IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Artificial Intelligence In Drug Discovery And Development

1 Sachin Arjun Kumbhar, 2 Praful Ananta Misal, 3 Shivani Mahendra Karale,

4 Laxmi Arun Ghube, 5 Prajakta Dattatray Dahiwal

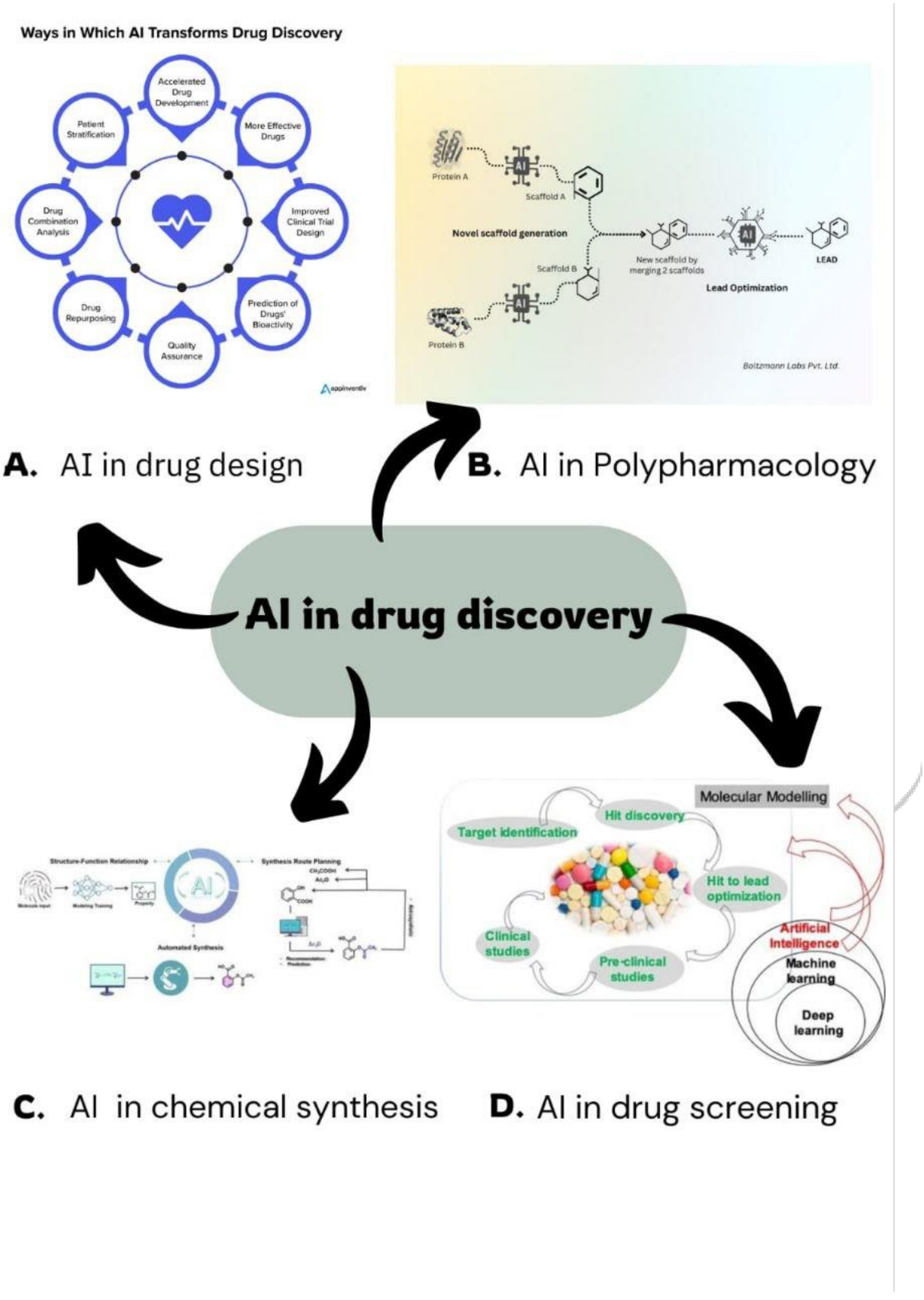
1,2,3,4,5 :- students of Shri Sant Gajanan Maharaj College of Pharmacy Buldhana

Abstract:- Artificial Intelligence (AI) has recently begun to significantly expand its applications across various sectors of society, with the pharmaceutical industry emerging as a leading beneficiary. This review underscores the impactful utilization of AI in multiple areas of the pharmaceutical sector, including drug discovery and development, drug repurposing, enhancing pharmaceutical productivity, and clinical trials, among others. These advancements not only reduce human workload but also expedite the achievement of targets. Additionally, the review discusses the tools and techniques employed in implementing AI, the current challenges faced, and potential solutions, as well as the future prospects of AI in the pharmaceutical industry.

