candidates (Rasul et al., 2022). To overcome these challenges, two emerging technologies show great potential for target identification: knowledge graphs (KGs) and large language models (LLMs) (S. Pan et al., 2024) (J. Z. Pan et al., 2023). By combining KGs and LLMs together, it is possible to leverage the complementary strengths of both technologies and to achieve a synergistic effect for target identification. KGs can provide structured knowledge and logical reasoning for LLMs, while LLMs can provide natural language understanding and generation for KGs. Together, they can enable a rational and integrative approach for target identification that is not biased by silos and more likely to have clinical success (S. Pan et al., 2024) (J. Z. Pan et al., 2023).

3.1. Knowledge Graphs

Knowledge graphs (KGs) are graphical representations of structured information, embodying objects and their relationships through nodes and edges (S. Pan et al., 2024). KGs can integrate multimodal data from various sources and domains into a unified and coherent structure that can capture rich factual knowledge. KGs can enhance target identification by providing external knowledge for inference and interpretability. For example, KGs can help to identify potential targets by linking them to diseases, pathways, phenotypes, compounds and by performing reasoning and querying over the graph structure.

3.2. Large Language Models

Large language models (LLMs) are neural network models that are trained on massive amounts of text data to learn the statistical patterns and semantic representations of natural language (S. Pan et al., 2024). LLMs can generate or evaluate natural language texts based on different types of inputs or tasks. LLMs can enhance target identification by providing natural language processing capabilities for text analysis and generation (S. Pan et al., 2024). For example, LLMs can help to identify potential targets by extracting relevant information from literature, patents and by generating natural language summaries or descriptions of targets (J. Z. Pan et al., 2023).

4. THE ROLE OF DATABASES IN TARGET-TO-HIT AND HIT- TO-LEAD STEPS

The target-to-hit and hit-to-lead stages are two important steps in the drug discovery process, in which potential drug candidates are identified and optimized based on their interaction with a specific biological target. These stages involve extensive use of databases that contain information about chemical compounds, biological targets, experimental results (Kiriiri, Njogu, and Mwangi, 2020). These databases play a crucial role in facilitating the search for potential hits, the evaluation of their properties and activities, the selection of promising leads, and the optimization of their structures and formulations. For instance, databases such as PubChem (Biotechnology Information, 2023), ChEMBL (Mendez et al., 2018), DrugBank (Wishart et al., 2017) provide vast amounts of data about chemical compounds including their structures, properties, activities, toxicity profiles. Similarly databases like UniProt (Consortium, 2022) or RCSB PDB (Berman et al., 2000) provide comprehensive information about protein sequences and structures which is crucial in target identification process. Therefore, databases are an integral part of fueling the modern computer-enhanced drug discovery process. (Kim et al., 2020).

Definition 1: Target
A target is a biomolecule that is involved in the pathophysiology of a disease, such as a DNA, RNA, protein or a lipid.
Definition 2: Hit
A hit is a compound that shows some biological activity against the target, usually measured by an in vitro assay.
Definition 3: Lead
A lead is a compound that has improved biological activity, selectivity, and safety profile compared to the hit, usually measured by an in *in vivo* assay.

The target-to-hit and hit-to-lead stages are challenging because they require the exploration and navigation of the vast chemical space, which is the set of all possible small molecules that could be synthesized or isolated from natural sources. The chemical space is estimated to contain between 10^{60} and 10^{200} molecules, depending on the criteria used to define them (Lipinski and Hopkins, 2004). However, only a tiny fraction of this space has been explored so far, as only about 107 molecules have been synthesized or isolated, and only about 10^2 molecules have been tested as drugs (Lipinski and Hopkins, 2004). To efficiently search the chemical space for hits and leads, it is necessary to apply various strategies and methods that can reduce the complexity and increase the diversity of the chemical space. Some of these strategies and methods include:

- using combinatorial chemistry or natural products as sources of diverse compounds;
- using high-throughput screening or virtual screening techniques to test large libraries of compounds against the target;
- using structure-based or ligand-based drug design approaches to design or optimize compounds based on the target or known active ligands;
- using AI or ML tools to generate or evaluate compounds based on data and models (Lipinski and Hopkins, 2004).

