A Zifo White paper

ARTIFICIAL INTELLIGENCE IN THE PHARMACEUTICAL INDUSTRY PART 1 - Drug Discovery & Preclinical Development

Zifo
SCIENTIFIC INFORMATICS

Introduction

This is the first of a series of white papers that looks at the impact of artificial intelligence (AI) on the Pharmaceutical Industry. Each paper will look at specific areas and provide background and examples of where AI is and could have impacts on the industry. The first part looks at drug research and preclinical development.

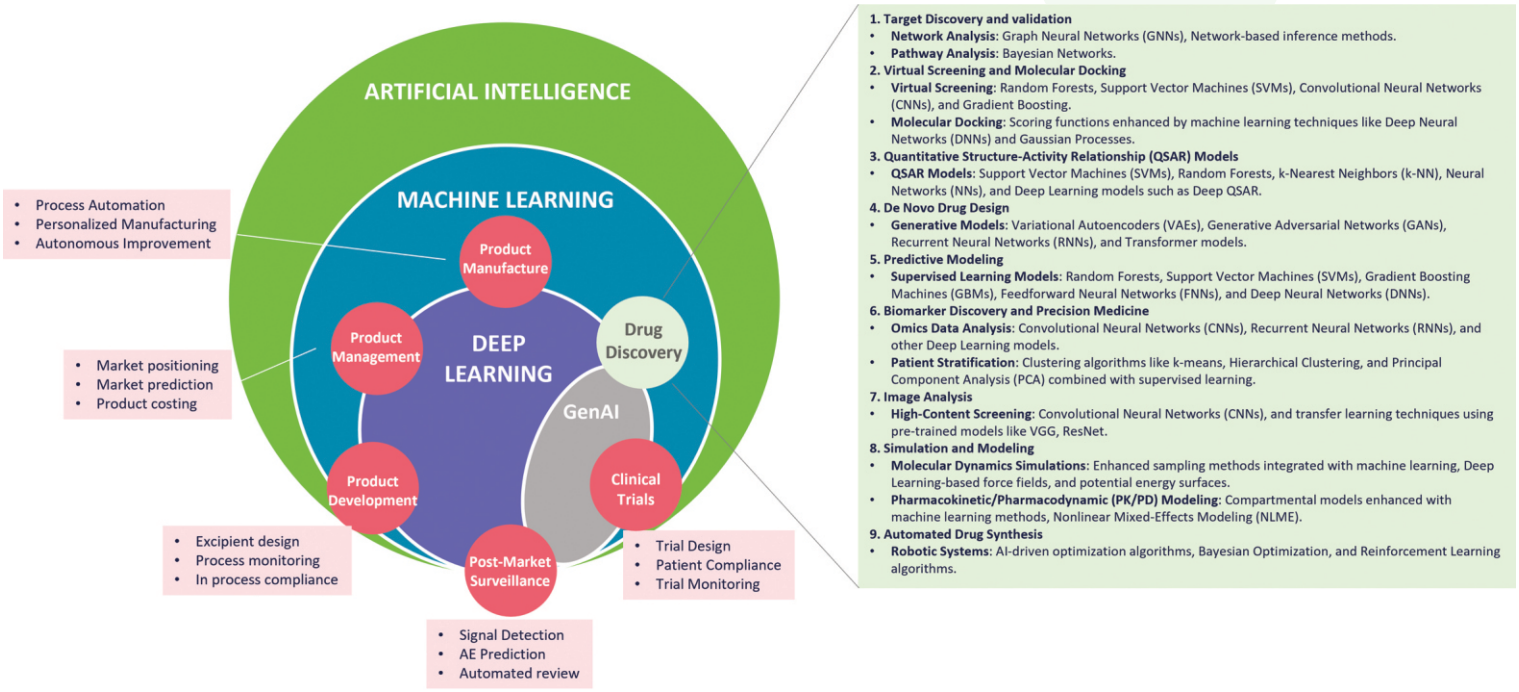
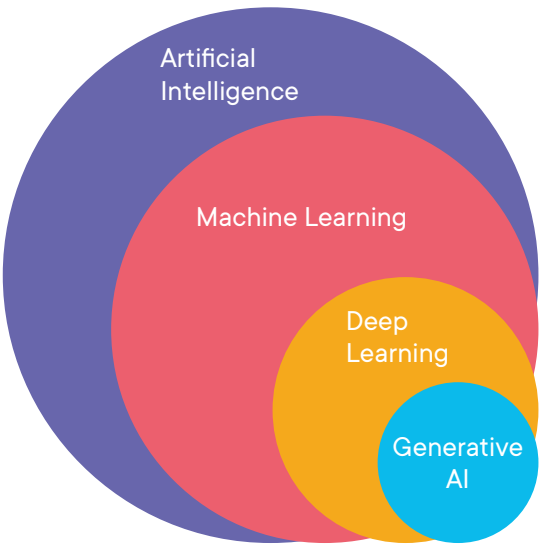


Figure 1. The various areas that this white paper series will look at – starting with research and preclinical development, then moving onto clinical trials, process development, product management, post market surveillance and final manufacturing.

The utility of AI within drug discovery and pre-clinical development heralds a new era of innovation in biomedical research. In the complex journey from the lab bench to the patient's bedside, the preclinical phase stands as a pivotal crossroad where potential drug candidates are identified, validated and rigorously analyzed before potential clinical testing in patients.