Keywords :- Artificially intelligence ,AI in pharmacy , AI in Primary Drug Screening, The cross-validation method, QSAR/QSPR and Structure-Based Modelling with Artificial Intelligence, Protein Structure and Function, Prediction of Protein Folding from Sequence, Prediction of Protein-Protein Interactions, Drug Repurposing, Virtual Screening, Activity Scoring.

Introduction :- Artificial Intelligence (AI) has emerged as a transformative force in various industries, and its application in drug discovery and development is particularly noteworthy. By leveraging advanced computational techniques, AI has the potential to significantly streamline the drug development process, from initial target identification to clinical trials and beyond. This integration not only accelerates the pace of discovering new therapeutic compounds but also enhances the precision and efficiency of drug repurposing efforts. In this review, we explore the pivotal role of AI in modern pharmaceutical research, examining the methodologies, tools, and techniques that are driving innovation in this field. Furthermore, we address the current challenges and propose potential solutions to fully realize the benefits of AI in drug discovery and development.

The integration of artificial intelligence in drug discovery and development has significantly advanced the pharmaceutical industry, driving a transformative change. This discussion explores the areas of AI integration, the tools and techniques employed, current challenges, and potential solutions.



- Protein Structure and Function

- Prediction of Protein Folding from Sequence

Protein dysfunctions are linked to many diseases. Studying protein structures enables the use of structure-based drug design strategies to identify small molecules that target specific proteins. However, obtaining the three-dimensional (3D) structures of proteins is currently resource-intensive. Therefore, developing algorithms to predict protein 3D structures from sequences is crucial. While sequence data for many proteins is available, accurately predicting their 3D structures de novo remains challenging. Recent advancements in deep learning have improved the prediction of protein secondary structures, backbone torsion angles, and residue contacts. For instance, a deep learning approach that combines one-dimensional (1D) and two-dimensional (2D) convolutional neural networks (CNNs) for predicting residue contacts has demonstrated superior performance in the 12th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP12). The ability of deep learning architectures to learn the relationship between sequence and structure through feature extraction holds promise for advancing 3D structure prediction.

- Prediction of Protein-Protein Interactions

Protein-protein interactions (PPIs) are essential for numerous biological processes and are implicated in various diseases. The String database, which houses approximately 1.4 billion PPIs obtained through both experimental and bioinformatics methods, is a valuable resource. The PPI interface, comprising protein-protein binding sites made up of numerous residues, represents a novel class of drug targets, distinct from traditional targets like G-protein coupled receptors (GPCRs), ion channels, kinases, and nuclear receptors. For example, the inhibitors of protein-protein Database (iPPI-DB) reports 1,756 non-peptide inhibitors across 18 PPI families. Targeting PPIs can expand the target space and enhance small molecule drug development while potentially reducing adverse effects by improving biological selectivity. For instance, compound DC_AC50 inhibits tumor cell proliferation by blocking copper ion transport within cells through interaction with copper-transfer interfaces, without affecting normal somatic cell survival.

Understanding PPI interfaces is vital for structure-based drug design. However, precise PPI information is often limited, prompting the development of computational methods for predicting these interfaces. Template-based methods are generally more reliable due to the conservation of PPI interfaces. For example, the eFindSite web server employs template-based, residue-based, and sequence-based features to develop support vector machine (SVM) and Naive Bayes Classifier (NBC) models for PPI interface prediction. Protein-protein docking methods, such as ZDOCK and SymmDock, predict PPI interfaces when the structures of interacting proteins are known. A key challenge is predicting conformational changes when two unbound proteins form a complex. Deep learning methods can effectively extract relevant sequence features to predict PPI interfaces, showing significant improvement over traditional machine learning methods like SVM.