Another important aspect of the target-to-hit and hit-to-lead stages is to consider the desirable properties that a drug candidate should have, such as solubility in water and lipids, bioavailability, stability, permeability, metabolism, toxicity. These properties are often referred to as the drug-likeness or developability of a compound, and they can influence the pharmacokinetics (”what the body does to the drug”) and pharmacodynamics (”what the drug does to the body”) of the drug (Lipinski and Hopkins, 2004). One of the most important properties is solubility in water and lipids, as it determines how well a compound can dissolve in aqueous or organic media. Solubility in water is essential for oral administration, as it affects how well a compound can be absorbed from the gastrointestinal tract into the bloodstream. Solubility in lipids is essential for crossing biological membranes, such as the blood-brain barrier or the cell membrane. A compound should have a balance between solubility in water and lipids, as too much or too little of either can compromise its bioavailability (the fraction of the administered dose that reaches the systemic circulation) or efficacy (the ability to produce a desired effect) (Lipinski and Hopkins, 2004). To facilitate the exploration and navigation of the chemical space, several public databases have been developed that store and provide information about millions of compounds and their biological activities. Two of the most popular databases are PubChem and ChEMBL. PubChem provides information about biological activities of small molecules across numerous assays while ChEMBL provides manually curated bioactivity data extracted from scientific literature (Kar and Leszczynski, 2023). These databases are invaluable resources for researchers involved in drug discovery as they provide access to vast amounts of data which can be used for various purposes such as virtual screening, data mining and machine learning (Rifaioglu et al., 2018). In addition to these databases, there are also specialized databases focusing on specific types of data. For example, BindingDB focuses on measured binding affinities, particularly of enzyme inhibitors (Liu et al., 2006), and ZINC is a free tool that focuses on providing commercially available compounds for virtual screening, particularly for drug discovery applications (Irwin et al., 2012). TTD combines detailed drug data with comprehensive drug target information (Zhou et al., 2021). These and several other databases provide more specific information which can be very useful, depending on research needs.

Moreover with advancements in artificial intelligence (AI) technologies there has been an increase in use of AI algorithms for drug discovery. AI can help in various stages from target identification to lead optimization by analyzing large datasets available in these databases (Kim et al., 2020). AI algorithms can predict potential targets for drugs identify potential drugs for known targets predict drug side effects among other things (Kim et al., 2020). Therefore databases play an integral role in modern drug discovery process.

4.1. ChEMBL and PubChem

ChEMBL is a database of bioactive molecules with drug-like properties, curated from scientific literature (Mendez et al., 2018). PubChem is a database of chemical structures and their associated biological activities, derived from various sources such as patents, journals, or web pages (Biotechnology Information, 2023). Both ChEMBL and PubChem can be used to screen for hits and to find leads by providing various tools and features that allow users to search, filter, analyze, and compare compounds. For example, users can search for compounds by name, structure, substructure, similarity, or target; filter compounds by activity, property, source, or annotation;

analyze compounds by clustering, diversity, or enrichment; compare compounds by structure-activity relationship (SAR), pharmacophore, or docking (O'Boyle et al., 2011) (Bolton et al., 2008). One of the most common and convenient ways to represent compounds in ChEMBL, PubChem and other databases is by using SMILES (Simplified Molecular Input Line Entry System). SMILES is a notation system that encodes the structure of a compound as a linear string of symbols. SMILES can represent the atoms, bonds, branches, rings, stereochemistry, and isotopes of a compound in a compact and human-readable way (Weininger, 1988). SMILES can also be easily converted to other formats or representations, such as molecular graphs, fingerprints, or images. This is the SMILES string for aspirin, CC(=O)OC1=CC=CC=C1C(=O)O, it can be broken down as follows: CC represents two carbon atoms bonded together, (=O) represents a carbon atom double-bonded to an oxygen atom, OC1=CC=CC=C1 represents a benzene ring with a carbonyl group attached at the C1 position, and C(=O)O represents a carbonyl group. The SMILES string has the necessary information to generate a Kekule´ diagram, such as the one in Figure 2.