Traditional drug discovery and preclinical development is fraught with high costs, prolonged timelines, and a considerable risk of failure. It is, however, a phase in which vast volumes of data are produced and analyzed. Enter modern AI: a transformative technology with the promise of revolutionizing the success of the entire process. By harnessing vast and diverse datasets, advanced algorithms, and unprecedented computational power, AI is poised to accelerate the output and success of drug discovery and preclinical development.

Whether it's predicting drug targets, optimizing the complementarity between molecular structures, or anticipating potential side-effects, AI promises to make drug discovery more efficient, cost-effective, and successful for patients. In this whitepaper, we delve deeper into this intersection of machine intelligence and molecular biology, to reveal how AI is enabling a transformative shift in the conception and discovery of new therapeutic modalities.



1956

Artificial Intelligence

The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence

1997

Machine Learning

Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions

2017

Deep Learning

A machine learning technique in which layers of neural networks are used to process data and make decisions

2021

Generative AI

Create new written, visual, and auditory content given prompts or existing data

Figure 2: Showing the history of Artificial Intelligence (AI), Machine Learning (ML), Deep Learning, and Generative AI represent different yet interconnected concepts within the field of computer science. AI is the broadest term, encompassing any technique that enables computers to mimic human intelligence using logic, if-then rules, decision trees, and machine learning. Machine Learning, a subset of AI, involves the development of algorithms that allow computers to learn and make predictions or decisions based on data rather than being explicitly programmed for a specific task. Deep Learning, a subset of ML, involves neural networks with many layers (hence "deep") that learn progressively higher-level features from data, enabling complex tasks like image and speech recognition. Generative AI, built upon deep learning techniques, focuses on creating new content – text, images, or music – that resembles human-generated content, often using advanced methods like Generative Adversarial Networks (GANs) or transformers. While AI and ML provide the foundational concepts and methods, Deep Learning and Generative AI represent more advanced and specialized applications within this broader framework.

AI-Driven Drug Discovery and Preclinical Development

Before exploring the utilities of AI within Pharma, it is helpful to quickly review the prevalent drug R&D paradigm. Developing a new drug typically begins with identifying a target, where researchers search for genes/proteins that are either directly or indirectly associated with a disease of interest. After selection of the most promising candidate, it must first be validated by demonstrating a potential therapeutic effect through modulation. Once thoroughly vetted, the search for a suitable therapeutic modality (e.g., small molecule, antibody, cell or gene therapy (C&GT) etc.) can begin.

In the case of small molecule drugs, screening assays, typically measuring binding affinity and/or protein activity, are developed and performed in high throughput automation environments. The "screening" aims to identify one or more lead compounds that can modulate the activity of the identified target in a therapeutically desirable manner. The screening phase is often assisted through structure-based drug design (SBDD) and virtual screening approaches. Lead compounds then undergo preclinical evaluation in model systems (e.g., in vivo, in vitro and in silico) to assess their potential pharmacological and toxicological behaviors. Data emerging from these efforts guide the lead optimization process through medicinal or computational chemistry. Upon completion, a single drug candidate is usually nominated, patents are filed, and an Investigational New Drug Application (IND) is submitted to regulators in preparation for three phases of clinical testing in human subjects.