Given the large buried area of PPI interfaces (1500–3000 Å^2), identifying druggable sites or local regions within these interfaces is essential. Hot spots, which contribute significantly to binding free energy, may represent druggable sites. Bai et al. used fragment docking and direct coupling analysis (FD-DCA) to identify druggable PPI sites. They developed iFitDock, a fragment docking tool to locate druggable hot spots in PPI interfaces. By clustering small hot spots into candidate binding sites and using a scoring function based on evolutionary conservation, they identified promising protein-protein binding sites. Identifying hot spots and designing small modulators targeting PPI interfaces is a promising approach for drug discovery.

- Drug Repurposing

Drug repurposing, also known as drug repositioning, involves identifying new uses for approved drugs. This approach can mitigate the time and risk associated with drug development. Many drugs have multiple targets, which can lead to diverse drug-disease interactions. For instance, Metformin, originally approved for type 2 diabetes, has been observed to potentially extend lifespan.

Key components in drug repurposing include the drug itself, the associated disease, drug targets, and disease genes. Network analysis is employed to illustrate these interactions. There are nine crucial types of networks in drug design: gene regulatory, metabolic, protein-protein, drug-target, drug-drug, drug-disease, target-disease, drug-adverse effect, and disease-disease networks. The fundamental hypothesis is that drugs with similar properties often share similar targets or effects. Due to the limitations of individual networks, integrating multiple networks into a heterogeneous network is crucial for effective drug repurposing. For instance, DTINet integrates data from various networks to predict new drug targets and indications, leading to the discovery of novel effects for existing drugs.

- Virtual Screening

Virtual screening utilizes algorithms and software to identify bioactive molecules from chemical libraries, offering an efficient method for discovering new hits and filtering out undesirable compounds early in drug development. Methods include docking-based, pharmacophore-based, similarity searching, and machine learning techniques. Structure-based virtual screening, like molecular docking, is effective when the 3D structure of a target protein is known. However, limitations such as inaccuracies in scoring functions and considerations of protein flexibility affect its efficacy.

Ligand-based virtual screening does not depend on 3D structural information and instead maps molecular features to bioactivity classes. Machine learning methods, including support vector machines (SVM), have demonstrated high accuracy and reduced false-hit rates. Recently, deep learning techniques have been applied to virtual screening due to their superior classification and feature extraction capabilities. For example, long short-term memory networks and adversarial autoencoders have been used to generate focused molecular libraries and identify potential anticancer agents.

- Activity Scoring

In molecular docking, scoring functions evaluate the binding affinity of drug-like molecules to targets. Machine learning-based scores, such as those derived from random forests (RF) and SVM, offer improved performance by effectively extracting geometric, chemical, and physical force field features. These models predict binding affinities based on experimental data, bypassing complex physical functions. Recent advancements include CNN-based methods for extracting features from protein-ligand interactions, which have shown better predictive power compared to traditional docking programs. Deep learning techniques like CNN can enhance predictive capabilities by learning complex features from basic compound-protein interactions.

In Silico Evaluation of ADME/T Properties

- Physical and Chemical Properties

Early detection of molecules with unfavorable physical or chemical properties significantly mitigates the risk of failure in drug discovery. Various deep learning approaches have been developed to address this issue. For instance, Duvenaud et al. utilized a combination of convolutional neural networks (CNN) and artificial neural networks (ANN) to predict solubility by extracting information directly from molecular graphs. This method demonstrated good predictive performance, with a mean absolute error (MAE) of 0.53 ± 0.07. Its interpretability is a notable advantage, allowing for the identification of fragments, such as hydrophilic R-OH groups, that contribute to molecule solubility.

Building on this work, Coley et al. employed a tensor-based convolutional embedding method to predict molecular aqueous solubility. Their model, which integrates bond-level and atom-level features into a molecular tensor, outperformed Duvenaud's model with an MAE of 0.424 ± 0.005. The use of more detailed atom-level information contributed to this improved performance.

Predicting the Caco-2 permeability coefficient (Papp) is crucial for assessing oral drug absorption. Wang et al. constructed prediction models using Boosting, SVM regression, partial least squares (PLS), and multiple linear regression (MLR) with 30 descriptors. Their Boosting model achieved the best results, with an R^2 of 0.81 and a root mean square error (RMSE) of 0.31 for the test set. The model adhered to the Organization for Economic Co-operation and Development (OECD) principles for quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR), ensuring model reliability.