ChEMBL and PubChem are valuable resources for target to hit and hit to lead stages in drug discovery, as they can provide access to a large and diverse subset of the chemical space that contains relevant and reliable information about compounds and their biological activities (Biotechnology Information, 2023) (Mendez et al., 2018). By using ChEMBL and PubChem in combination with other strategies and methods mentioned earlier, it is possible to identify and optimize hits and leads that can more likely advance to the next stages of drug development.

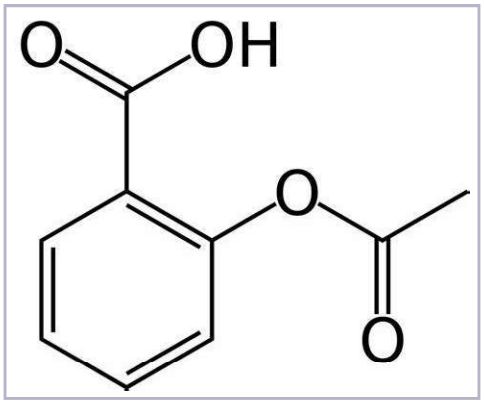


Figure 2. The Kekul' diagram of aspirin, showing the acetyl and salicylic acid groups. Aspirin is a weak acid that can form salts with bases, such as sodium salicylate. Aspirin is also an ester that can undergo hydrolysis to produce acetic and salicylic acids. Aspirin is synthesized by reacting salicylic acid with acetic anhydride in the presence of a catalyst, such as sulfuric acid. The Figure 2 is in the public domain.

5. AI ENHANCED LEAD OPTIMIZATION

Artificial Intelligence (AI) has been applied to various domains and problems, including drug discovery and lead optimization (Deng et al., 2021). AI can enhance lead optimization by providing advanced tools and methods that can handle large and complex datasets of compounds and their activities, and that can generate or evaluate novel compounds with desired properties (Bleicher et al., 2022). Neural networks are one of the most popular and successful types of AI models that have been used for lead optimization. Neural networks are computational models that are inspired by the

structure and function of biological neurons. Neural networks consist of layers of artificial neurons that are connected by weights that represent the strength of the connections. Neural networks can learn from data by adjusting the weights according to a learning algorithm that minimizes a loss function that measures the error between the predicted and actual outputs. Neural networks can approximate any complex function or relationship between inputs and outputs, making them suitable for modeling Structure-Activity Relationship (SAR) and Quantitative Structure-Activity Relationship (QSAR) (Deng et al., 2021). Neural networks have decades of track record in computational chemistry, dating back to the early 1990s when they were first applied to QSAR problems (Deng et al., 2021). Since then, neural networks have been improved and diversified by incorporating different architectures, techniques, and applications. Some of the most common types of neural networks that have been used for lead optimization are:

- **Feed-forward neural networks (FFNNs):** These are the simple stand most basic type of neural networks, in which the information flows from the input layer to the output layer through one or more hidden layers in a forward direction. FFNNs can be used for regression or classification tasks, such as predicting the activity or toxicity of compounds based on their molecular descriptors or fingerprints (Deng et al., 2021).

- **Recurrent neural networks (RNNs):** These are a type of neural networks that have feedback loops or connections that allow information to flow backward from later layers to earlier layers. RNNs can capture temporal or sequential information from data, such as text or speech. RNNs can be used for sequence generation or analysis tasks, such as generating or optimizing SMILES strings of compounds based on their desired properties (Deng et al., 2021).

- **Long short-term memory (LSTM) networks:** These are a special type of RNNs that have memory cells or units that can store or forget information over long periods of time. LSTM networks can overcome the problem of vanishing or exploding gradients that affect RNNs when dealing with long sequences. LSTM networks can be used for sequence generation or analysis tasks, such as generating or optimizing SMILES strings of compounds based on their desired properties (Deng et al., 2021).