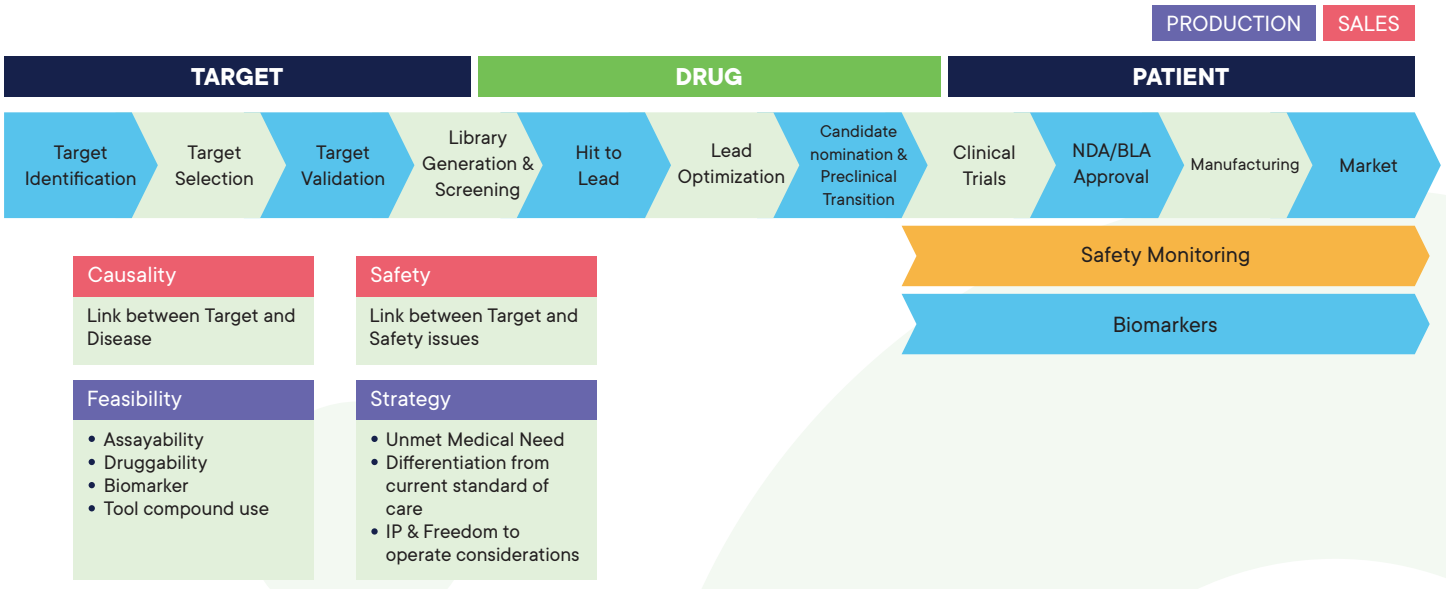


Figure 3: The overall drug discovery starts with identifying a drug target. After validation of the target's disease association and "drugability", screening typically takes place to identify potential "hits". As the process continues, fewer candidates are worked on. The next step is where the drug is evaluated for pharmacological and toxicological profiles with in vitro and in vivo approaches. Once an optimized lead is chosen, patents are applied for, and an Investigational New Drug Application (IND) is submitted. If approved, the sponsor advances to the first of the human clinical testing phases.

For the purpose of this whitepaper, we focus on the utility of AI in the stages prior to clinical trials. These application areas are best categorized into three core activities.

1. **Target Identification and Validation:** Finding and confirming biological molecules implicated in a disease.
2. **Hit-to-Lead Identification:** Testing millions of compounds to find potential 'hits' that can modulate a target and then optimizing for efficacy and safety to generate the 'lead' drug candidate(s).
3. **Preclinical Development:** Testing the lead compounds in model systems to better understand safety and efficacy. While several leads may be tested, typically one emerges as the drug candidate for clinical trials in patients.

A. Target Identification & Validation

Targets are biological molecules, typically proteins, that can be used to rectify the molecular mechanisms associated with human disease, normally through one of several possible therapeutic modalities. Perhaps the most straightforward AI-based method for identifying drug targets is analyzing vast and varied multi-omics datasets (e.g., genomics, transcriptomics, metabolomics). This approach enhances comprehension of disease biology and highlights the connections between diseases and potential targets. A notable instance is TargetDB [1], which integrates public data for specific targets and employs machine learning to rank target drugability. Its method and scoring framework offer valuable metrics for evaluating potential ligand affinities and determining the priority of drug targets for further development. Other laboratory-based approaches, such as phenotypic screening, have also been improved through Convolutional Neural Networks (CNNs). Thus, selecting a specific method hinges on the discovery approach, volume of data, and the research objective being addressed.

Given that targets operate in large multi-component molecular pathways, analysis of these protein networks has emerged as a dominant strategy in target identification. Such network representations lend themselves to a) network-based analyses [2] and b) machine learning approaches [3], particularly in connection with knowledge graphs (see below). A knowledge graph (KG) [4] is a knowledge base that uses graph-structured data models to integrate data in so-called Graph Databases [5]. Because of the highly interconnected nature of biomedical data, the pharmaceutical industry has been one of the early adopters of graph databases, enabling more natural knowledge models, improved data integration workflows, visualization and analysis facilities.

In drug discovery, there are three essential concepts to a KG: nodes, edges and meta-paths. Nodes represent different entities, such as genes, proteins, drugs, diseases, or biological pathways. Edges, on the other hand, define relationships between nodes. For example, an edge might indicate a "gene-protein" relationship, a "drug-target" interaction, or a "protein-protein interaction." Finally, meta-paths describe predefined sequences of node types and relationships that capture specific patterns of interactions or associations within the KG. These paths are denoted using a combination of node types and relationship types. For example, a simple meta-path in a gene-protein-drug KG could be: "Gene - Protein - Drug." By structuring complex biological relationships in this manner, KGs allow researchers to traverse through specific relationships and identify meaningful patterns. Meta-paths, for example, can be used to compare entities based on shared paths, enabling the identification of similar proteins, drugs, or diseases that are involved in similar biological processes. They can also be used to generate features for machine learning models in drug discovery. These features capture structural patterns in

the graph and enhance predictive modeling tasks. Finally, researchers can use meta-paths to validate hypotheses by examining the presence or absence of specific paths in the KG, helping to confirm or refute potential drug-target associations. Overall, meta-paths provide an interpretable way to analyze the complex relationships within drug discovery KGs, facilitating the discovery of new insights and aiding decision-making processes.

Tool	Description	Availability	URL
BioGraph	BioGraph is a bioinformatics platform that leverages KG's and network analysis for drug discovery. It integrates diverse biological data sources to create a comprehensive KG that can be used to identify potential drug targets and biomarkers.	Free	http://biograph.pa.icar.cnr.it/
Cytoscape	Cytoscape is an open-source bioinformatics software platform widely used for network visualization and analysis. It provides various plugins and extensions that enable the construction and exploration of Kg's in the context of biological networks.	Free	https://cytoscape.org/
Neo4j	Neo4j is a popular graph database management system that allows the storage and querying of large-scale KG's. It can be used to build and analyse complex biological networks, aiding in target discovery and other life sciences research.	Free & Commercial	http://neo4j.org
GraphDB	GraphDB is a semantic graph database that offers advanced graph querying and reasoning capabilities. It is utilized in life sciences and biomedical research for building and querying KG's to aid in target discovery and drug development.	Free & Commercial	https://www.ontotext.com/products/graphdb/download/
OrientDB	OrientDB is a multi-model graph database that supports graph, document, and key-value data models. It can be used to create and explore KG's for various applications, including target discovery.	Free	http://orientdb.org/
Smartgraph	SmartGraph is an innovative platform that utilizes state-of-the-art technologies such as a Neo4j graph-database, Angular web framework, RxJS asynchronous event library and D3 visualization to address the needs of systems pharmacology.	Free	https://smartgraph.ncats.io/

Table 1: Overview of Knowledge Graph technologies that can be employed for Target Identification.

i) Network Analysis Methods

In network analysis, so-called "centrality measures" are used to identify the most important vertices within a graph [6]. For drug target identification, centrality metrics can help pinpoint proteins or genes that play crucial roles in biological networks. The following centrality measures are particularly relevant in this regard.