- Absorption, Distribution, Metabolism, and Excretion

Drug absorption is the process by which drugs enter the bloodstream from the site of administration. Bioavailability, an essential pharmacokinetic parameter, reflects the extent of absorption. Tian et al. developed an MLR model to predict bioavailability using structural fingerprints and molecular properties from a dataset of 1,014 molecules. Their model, which incorporated genetic function approximation for automatic selection of molecular properties, showed a correlation coefficient of 0.71 and an RMSE of 0.2355.

Drug distribution, the process by which drugs circulate to interstitial and intracellular fluids, is quantified by the volume of distribution at steady state (VDss). Lombardo and Jing used PLS and random forest (RF) models to predict VDss from a dataset of 1,096 molecules. Their model had a 50% success rate within a 2-fold error on the external test set, highlighting the challenge of predicting VDss from molecular structure alone due to various influencing factors.

Metabolism involves the transformation of drugs, potentially leading to loss of function or the production of toxic metabolites. Accurate prediction of metabolic sites can guide structural optimization for metabolic stability. Machine learning methods have been employed to predict sites of metabolism by different enzymes, such as cytochrome P450s (CYP450s) and UDP-glucuronosyltransferases (UGTs). XenoSite, based on neural networks, predicts CYP450 metabolism sites with 87% accuracy and also utilizes a neural network trained on UGT metabolism data.

Drug excretion is the elimination of drugs and metabolites from the body. While water-soluble metabolites are typically excreted easily, some drugs can be directly excreted. Lombardo et al. used principal component analysis (PCA) to predict primary clearance mechanisms, achieving an 84% predictive accuracy. Their

subsequent PLS model for total human clearance performed well and was competitive with animal scaling methods.

- Toxicity and ADME/T Multi-Task Neural Networks

Toxicity is a major concern in drug development, causing attrition of approximately one-third of lead compounds. Predicting toxicity is critical for optimizing lead compounds and reducing development risks. Traditional methods, relying on expert knowledge and structural alerts, often result in false positives and incomplete feature coverage. Deep learning models, however, offer improved performance in toxicity prediction due to their ability to handle diverse chemical characteristics and automatically extract features.

Xu et al. developed an acute oral toxicity prediction model using molecular graph encoding with convolutional neural networks (MGE-CNN). This model outperformed previous SVM-based models and allows for flexible adjustments of molecular fingerprints. Xu et al. also mapped toxicological features to atomic levels, highlighting fragments related to structural alerts.

Mayr et al. created a multi-task deep neural network (DNN) model, DeepTox, to predict toxicity, which outperformed other models in the Tox21 challenge. The multi-task approach, sharing parameters across related tasks, generally provides better performance by learning common features.

Combining ADME/T predictions in a multi-task neural network framework can enhance predictive performance across these tasks. Kearnes et al. compared single-task and multi-task neural networks using ADME/T experimental datasets and found that multi-task models delivered superior results.

- QSAR
- QSAR/QSPR and Structure-Based Modeling with Artificial Intelligence

QSAR/QSPR modeling has evolved significantly since its inception over 50 years ago. These computational models have had a profound impact on drug discovery, particularly in the successful prediction of biological activity and pharmacokinetic parameters such as absorption, distribution, metabolism, excretion, and toxicity (ADMET). For ligand-based QSAR/QSPR modeling, the structural features of molecules (e.g., pharmacophore distribution, physicochemical properties, and functional groups) are often translated into machine-readable numbers using molecular descriptors. These descriptors aim to capture a variety of aspects of the underlying chemical structure. Over time, QSAR/QSPR approaches have moved from simpler models, like linear regression and k-nearest neighbors, to more advanced machine learning techniques such as support vector machines (SVM) and gradient boosting methods (GBM). These advanced techniques aim to address complex and potentially nonlinear relationships between chemical structures and their properties, though often at the expense of interpretability.