- **Transformers:** These are a type of neural networks that use attention mechanisms to focus on different parts of the input or output sequences based on their relevance or importance. Transformers do not use recurrent or convolutional layers, but rather rely on self-attention and cross-attention layers to encode and decode sequences. Transformers can be used for sequence generation or analysis tasks, such as generating or optimizing SMILES strings of compounds based on their desired properties (Deng et al., 2021).

AI enhanced lead optimization is an emerging and promising field that can accelerate and improve the drug discovery process by providing novel and efficient methods for structure-property prediction with simple 2D-based chemical representations of small molecules (Bleicher et al., 2022). Furthermore, AI/ML methods in drug discovery are maturing and their utility and impact is likely to permeate many aspects of drug discovery including lead finding and lead optimization (Bleicher et al., 2022).

5.1. Neural Networks

Drug repurposing, also known as drug reprofiling, is the process of finding new uses for existing drugs. This can be done by exploring new indications for approved drugs, developing new formulations or drug combinations of existing drugs, or optimizing existing drugs by modifying their structure to improve their efficacy, reduce side effects, or alter their pharmacokinetic properties (Jarada, Rokne, and Alhajj, 2020). This process plays an important role in optimizing the pre-clinical process of developing novel drugs by saving time and cost compared to the traditional de novo drug discovery processes (Jarada, Rokne, and Alhajj, 2020). As AI models continue to improve and as more data becomes available, their impact on drug discovery is likely to increase (Akshaya and Deva, 2022). Neural networks, a subset of machine learning, have found extensive applications in the field of medical chemistry. They are used for tasks such as predicting molecular properties, generating novel drug candidates, and modeling complex biological systems (Akshaya and Deva, 2022). Neural networks are computational models inspired by the human brain. They are capable of learning complex patterns and making predictions based on these patterns. In medical chemistry, they are used to model and predict various chemical and biological phenomena (Akshaya and Deva, 2022). One of the key applications of neural networks in medical chemistry is predicting molecular properties. These properties include characteristics like solubility, toxicity, and binding affinity. Accurate prediction of these properties is crucial for drug discovery and development (Akshaya and Deva, 2022). Neural networks can also be used to generate novel drug candidates. This is done by training the network on a large dataset of known drugs and then using it to generate new molecules that are likely to have desirable properties (Akshaya and Deva, 2022). Finally, neural networks can be used to model complex biological systems. This includes aspects like protein folding, metabolic pathways, and cell signaling networks. These models can help scientists understand how these systems work and how they can be manipulated for therapeutic purposes (Akshaya and Deva, 2022). Neural networks have the potential to revolutionize medical chemistry by providing powerful tools for prediction, generation, and modeling. As our understanding of these networks improves and as more data becomes available, their impact on medical chemistry is likely to grow (Akshaya and Deva, 2022).

5.1.1. Transformer-Based Models and Chemoınformatıcs

The vast expanse of chemical space, which encompasses all possible small organic molecules, is estimated to contain up to 1060 unique structures. This is a number so large that it dwarfs the number of atoms in the observable universe (Lipinski and Hopkins, 2004). Navigating this space in search of compounds that can modulate biological systems is one of the grand challenges in biology and medicine (Lipinski and Hopkins, 2004). Chemoinformatics is the field that develops tools and methods to navigate the chemical space. It uses computational techniques to predict the structure, activity, and properties of molecules (Lipinski and Hopkins, 2004). High-throughput screening (HTS) is a method for scientific experimentation especially used in drug discovery. It involves testing a large number of potential drug compounds for activity against biological targets (Lipinski and Hopkins, 2004). The advent of transformer-based models, such as the BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer), has revolutionized many fields, including chemoinformatics (Wang et al., 2022). These models have the potential to significantly accelerate the drug discovery process by predicting drug-target interactions, generating novel drug candidates, and optimizing existing drugs (Wang et al., 2022). One of the key

applications of transformer-based models in drug discovery is predicting drug-target interactions. These models can learn complex patterns in data and predict how a drug will interact with various targets in the body (Wang et al., 2022). Transformer-based models can also be used for de novo drug design. This involves generating novel drug candidates from scratch. The models can learn the chemical language of drug-like compounds and generate new compounds that are likely to have desired properties (Wang et al., 2022).