- **Degree Centrality:** This is the simplest form of centrality and is based on the number of edges incident upon a node (i.e., the number of ties a node has). In the context of protein-protein interaction networks, a protein with high degree centrality would be one that interacts with many other proteins.
- **Betweenness Centrality:** Measures the extent to which a vertex lies on paths between other vertices. Nodes with high betweenness may have significant influence within a network by virtue of their control over information passing between others. Proteins with high betweenness centrality may be critical for the communication between different parts of a biological network.
- **Closeness Centrality:** Indicates how close a node is to all other nodes in the network, calculated based on the sum of the shortest paths between a node and all other nodes. In biological contexts, proteins with high closeness centrality might be central in terms of signal propagation or metabolic pathways.
- **Eigenvector Centrality (and its variant, PageRank):** Instead of assuming all nodes are equal, eigenvector centrality gives nodes a score proportional to the sum of the scores of their neighbors. A node is considered important if it is connected to other important nodes. For drug target identification, it can be used to identify proteins that might not have many interactions (low-degree centrality) but are crucial because they interact with other highly connected proteins.
- **Katz Centrality:** This is a measure of the number of all walks ending at a given node, considering more weight to short walks. It is somewhat similar to eigenvector centrality but introduces a damping factor to give more importance to nearby nodes.
- **Subgraph Centrality:** Measures the number of closed walks of different lengths starting and ending at a node. It can identify nodes that are involved in multiple loop structures, often crucial in metabolic or signaling pathways.

ii) Machine Learning Approaches

When combined with machine learning, KGs can facilitate the discovery of new relationships, the prediction of drug-target interactions, and the understanding of complex biological mechanisms [7]. The following represents the most promising ML methods for target identification.

- **Graph Neural Networks (GNNs):** GNNs are deep learning models designed to process data on graphs. They operate by propagating node information along edges, enabling the capture of graph topology and node interactions. GNNs can be used to predict how certain drug molecules might interact with potential protein targets or to infer new relationships between entities.

- **Embedding-Based Methods:** These techniques generate vector representations (embeddings) for nodes and relationships in a graph. Common methods include TransE, DistMult, and ComplEx. Once entities and relationships in a knowledge graph are embedded in a continuous vector space, similarities and potential relationships can be deduced. For drug target identification, embeddings can help in predicting novel drug-target interactions or potential side-effects based on proximity in the embedding space.
- **Random Walks and Path-Based Methods:** These methods leverage the paths or sequences of nodes in a graph. Random walks, for example, involve moving from one node to another randomly, capturing the local neighborhood structure. Random Walk with Restart (RWR) can be used to prioritize potential drug targets based on the likelihood of reaching them from a given disease node.
- **Matrix Factorization:** Matrix factorization decomposes a large matrix (like an adjacency matrix of a graph) into the product of two lower-dimensional matrices. When applied to knowledge graphs, matrix factorization can uncover latent relationships, potentially revealing hidden drug-target interactions.
- **Relational Learning:** Relational learning techniques, like inductive logic programming, are designed to deal with structured data and can use background knowledge (like the relationships in a knowledge graph) to improve predictions. Relational learning can be used to infer new relationships in the graph, potentially identifying new drug-target interactions based on existing knowledge.
- **Semi-Supervised and Transfer Learning:** These techniques leverage both labeled and unlabeled data or transfer knowledge from one domain to assist learning in another. Given that biomedical datasets often have limited labeled data, these methods can be crucial in drug target identification, using known drug-target interactions to predict unknown ones.

By employing these machine learning methods on KGs, researchers can efficiently analyze vast amounts of interconnected biomedical data to identify promising drug targets, making the drug discovery process more efficient and data-driven.

The Pharos platform is a good example of an ML-ready knowledge graph developed by the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH). It aims to integrate and organize diverse data related to drug targets and diseases and provides a user-friendly interface to explore relationships, making it a valuable resource for target identification and pathway analysis. The system offers several useful features and functionalities, including Data Integration, Network Visualization, Target Prioritization, Disease-Gene Associations, Pathway Analysis, Data curation and Data Export.