Deep learning, although not new, gained significant traction in chemoinformatics following the success of neural networks in the 1990s and their breakthrough in the Merck Molecular Activity Challenge in 2012. Deep learning methods, including graph neural networks and recurrent neural networks, offer several advantages. Notably, these networks can perform automatic feature extraction during training. Graph neural networks, in particular, generate internal context-specific representations of molecular structures by learning latent atom and bond representations. Deep learning is promising for tasks that classical descriptors were not initially designed for, such as modeling peptides, macrocycles, and proteolysis-targeting chimeras (PROTACs). Additionally, deep learning is advantageous for multitask learning, which aims to find common

internal representations useful for related endpoints. This is particularly beneficial in drug discovery, which is a multiparameter optimization challenge.

However, deep learning has limitations, particularly its poor performance in scenarios with medium-to-low data availability. Some chemogenomic-based approaches might provide insights in these scenarios by leveraging additional genomic or biological interactome data. Advances in few-shot learning and meta-learning hold promise in mitigating data scarcity issues. Additionally, purely data-driven approaches for molecular property predictions face fundamental limitations in extrapolating and making reliable predictions for unseen compound classes. Physics-inspired machine learning approaches and active learning strategies provide tools to overcome these limitations, though their success depends on how well they handle data sparsity.

Deep learning models are often criticized for their difficulty in debugging and their 'black-box' nature. In contrast, manually developed domain-specific features can integrate background knowledge in a more human-intelligible way. Explainable AI techniques, including feature attribution and attention-based networks, could help bridge the gap between deep learning and drug discovery specialists, making close collaboration between these fields essential.

Another drawback of deep learning approaches is their high computational cost. Without specialized hardware, deep learning typically requires longer training and evaluation times compared to other machine learning approaches. However, deep learning models can learn in an online setting using stochastic gradient descent optimization, making them suitable for big data scenarios. Deep learning also tends to require more human expertise for practical applications compared to other methods. For instance, training a well-performing random forest model requires relatively less effort for hyperparameter tuning compared to deep learning models.

Moreover, neural networks might provide correct answers for misleading reasons and tend to produce overly confident predictions, even when they are wrong. This issue might be alleviated with the adoption of uncertainty estimation techniques, such as Bayesian neural networks or ensemble learning.

Significant progress has also been made in structure-based prediction of protein-ligand activities. This field has transitioned from classical approaches, which modeled explicit mathematical relationships of protein-ligand complexes, to more advanced and flexible nonlinear models like random forests and SVMs. Deep learning has further advanced this field, with techniques inspired by computer vision and image recognition being adapted for bioactivity prediction. Recent research focuses on overcoming theoretical limitations of three-dimensional convolutional neural networks, such as the lack of rotational invariance, with new neural network architectures like Euclidean Neural Networks and SchNet.

The growth of deep learning applications in drug discovery necessitates diligent data curation and proper benchmarking of newly developed models. The availability and size of chemical compound libraries have improved, with databases like ZINC and ChEMBL serving as starting points for ligand-based projects. Structure-based modeling has benefited from databases such as PDBbind and BindingDB, which provide detailed structural information on protein-ligand complexes. Standardized assessments of machine learning methodologies, such as the MoleculeNet benchmarking suite, aim to facilitate model testing by providing timely evaluations of popular deep learning architectures. However, most structural activity/property relationship data are still generated by commercial research organizations, publishers, and pharmaceutical

companies, which often keep this data confidential. Efforts are underway to develop federated and IP-preserving learning techniques to overcome these limitations.

Model evaluation practices are evolving, with alternatives to pseudo-random performance testing, such as scaffold-based or time-based splits, offering more informative assessments. Prospective applications are considered the gold standard for model benchmarking, though they are not without biases. Despite the lack of benchmarking consensus, machine-learning scoring functions have shown promise in virtual screening campaigns. Proper performance metrics for classification and regression models, and their limitations, continue to be a focus of dedicated efforts.

- De Novo Drug Design with Artificial Intelligence

De novo design, the creation of novel molecular entities with desired pharmacological properties from scratch, is one of the most challenging computer-assisted tasks in drug discovery due to the vastness of the chemical space of drug-like molecules (estimated to range from \(10^{60}\) to \(10^{100}\)). This process faces the issue of combinatorial explosion because of the numerous atomic types and molecular topologies that can be investigated. Approaches to de novo design can be ligand-based, structure-based, or a combination of both, depending on the guiding information used.