6. CONCLUSION

In conclusion, this chapter has taken us on a journey through the complex and fascinating world of drug discovery, from the initial pipelines to the preclinical trials. We have explored how computer-based method- ologies play crucial roles in enhancing target identification, optimizing lead compounds, and navigating the vast chemical space (Savage, 2021). In ”The Pipeline(s) to a New Drug”, we delved into the intricate processes involved in drug discovery and development. We highlighted the idea that these processes are not linear but rather a series of interconnected steps that require careful planning and execution (Savage, 2021). The sections ”AI Enhanced Target Identification” and ”The Role of Databases in Target-to-Hit and Hit-to-Lead Steps” demonstrate the importance of ML/AI and databases in modern drug discovery. Compu- tational algorithms can sift through vast amounts of data to identify potential drug targets, while databases provide a wealth of information that can guide the hit-to-lead optimization process (Savage, 2021). In ”AI Enhanced Lead Optimization”, we explored how AI technologies can generate compounds designed specifically to successfully enter clinical trials (Savage, 2021). Indeed, AI has been revolutionizing the field of drug discovery by streamlining the lead optimization process. By predicting the properties of potential drug candidates, AI can assist researchers in designing drugs that are not only more effective but also safer (Savage, 2021). This is achieved through the use of machine learning algorithms that can analyze vast amounts of data and identify patterns that unassisted humans would overlook. Understanding the chemical space is crucial in the drug discovery field (Lipinski and Hopkins, 2004). This vast space, which contains all possible small organic molecules, is key to finding compounds that can modulate biological systems. However, navigating this space is a complex task due to its size and complexity. Advanced AI models, such as those mentioned in this chapter, have shown great promise in drug discovery. These models leverage neural networks and transformer-based architectures to predict drug-targets, generate novel drug candidates, and model interactions in complex biological systems. As we move forward, the materialized chemical space will continue to expand with the continuous synthesis of new compounds. This expansion will pro- vide more opportunities for finding novel drugs but will also pose challenges in terms of navigation and analysis. Advanced computational algorithms will be crucial in efficiently exploring this expanding space. As these models become more sophisticated, their predictions will become more accurate and reliable. This enhancement will be seen in all stages of the pipeline to a new drug, from target identification to lead optimization and drug development (Deng et al., 2021). Transformer-based models and their subsequent developments are likely to lead to future advancements. On the wet lab, the integration of multi-omics data (Genomics, proteomics, metabolomics) at a single-cell scale will provide a more holistic view of biological systems. This improved understanding of disease mechanisms will aid in the identification of novel drug targets, enabling the development of more effective and targeted therapies. With advancements in computational algorithms and an increased understanding of the human individual’s genetic makeup, its interactions with the environment

and other biomes, we are moving towards truly personalized medicine (Figure 3). This means that drugs could be tailored to individuals based on their unique biological system, genetic and biochemical makeup, improving efficacy and reducing side effects. Drug discovery is a multi-disciplinary field that encompasses biology, chemistry, informatics, and several other sciences. In the future, we can expect even greater collaboration between these disciplines, leading to more integrated and efficient drug discovery pipelines (Kim et al., 2020). In conclusion, the future of drug discovery looks promising with the integration of emerging computational technologies. However, it also poses new challenges that need to be addressed through continuous research and innovation. The journey towards finding new drugs is complex and arduous, but with these advancements, we are moving closer to our goal everyday.

Figure 3. A futuristic image illustrating the concept of personalized medicine, with elements representing computational algorithms, genetic makeup, and interactions with the environment and biomes. This image was generated by an artificial intelligence model (DALL-E) from a text prompt.

7. GLOSSARY

ADME: is an acronym in pharmacokinetics and pharmacology for Absorption, Distribution, Metabolism, and Excretion. These are four criteria which largely influence the performance and pharmacological activity of the compound as a drug.

Artificial Intelligence (AI): The branch of computer science that deals with creating machines or systems that can perform tasks that normally require human intelligence, such as reasoning, learning, and decision making.