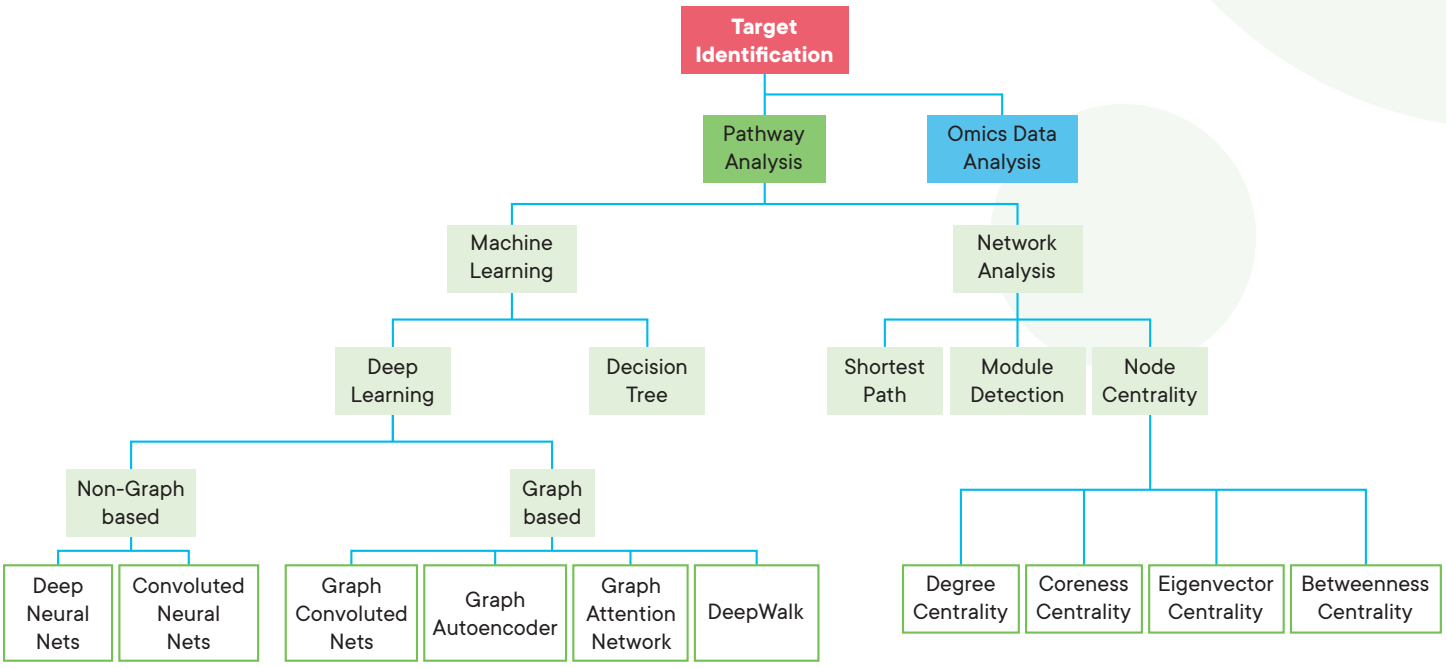


Figure 4: Overview of the main methods (Machine Learning based and Network-based), used in combination with knowledge graphs, to identify drug targets.

B. Hit-to-Lead identification

Once a protein target is identified, the challenge shifts to identifying small molecule chemistries that can bind and modulate the target in the desired manner. Traditional wet-lab techniques can screen millions of molecules in a matter of days. Yet, the task of identifying hit candidates remains enormous. This complexity arises because the universe of potential drug-like compounds has been estimated at around 1060; a million times more than all the atoms on Earth. For this reason, the quest to identify hit candidates was one of the initial applications of machine learning in drug discovery. We can computationally predict which molecules might bind effectively to a target through virtual screening. A technique called "molecular docking" determines how well the physicochemical properties of potential molecules match those of the target protein [9]. Though not exclusively an ML technique, molecular docking tools often incorporate ML to rank and predict how well potential drug molecules bind to their target proteins. The capacity of virtual screening eclipses that of wet-lab methods, spanning from 109 molecules in accessible virtual platforms to an astounding 1015 in specialized pharma collections, which is 4 to 9 orders of magnitude greater than the 106 molecules typically processed in wet-lab screenings. Quantitative Structure-Activity Relationship (QSAR), which uses statistical methods to predict the biological activity based on the chemical structure, provides another important element. Traditional ML techniques like linear regression, Support Vector Machines (SVM), and random forests are used in QSAR modeling to predict the biological activity of molecules based on their chemical structure. Beyond these methods, a variety of other more advanced ML approaches are now being applied, including.

i) Deep Learning

Deep learning can integrate various data types (e.g., genomic data, proteomic data, and chemical structure data) in hit identification, providing a holistic approach to understanding and predicting compound behavior. This capacity of deep learning to model complex, non-linear relationships from vast amounts of data has rendered it an invaluable tool for hit identification.

- **High-Throughput Screening Data Analysis:** Deep learning models can process and analyze vast datasets generated from High-Throughput Screenings (HTS) to identify potential hits.
- **Compound Property Prediction:** Deep learning models, especially Convolutional Neural Networks (CNNs), can be trained to predict various properties of compounds, such as solubility, toxicity, and binding affinity. This aids in filtering and prioritizing compounds before expensive and time-consuming experimental validation.
- **Molecular Representation and Analysis:** Molecular structures can be transformed into grid-like representations or vectorized forms that can be fed into deep learning models. These models can then be used for tasks like predicting bioactivity or determining molecular properties. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks, for example, can handle sequences, making them useful for sequential representations of molecules, like SMILES strings.
- **Generative Models for Novel Molecules:** Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are deep learning architectures that can design novel molecular structures. The generator creates new molecular structures, while the discriminator assesses them based on desired properties. They can be trained to generate molecules with desired properties, leading to the creation of potentially novel hits that might not have been considered using traditional methods.
- **Bioactivity Prediction:** Deep learning can predict how likely a compound is to interact with a specific biological target, giving insights into its potential therapeutic value. This is crucial in the hit identification stage to ensure the selected compounds have a high likelihood of desired biological activity. Given that molecules can be represented as graphs, Graph Neural Networks (GNNs) have been applied to predict various properties of molecules, from solubility to potential bioactivity, aiding in the identification of promising hits.