- Ligand-Based Methodologies

Ligand-based methodologies fall into two main categories:

1. Rule-Based Approaches: These use a set of construction rules for molecule assembly from 'building blocks' such as reagents or molecular fragments. An early example is the Topliss scheme, which generates analogs of an active lead compound to maximize potency. Modern methods apply molecular transformations for optimization, like matched molecular pairs or rules-of-thumb for functional group and molecular framework modification. Synthesis-oriented approaches include synthesis rules for building block assembly and ligand generation, useful for designing synthetically accessible libraries.

2. Rule-Free Approaches: These aim to generate molecules with desired properties directly, without construction rules. Contemporary methods often employ generative deep learning models, which sample new molecules from a learned latent molecular representation. This concept, dating back to the 'inverse QSAR' problem of the early 1990s, uses existing QSAR models to identify descriptor values for desired properties and generates molecules accordingly. Generative deep learning models, such as those borrowed from natural language processing (using SMILES syntax), recurrent neural networks, variational autoencoders, and generative adversarial networks, have been popular. These models can leverage additional information, like three-dimensional shape, drug-likeness, synthesizability, molecular descriptors, and gene expression signatures.

- Evaluation and Challenges

The rapid development of generative neural network approaches has led to an increase in ligand-based design methods, with over 40 new models developed in recent years. This proliferation has driven efforts to evaluate and benchmark these approaches in a standardized manner. Platforms like MOSES and GuacaMol implement various generative models and provide metrics for comparison, focusing on validity, novelty, similarity to known compounds, and scaffold and fragment diversity.

- Rule-Based vs. Rule-Free Approaches

Both approaches have distinct advantages. Rule-based methods generate readily synthesizable molecules with desired properties by relying on preexisting knowledge. However, the chemical diversity is limited by hard-coded rules and chosen building block libraries. Rule-free methods, learning directly from data, theoretically explore a broader chemical space but may produce compounds difficult to synthesize. Mixed approaches, combining rule-free and rule-based methods, show promise in designing novel, bioactive, and synthesizable molecular entities.

- Structure-Based Generative Design

Most deep-learning-based de novo design studies have focused on ligand-based approaches. Structure-based generative design, which uses information about ligand-binding sites, offers a complementary research direction for targeting orphan receptors and unexplored macromolecules. While not extensively permeated by deep learning yet, initial developments consider the shape and properties of the binding pocket for ligand design.

3. Automated Synthesis Planning with Artificial Intelligence

The majority of known organic compounds can be synthesized using a limited set of robust reactions. However, achieving reliable and fully automated synthesis planning in chemistry remains a challenge. This is largely due to the extensive chemistry expertise required for efficient forward and retrosynthetic planning. AI-driven synthesis planning has a long history, dating back to the 1970s with the advent of computer-aided retrosynthetic prediction. Advances in computational power, big data, and novel algorithms for deep learning and optimization have revitalized AI's role in synthetic organic chemistry.

In retrosynthesis, the primary goal is to recursively design efficient synthetic routes for a target molecule. Rule-based methods have been particularly valuable in this area, suggesting retrosynthetic pathways via reaction mechanism encoding and skeletal building. However, these methods are limited by their dependence on explicitly defined chemical transformations, which typically require manual construction and curation.

Recent research has drawn inspiration from natural language processing methods, such as sequence-to-sequence models and transformer models, driven by the observation that the rank distribution of fragments in organic molecules is similar to that of words in the English language. Rule-free approaches use text-based representations of products (e.g., SMILES) and process them via an encoder-decoder architecture to predict corresponding synthetic precursors at a one-step reaction distance. Improvements over this architecture include tiered neural networks, which partition the retrosynthesis prediction problem into reaction type classification and reaction rule selection steps. This separation has been shown to achieve performance gains over previous baselines.

While many methods focus on the linear one-step retrosynthesis problem, real-world scenarios involve rapidly exploding combinatorial problems. Inspired by progress in reinforcement learning, a significant breakthrough in recent years has been the use of sophisticated search methods, such as Monte Carlo Tree Search, to efficiently navigate chemical reaction spaces. One study elucidated both reactants and reagents using transformer models for one-step precursor predictions, combined with hyper-graphs to represent synthetic pathways. These hyper-graphs are explored with beam search, aided by a Bayesian-like probability scheme that biases toward suggesting chemically simpler precursors.