Attention mechanism: A technique that allows a neural network to focus on different parts of the input or output sequences based on their relevance or importance.

Biomolecule: A molecule that is involved in the structure or function of living organisms, such as proteins, nucleic acids, lipids, or carbohydrates.

Chemical space: The set of all possible small organic molecules that can be synthesized or isolated from natural sources.

Clinical trial: A type of research study that tests the safety and efficacy of a new drug or intervention in human volunteers, usually following a predefined protocol and regulations.

Computer-aided drug design (CADD): The use of computational methods and tools to assist in the dis- covery and optimization of new drugs, such as by predicting molecular properties, interactions, and activities.

Database: A collection of organized and structured data that can be accessed, manipulated, and updated by computer programs.

Drug attrition: The failure of a drug candidate to progress through the development stages due to various reasons, such as lack of efficacy, safety issues, or commercial viability.

Drug discovery: The process of finding and developing new drugs that can treat or prevent diseases.

Drug-likeness The property of a compound that indicates how well it can become a successful drug, based on factors such as solubility, bioavailability, stability, etc.

Generative model: A type of machine learning model that can generate new data or samples based on existing data, such as by creating novel drug candidates that match certain criteria.

Hit-to-lead: The stage of drug discovery in which potential drug candidates are optimized based on their biological activity, selectivity, and safety profile.

Knowledge graph (KG): A graph-based representation of knowledge that consists of entities, relations, and attributes.

Large language model (LLM): A type of deep neural network that can process natural language and gen- erate text based on large amounts of data.

Lead compound: A chemical compound that has some desirable biological activity and can be further optimized to become a drug candidate.

Lead optimization: The stage of drug discovery in which potential drug candidates are improved based on their biological activity, selectivity, safety, and drug-likeness.

Multi-omics data: Data that integrates information from different levels of biological organization, such as genes, proteins, metabolites, etc.

Neglected tropical diseases (NTDs): Agroupofinfectiousdiseasesthataffectmorethanonebillionpeople in low- and middle-income countries, often causing chronic disability and poverty. They are neglected because they receive less attention and funding than other diseases.

Neuralnetwork: A type of machine learning model that consists of layer sofartificial neurons that canlearn from data and perform complex tasks.

Omics data: Data that captures information from different levels of biological organization, such as ge- nomics, transcriptomics, proteomics, metabolomics, etc.

Personalized medicine: An approach to medicine that takes into account the individual characteristics of each patient, such as their genetic makeup, lifestyle, and environment.

Quantitative Structure-Activity Relationship (QSAR): The mathematical expression of the SAR using molecular descriptors or features as variables and biological activity as the dependent variable.

SMILES: A notation system that represents the structure of a molecule using a string of symbols.

Structure-Activity Relationship (SAR): The relationship between the chemical structure and the biological activity of a compound or a series of compounds.

Target identification: The initial step of most modern drug discovery processes, in which a specific biolog- ical target is selected for which a drug can be designed.

Target-to-hit: The stage of drug discovery in which potential drug candidates are identified based on their interaction with a specific biological target.

Transformer-based model: A type of deep neural network that uses attention mechanisms to learn the relationships between different parts of an input sequence, such as words or molecules.

ACKNOWLEDGES

This work was supported by the project “Strategic Necessity: Molecule to Drug” with the grant number 2022-1-TR01-KA220-VET-000088373.

REFERENCES

Ban, T.A. The role of serendipity in drug discovery. Dialogues Clin Neurosci. 8, 335-344. (2006)

Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T.G., Fan, J., Garmiri, P., da Costa Gonzales, L.J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D., Kandasaamy, S., Lock, A., Luciani, A., Lugaric, M., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Mishra, A., Moulang, K., Nightingale, A., Pundir, S., Qi, G., Raj, S., Raposo, P., Rice, D.L., Saidi, R., Santos, R., Speretta, E., Stephenson, J., Totoo, P., Turner, E., Tyagi, N., Vasudev, P., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A.J., Aimo, L., Argoud-Puy, G., Auchincloss, A.H., Axelsen, K.B., Bansal, P., Baratin, D., Batista Neto, T.M., Blatter, M.C., Bolleman, J.T., Boutet, E., Breuza, L., Gil, B.C., Casals-Casas, C., Echioukh, K.C., Coudert, E., Cuche, B., de Castro, E., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gaudet, P., Gehant, S., Gerritsen, V., Gos, A., Gruaz, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Kerhornou, A., Le Mercier, P., Lieberherr, D., Masson, P., Morgat, A., Muthukrishnan, V., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Poux, S., Pozzato, M., Pruess, M., Redaschi, N., Rivoire, C., Sigrist, C.J.A., Sonesson, K., Sundaram, S., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Zhang, J. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 51, 523-531. (2023)

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. The Protein Data Bank. Nucleic Acids Res. 28, 235–242. (2000)

Bleicher, L.S., van Daelen, T., Honeycutt, J.D., Hassan, M., Chandrasekhar, J., Shirley, W., Tsui, V., Schmitz, U. Enhanced utility of AI/ML methods during lead optimization by inclusion of 3D ligand information. Front Drug Discov. 2, 1-12. (2022)

Bolton, E.E., Wang, Y., Thiessen, P.A., Bryant, S.H. PubChem: Integrated Platform of Small Molecules and Biological Activities. Annu Rep Comput Chem. 4, 217-241. (2008)

Boolell, M., Allen, M.J., Ballard, S.A., Gepi-Attee, S., Muirhead, G.J., Naylor, A.M., Osterloh, I.H., Gingell, C. Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. Int J Impot Res. 8, 47-52. (1996)

Brown, D.G., Wobst, H.J., Kapoor, A., Kenna, L.A., Southall, N. Clinical development times for innovative drugs, Nat Rev Drug Discov. 21, 793-794. (2022)

Deng, J., Yang, Z., Ojima, I., Samaras, D., Wang, F. Artificial intelligence in drug discovery: Applications and techniques, Brief Bioinform. 23, 1-19. (2022)

Druker, B.J. Imatinib as a paradigm of targeted therapies. Adv Cancer Res. 91, 1-30. (2004)

FDA. Artificial Intelligence and Machine Learning (AI/ML) for Drug Development. U.S.FDA. 167973, 1-31. (2023)

Fleming, A. THE DISCOVERY of penicillin. Br Med J. 4915, 711. (1955)

Irwin, J.J., Sterling, T., Mysinger, M.M., Bolstad, E.S., Coleman, R.G. ZINC: A free tool to discover chemistry for biology, J. Chem. Inf. Model. 52, 1757–1768. (2012)

Jarada, T.N., Rokne, J.G., Alhajj, R. A review of computational drug repositioning: Strategies, approaches, opportunities, challenges, and directions. J Cheminform. 12, 1-23. (2020)

Kar, S., Leszczynski, J. Databases for Drug Discovery and Development. Trends Comput. Model. Drug Discov. 35, 269-298. (2023)

Karthikeyan, A., Priyakumar, U.D. Artificial intelligence: machine learning for chemical sciences. J. Chem. Sci. 134, 1-20. (2022)

Kim, H., Kim, E., Lee, I., Bae, B., Park, M., Nam, H. Artificial Intelligence in Drug Discovery: A Comprehensive Review of Data-driven and Machine Learning Approaches. Biotechnol Bioprocess Eng. 25, 895-930 (2020).

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E. PubChem 2023 update. Nucleic Acids Res. 51, D1373–D1380. (2023)

Kiriiri, G.K., Njogu, P.M., Mwangi, A.N. Exploring different approaches to improve the success of drug discovery and development projects: a review. Futur J Pharm Sci. 6, 1-12. (2020)

Kolluri, S., Lin, J., Liu, R., Zhang, Y., Zhang, W. Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: a Review. AAPS J. 24, 1-10. (2022)

Lipinski, C., Hopkins, A. Navigating chemical space for biology and medicine. Nature. 432, 855–861. (2004)

Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K. BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 35, D198–D201. (2007)