ii) Reinforcement Learning

Reinforcement Learning (RL) in the context of hit identification is relatively novel, but it shows promise. Applied to optimize molecular properties by iteratively improving the molecular design as screening progresses, an RL model can prioritize which compounds to test next based on the feedback from previous screenings. This maximises the chances of identifying hits with fewer tests. RL can also be used in tandem with deep generative models to design novel molecules with desired properties. In this setup, the RL agent iteratively proposes molecular structures, receives feedback on their potential as hits (either from experimental data or predictive models), and refines its proposals to improve the overall structure. It can also be integrated with other machine learning models, such as those predicting compound bioactivity, toxicity, or solubility.

Beyond the structural considerations, RL can also be employed to optimize experimental conditions for screening, such as reaction times, concentrations, or temperatures, maximizing the likelihood of successful hit identification. While the potential of RL in hit identification is vast, it is still an emerging area in this domain. Data sparsity, noisy experimental results, and the vastness of the chemical space pose significant challenges. However, with the rapid advancements in RL algorithms and growing investment in AI-driven drug discovery, these challenges will likely be addressed over the coming years.

iii) Transfer Learning

Pre-trained models from one task are fine-tuned on a smaller, specific dataset related to the drug discovery process. This method capitalizes on knowledge from one domain to enhance learning in another, potentially streamlining the hit identification and lead optimization processes. Some of the most prevalent applications include.

- **Chemical Space Embedding:** Deep learning models trained on large chemical databases can learn meaningful representations of chemical space. These representations can be used as starting points for models tasked with identifying potential hits or optimizing leads for a specific target or therapeutic area.
- **Drug-Target Interaction Models:** Models trained on extensive datasets of known drug-target interactions can be fine-tuned for specific protein families or disease areas, allowing for more precise predictions in hit identification.
- **Predictive Toxicology and ADMET:** Models trained on broad toxicity datasets can be adapted for predicting specific toxicological endpoints or optimizing lead compounds for desired ADMET (absorption, distribution, metabolism, excretion, toxicity) properties.
- **Molecular Property Prediction:** General molecular property prediction models can be fine-tuned on specific compound classes or therapeutic areas to improve the accuracy of predictions.

iv) Active Learning

Active learning is a semi-supervised machine learning approach that seeks to efficiently use labeled data by prioritizing the selection of the most informative samples for labeling, thereby potentially reducing the number of required experiments or assays. In the context of hit identification and lead optimization in drug discovery, active learning can play a crucial role.

- **Prioritizing Compound Testing:** At the outset, there may be millions of potential compounds to test. Running experimental assays for all of them would be costly and time-consuming.

  Active learning algorithms, using preliminary data, can rank these compounds based on how likely they are to be hits (i.e., having desired biological activity) or how uncertain the model is about their activity. Compounds that the model deems most informative (e.g., those near decision boundaries or with high uncertainty) are then selected for experimental testing.
- **Iterative Feedback Loop:** Once the selected compounds are tested, the experimental results (labels) are fed back into the machine learning model. The model is then updated or retrained with this new data. This iterative process continues, with the model becoming increasingly accurate over time, and fewer and fewer experimental assays are required to identify promising hits or optimize leads.

- **Optimizing Experimental Design:** In lead optimization, compounds might need chemical modifications to enhance their properties (e.g., potency, selectivity, solubility, etc.). Active learning can guide which modifications to try next by analyzing the structure-activity relationship landscape and suggesting the most informative modifications.
- **Addressing Data Scarcity:** Often, experimental assays are expensive or time-consuming, leading to limited labeled data. Active learning can be particularly useful in such scenarios, maximizing the information gained from each experiment.
- **Integration with Other Techniques:** Active learning can be combined with other techniques like transfer learning, where a model pretrained on a related task is fine-tuned using actively selected samples. This can accelerate the learning process. Similarly, it can be integrated with techniques like molecular docking or virtual screening to prioritize which molecules to simulate or test in-depth.

In summary, active learning acts as a bridge between computational methods and wet-lab experiments, ensuring that computational resources and experimental efforts are focused on the most promising or uncertain areas of the search space. This can lead to faster and more cost-effective hit identification and lead optimization.

Case study: AlphaFold

AlphaFold is an advanced artificial intelligence (AI) program developed by DeepMind, designed to predict the three-dimensional structures of proteins based on their amino acid sequences. The core methods that underpin AlphaFold's success include deep learning, specifically the use of deep neural networks (DNNs), and attention mechanisms. DNNs are capable of learning hierarchical representations of data, which is essential for understanding the complex relationships between amino acid sequences and protein structures. It also uses several types of neural networks, including Convolutional Neural Networks (CNNs) for processing spatial information and fully connected networks for integrating information across the protein sequence.