Forward synthesis planning differs from retrosynthesis in that it often requires information from reactions that yield no product. Current chemical reaction databases are heavily skewed toward productive reaction data, creating a demand for additional data, such as experimental conditions and side-product information. Efforts are being made to expand reaction databases with negative outcomes to create new customized data compilations for automated synthesis planning.

Earlier approaches ranked candidate products using hard-coded reaction templates derived from data. Proof-of-concept machine learning methods ranked reaction templates when details of reactants and reagents were provided. Newer approaches view the chemical reaction prediction problem as a graph transformation task, ranking products directly. Advances in quantum mechanics have also led to approaches using first-principle calculations to evaluate reaction energy barriers, although these are computationally prohibitive for medium-to-large systems. Quantum-mechanical machine learning may help bridge this gap in the future.

Template-free forward synthesis prediction has seen the rise of natural language processing approaches based on transformer or recurrent neural network architectures, achieving top-1 reactant accuracy above 90%.

Other deep learning approaches encode reaction prediction as an electron rearrangement task, using message-passing neural networks, though this method requires filtering out reactions where electron flow is not directly identifiable, excluding many relevant organic reactions.

- Machine Learning Strategies and Programs for Drug Design

Methods of Molecular Representation

In drug design, molecular representations such as molecular fingerprints, numbers, ASCII strings, and graphs are utilized as input features for machine learning methods.

Molecular Fingerprints encode molecular attributes as binary sequences where a "1" indicates the presence of a particular attribute and a "0" indicates its absence. These fingerprints are widely used to predict molecular properties and assess molecular similarity due to their simplicity and effectiveness. Commonly used 2D structure-based molecular fingerprints include:

- Molecular ACCess System (MACCS)

- Extended-Connectivity Fingerprint (ECFP)

- Functional Class Fingerprint (FCFP)

- Molprint2D

For instance, MACCS fingerprints have been used to train autoencoder models for identifying anti-cancer molecules.

Molecular Graphs have long been employed by chemists to qualitatively analyze molecular structures. Recent advancements in artificial intelligence (AI) have enabled quantitative analysis through convolutional neural networks (CNNs). CNNs can automatically extract features from molecular graphs for bioactivity prediction, toxicity assessment, physicochemical property evaluation, and protein-ligand affinity estimation. Graph convolutional methods offer flexibility, as the graph architecture can be tailored to specific tasks. These methods can be integrated with neural networks for simultaneous feature extraction and model training. Notable graph convolutional fingerprints include:

- Duvenaud's Fingerprints: Based on atomic radiation methods, where atomic and bond features are encoded and used to generate initial molecular feature vectors.

- Kearnes's Fingerprints: Based on atoms, bonds, and pairwise relationships.

- Coley's Fingerprints: Based on molecular tensors.

Duvenaud's graph CNN fingerprints, for example, generate interpretable molecular features, with successful implementations in the DeepChem toolbox demonstrating superior performance compared to other models.

Recursive Neural Networks (RNNs) can also represent molecules effectively. For instance, Urban's recursive networks have shown improved prediction accuracy on public datasets compared to other methods.

- String Representations of small molecules include:

- Wiswesser Line-Formula Notation (WLN)

- SYBYL Line Notation (SLN)

- SMILES (Simplified Molecular Input Line Entry System)

- International Chemical Identifier (InChI)

Among these, SMILES is particularly popular and supported by numerous programs (e.g., ChemDraw, Cheopy, RDKit) and databases (e.g., PubChem, ZINC). RNNs can learn SMILES coding grammar and convert it into molecular graphs or use it directly to predict molecular properties.

Molecular Descriptors refer to structural or physicochemical properties of molecules and can be derived from molecular encoding or experimental data. The appropriate selection of descriptors is crucial for enhancing model efficiency, generalization, and interpretability. Common software tools for calculating molecular descriptors include:

- Dragon

- Cheopy

- PaDEL

- Cinfony

Transfer Learning for Low Data

Deep learning techniques have demonstrated considerable potential in drug design due to their robust data mining capabilities. However, these methods typically require large amounts of training data, which limits their application in scenarios with limited data availability. For instance, predicting the bioactivity of new molecules is challenging with minimal activity data, as it may not capture sufficient chemical diversity.