McKee, S.A., Sane, D.C., Deliargyris, E.N. Aspirin resistance in cardiovascular disease: A review of prevalence, mechanisms, and clinical significance. Thromb Haemost. 88, 711-715. (2002)

Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C.J., Segura-Cabrera, A., Hersey, A., Leach, A.R. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 47, D930-D940. (2019)

Moffat, J.G., Vincent, F., Lee, J.A., Eder, J., Prunotto, M. Opportunities and challenges in phenotypic drug discovery: An industry perspective. Nat Rev Drug Discov. 16, 531–543. (2017)

O'Boyle, N.M., Guha, R., Willighagen, E.L., Adams, S.E., Alvarsson, J., Bradley, J.C., Filippov, I. V., Hanson, R.M., Hanwell, M.D., Hutchison, G.R., James, C.A., Jeliazkova, N., Lang, A.S.I.D., Langner, K.M., Lonie, D.C., Lowe, D.M., Pansanel, J., Pavlov, D., Spjuth, O., Steinbeck, C., Tenderholt, A.L., Theisen, K.J., Murray-Rust, P. Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on. J Cheminform. 3, 1-16. (2011)

Owen, M.R., Doran, E., Halestrap, A.P. Evidence that metformin exerts its anti-diabetic effects through inhibition of complex 1 of the mitochondrial respiratory chain. Biochem J. 348, 607–614. (2000)

Pan, J.Z., Razniewski, S., Kalo, J.-C., Singhania, S., Chen, J., Dietze, S., Jabeen, H., Omeliyanenko, J., Zhang, W., Lissandrini, M., Biswas, R., De Melo, G., Bonifati, A., Vakaj, E., Dragoni, M., Kessler, B., Graux, D., Hogan, A., Horrocks, I., Hotho, A., Kagal, L. Large Language Models and Knowledge Graphs: Opportunities and Challenges. TGDK. 1, 1-38. (2023)

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. TKDE. 36, 3580-3599. (2024)

Patel, L., Shukla, T., Huang, X., Ussery, D.W., Wang, S. Machine Learning Methods in Drug Discovery. Molecules. 25, 1-17. (2020)

Rasul, A., Riaz, A., Sarfraz, I., Khan, S.G., Hussain, G., Zara, R., Sadiqa, A., Bushra, G., Riaz, S., Iqbal, M.J., Hassan, M., Khorsandi, K. Target identification approaches in drug discovery. Drug Target Select Valid. 5, 41-59. (2022)

Rifaioglu, A.S., Atas, H., Martin, M.J., Cetin-Atalay, R., Atalay, V., Doğan, T. Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases. Brief Bioinform. 27, 1878-1912. (2019)

Savage, N. Tapping into the drug discovery potential of AI. Biopharma Dealmakers. 6, B37-B39. (2021)

Scannell, J.W., Blanckley, A., Boldon, H., Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 11, 191–200. (2012)

Seto, B. Rapamycin and mTOR: a serendipitous discovery and implications for breast cancer. Clin Trans Med. 1, 29. (2012)

Vézina, C., Kudelski, A. Rapamycin (AY-22,989), a new antifungal antibiotic. I. taxonomy of the producing streptomycete and isolation of the active principle. J Antibiot (Tokyo). 28, 721-726. (1975)

Wang, Y., Zhao, H., Sciabola, S., Wang, W. cMolGPT: A Conditional Generative Pre-Trained Transformer for Target-Specific De Novo Molecular Generation. Molecules. 28, 1-14. (2023)

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

Winkler, D.A. Use of Artificial Intelligence and Machine Learning for Discovery of Drugs for Neglected Tropical Diseases. Front Chem. 15, 1-15. (2021)

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., MacIejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, Di., Pon, A., Knox, C., Wilson, M. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. 46, D1074–D1082. (2018)

Zhou, Y., Zhang, Y., Lian, X., Li, F., Wang, C., Zhu, F., Qiu, Y., Chen, Y. Therapeutic target database update 2022: Facilitating drug discovery with enriched comparative data of targeted agents. Nucleic Acids Res. 50, D1398-D1407. (2022)