A key feature of AlphaFold is its use of attention mechanisms, specifically a variant known as the Transformer architecture. Originally developed for natural language processing tasks, the Transformer allows the model to focus on different parts of the amino acid sequence, understanding how distant amino acids might interact to influence the protein's final structure. This ability to capture long-range interactions is crucial for accurate protein structure prediction. It also employs an iterative refinement process, where the model progressively refines its predictions of the protein structure. This process allows the model to improve the accuracy of its predictions by repeatedly adjusting and updating the structure based on the learned relationships between sequence features and structural outcomes. Multiple Sequence Alignments (MSA) are also provided as part of its input data, which help the model to identify evolutionarily conserved patterns, providing critical information about the importance of specific amino acids and their roles in the protein's structure. Finally, the model is trained in an end-to-end manner, directly learning to predict the final protein structure from the amino acid sequence and MSA inputs. This approach allows the model to automatically learn the most relevant features for structure prediction, without the need for manual feature engineering.

Using this ensemble of approaches, AlphaFold has demonstrated remarkable accuracy in predicting protein structures, making it a highly valuable tool in the field of Structure Based Drug Design (SBDD). The success of AlphaFold in the Critical Assessment of Structure Prediction (CASP) competitions, particularly CASP14, where it achieved median accuracy scores comparable to experimental methods, has underscored its potential utility in drug discovery processes. In protein engineering, AlphaFold's ability to predict protein structures with high accuracy is allowing scientists to design new proteins or modify existing ones with desired properties by

understanding how changes in amino acid sequences affect protein structure and function. This can lead to the development of enzymes with enhanced catalytic activity, stability, or specificity, as well as the creation of novel proteins with potential therapeutic applications. While AlphaFold brings substantial benefits to SBDD, there are challenges. For example, the dynamic nature of proteins means they can adopt multiple conformations. AlphaFold's static predictions may not capture all functionally relevant states. Moreover, drug molecules might bind to less stable conformations, which could be missed by predictions focusing only on the most stable structure. Predicting how proteins interact with each other or with small molecule ligands, especially in complex and dynamic environments, also remains challenging. Thus, despite the high accuracy of AlphaFold's predictions, experimental validation remains essential, especially for novel or complex targets.

C. Preclinical Development

Model organism testing is a fundamental step in the preclinical drug development pipeline. These animal-based experiments evaluate the Safety, Pharmacokinetics, and Pharmacodynamics of the development candidate with a view to trials in humans. However, there are ethical, cost, time, and scientific concerns associated with animal testing that have spurred increasing interest in the utility of computational methods. ML offers a suite of tools that can complement and, in some cases, potentially replace some aspects of model organism testing, including the following.

- **In Silico Predictive Models:** One of the primary reasons for model organism testing is to determine the potential toxic effects of compounds. ML can analyze vast datasets of known toxic and non-toxic compounds to predict the toxicity of new compounds based on their molecular features. ML models, trained on extensive datasets of drug interactions and their outcomes, can predict the efficacy of a new compound in a particular biological/disease setting.
- **Drug Repurposing:** By analyzing existing data of approved drugs, ML can identify compounds that might be effective against new targets or diseases, reducing the need for extensive animal testing as these drugs have already undergone safety assessments.
- **Data Augmentation and Transfer Learning:** Experimental results from one set of compounds can be used to inform predictions about another set, especially if there's overlap in their chemical space. For instance, if a subset of compounds has been tested in animals, the results can be used to predict outcomes for related compounds without direct testing. Transfer learning, where models trained on one task are fine-tuned for another, can be particularly effective in this context.
- **Organoids and Lab-on-a-Chip Systems:** Organoids are miniaturized, simplified versions of organs produced in vitro in three dimensions that show realistic micro-anatomy. ML can analyze data from organoid systems to predict how drugs might interact in a more complex organism. Lab-on-a-chip systems that mimic the physiological response of entire organs can be combined with ML models for high-throughput drug screening.
- **Phenotypic Screening:** Advanced imaging combined with ML can enable high-throughput phenotypic screening of compounds in cell cultures. ML algorithms can identify subtle changes in cell morphology, behavior, or molecular expression levels that might be indicative of drug effects.
- **Understanding Mechanism of Action (MoA):** Deciphering the MoA of a compound can be complex. ML models, especially those like deep learning, can identify patterns in data to suggest potential pathways or targets that a drug might be affecting.

- **Bridging In Vitro to In Vivo:** ML can be used to translate findings from in vitro (test tube) experiments to predict in vivo (whole organism) outcomes, which could potentially reduce the need for some animal tests.
- **Patient-derived Xenograft Models:** While this still involves animal testing, it is a more targeted approach. Tumor samples from patients are grafted onto mice, and ML can be used to predict which drugs or drug combinations might be most effective for that specific tumor profile.
- **Mining Existing Databases:** With the plethora of biological and chemical databases available, ML can mine these resources to draw connections, predict interactions, and provide insights without the need for new experiments.

While ML holds promise in refining, reducing, and potentially replacing some animal tests, it's essential to note that current technology cannot entirely replace the insights derived from whole-organism testing. Human biology is extremely complex, and in silico models, while informative, have limitations. Thus, the combination of AI/ML with traditional methods offers the most effective and ethical approach to drug discovery and development.

The AI-augmented Laboratory

AI has emerged as a pivotal tool in preclinical drug development, particularly in the context of optimizing laboratory operations, efficiency, and enhancing the robustness of experimental outcomes. The following represent some of the successful applications of ML for laboratory optimization during preclinical stages.

- **Robotics and Automation:** ML-driven robotic systems can handle tasks like liquid handling, plate movements, and sample processing. By learning from previous operations, these systems can improve their accuracy and efficiency over time.
- **Predictive Maintenance:** Algorithms can predict when equipment needs maintenance or is likely to fail by analyzing usage patterns and performance metrics. This reduces downtime and ensures consistent experimental conditions.
- **Quality Control:** ML algorithms can automatically analyze experimental outputs against expected benchmarks to identify deviations or quality issues in real-time.
- **Data Management and Integration:** ML can assist in integrating data from different instruments and platforms, ensuring consistent data annotation and reducing errors. It can also highlight anomalies or outliers in datasets.
- **Resource Allocation:** By analyzing the usage patterns of various lab resources, ML can predict future needs and optimize the allocation of resources such as reagents, equipment time, or personnel.
- **Lab Safety Monitoring:** ML algorithms can monitor lab environments, detecting any deviations from safety norms, such as unusual gas concentrations, temperature changes, or other potential hazards.

Conclusions

AI's integration into preclinical drug discovery is no longer a futuristic aspiration, but rather an ever-evolving practical reality. As the technology matures to augment human expertise, drug discovery will likely experience transformative changes that will benefit the industry and patients worldwide. AI/ML has already begun to radically expedite preclinical discovery and validation processes, such as predicting drug targets, optimizing lead compounds, and simulating in vivo drug effects using in silico candidate PK/PD models, safety and toxicity. With the ongoing expansion of biological datasets and refinement of algorithms, the precision and accuracy of these AI-driven predictions are likely to improve dramatically with time. However, it's crucial to recognize that while AI can streamline and enhance many aspects of the entire drug discovery process, it currently best serves as a powerful augmentation approach.

One of the primary hurdles in harnessing AI's potential is obtaining sufficient, high-quality, and relevant data. Unlike other sectors, where vast amounts of data are available to train machine learning models, drug discovery is challenged with sparse and fragmented datasets. Critical in this regard is the availability of high-quality human-specific clinical and molecular data, particularly when associated with information on treatment outcomes. The scarcity of such data emphasizes the need for continual data evolution, ensuring the data quality and relevance is appropriate for AI application in studying human disease. To address the notorious "garbage-in, garbage-out" issue, there is a pressing requirement to improve the availability of patient outcomes data and refine the representation of complex molecular structures, allowing AI systems to extract vital clinical hypotheses directly from data.

Despite such challenges, the benefits of AI/ML are undeniable. In an era where conventional drug development is costly and fraught with high failure rates, AI provides a beacon of technological hope. It transforms cost-efficiency around activities such as virtual screening. It improves the risk-reward spectrum, particularly for drugs that traditional methods might have overlooked (e.g., the application of AI in drug repositioning). Today, AI's drug discovery capabilities enable researchers to offload and improve routine tasks and usher in a new era of efficient, innovative, and cost-effective drug research and development.

References

1. Stephane De Cesco, John B Davis, Paul E Brennan. **TargetDB: A target information aggregation tool and tractability predictor**. PLoS One 2020 Sep 2;15(9):e0232644. doi: 10.1371/journal.pone.0232644.
2. Miguel Castresana-Aguirre, Dimitri Guala, Erik L L Sonnhammer. **Benefits and Challenges of Pre clustered Network-Based Pathway Analysis**. Front Genet. 2022 May 10:13:855766. doi: 10.3389/fgene.2022.855766.
3. Zhoumeng Lin, Wei-Chun Chou. **Machine Learning and Artificial Intelligence in Toxicological Sciences**. Toxicol Sci. 2022 Aug 25;189(1):7-19. doi: 10.1093/toxsci/kfac075.
4. Brook E Santangelo, Lucas A Gillenwater, Nourah M Salem, Lawrence E Hunter. **Molecular cartooning with knowledge graphs**. Front Bioinform. 2022 Dec 8:2:1054578. doi: 10.3389/fbinf.2022.1054578.
5. Nilesh Kumar, Shahid Mukhtar. **Building Protein-Protein Interaction Graph Database Using Neo4j**. Methods Mol Biol. 2023:2690:469-479. doi: 10.1007/978-1-0716-3327-4_36.
6. Pourya Naderi Yeganeh, Christine Richardson, Erik Saule, Ann Loraine, M Taghi Mostafavi. **Revisiting the use of graph centrality models in biological pathway analysis**. BioData Min. 2020 Jun 16:13:5. doi: 10.1186/s13040-020-00214-x.
7. Wei Zhang, Jeremy Chien, Jeongsik Yong, Rui Kuang. **Network-based machine learning and graph theory algorithms for precision oncology**. NPJ Precis Oncol. 2017 Aug 8;1(1):25. doi: 10.1038/s41698 017-0029-7.
8. Timothy Sheils, Stephen L Mathias , Vishal B Siramshetty,et al. **How to Illuminate the Druggable Genome Using Pharos**. Curr Protoc Bioinformatics. 2020 Mar;69(1):e92. doi: 10.1002/cpbi.92.
9. Laurent Hoffer, Christophe Muller, Philippe Roche, Xavier Morelli **Chemistry-driven Hit-to-lead Optimization Guided by Structure-based Approaches**. Mol Inform. 2018 Sep;37(9-10):e1800059. doi: 10.1002/minf.201800059.

If you're interested in Artificial Intelligence in Preclinical Drug Discovery, please get in touch with us at **info@zifornd.com**

About the Authors

David Jackson

Principal Consultant -
Strategic Consulting and Advisory

Paul Denny Gouldson

Chief Scientist

Aref Abdollah Aghebatrafat

Senior Consultant -
Strategic Consulting and Advisory

Zifo
SCIENTIFIC INFORMATICS