Transfer learning addresses this issue by leveraging knowledge from related data sources. Similar to how human experts apply previously acquired knowledge to new problems, transfer learning aims to replicate this capability. The core principle involves utilizing knowledge from past tasks to improve performance on a related target task with limited data.

One-shot learning is a related approach that focuses on deep learning methods requiring only a few training samples. It enables the transfer of information between relevant but distinct tasks by learning meaningful distance metrics. Altae-Tran et al. developed a one-shot learning method combining iterative refinement of long short-term memory networks with graph convolutional networks for low-data scenarios. This model has shown superior performance compared to traditional methods such as random forests on datasets like Tox21 and SIDER. However, when trained on toxicity data to predict side effects, this model may fail due to the weak relevance between the datasets.

- The cross-validation method

The cross-validation method is used to assess model performance, with random-split cross-validation being a common approach. However, this method can be overly optimistic in estimating predictive performance because it mixes data from different time periods, potentially diluting the impact of covariate changes in drug development. An alternative is time-split cross-validation, where data is divided into training and test sets based on the temporal order of experiments. Research has shown that time-split cross-validation provides a more accurate estimate of predictive value compared to random-split methods. For instance, time-split cross-validation has been found to yield R^2 values that more closely reflect true prospective predictions. Consequently, it is advisable to use time-split cross-validation in drug discovery when temporal data is available, as demonstrated in studies evaluating the performance of deep neural networks in simulating the hit-to-lead process.

Training deep neural networks presents challenges due to their complex architectures and numerous parameters. These difficulties are exacerbated when sample sizes are limited or feature matrices are sparse, often resulting in suboptimal local minima and unsatisfactory accuracy. To address these issues, unsupervised pre-training methods, such as deep belief networks, have been proposed to enhance parameter initialization. Research indicates that these methods are more effective than random initialization. Additionally, dropout strategies have been shown to effectively prevent overfitting in QSAR datasets. Furthermore, the ReLU

activation function is preferred over the sigmoid function for QSAR tasks due to its advantages in mitigating the vanishing gradient problem and avoiding local minima.

- AI in drug screening
- AI in Primary Drug Screening

- Sorting and Classification of Cells through Image Analysis

AI has proven highly effective in image recognition, particularly in identifying distinct objects or features within images. Traditional visual inspection methods are often inefficient and labor-intensive, especially when dealing with large datasets. AI-based computing technologies are well-suited for such applications. In cell target classification or diagnosis, AI models must be trained to automatically identify and categorize different cell types based on their features. For instance, to classify breast cancer cells, images are first segmented from their backgrounds through contrast adjustments. Features such as Tamura texture and wavelet-based texture are then extracted and reduced in dimension using principal component analysis (PCA). AI models, such as least-squares support vector machines, are then trained for classification tasks.

In cell sorting, AI-driven image analysis must be rapid enough to allow robots to accurately separate different cell types. Modern image-activated cell sorting (IACS) devices employ optical, electrical, and mechanical measurements to facilitate high-speed sorting. These systems utilize AI-based deep neural network (DNN) algorithms for real-time image processing and decision-making, often within milliseconds. This approach has demonstrated high specificity and sensitivity in sorting tasks involving Chlamydomonas reinhardtii and human platelets.

AI is also making strides in interpreting computerized electrocardiography (ECG), streamlining the diagnostic process by reducing the reliance on manual inspection by practitioners. The use of deep learning (DL) algorithms with digital ECG data has significantly enhanced the accuracy and scalability of automated ECG analysis.

- AI in Secondary Drug Screening

- Predictions of Physical Properties

In drug design, selecting candidates with optimal properties—such as bioavailability, bioactivity, and toxicity—is crucial. Physical properties like melting point and partition coefficient (logP) play a significant role in determining a drug's bioavailability. AI algorithms use molecular representations, such as molecular

fingerprints, SMILES strings, and Coulomb matrices, to predict these properties. Deep neural networks (DNNs) are employed in a two-stage process: a generative stage to create feasible molecular structures and a predictive stage to estimate molecular properties. This approach, which may incorporate reinforcement learning, facilitates the design of drugs with desirable characteristics.

- Predictions of Bioactivity

Matched molecular pair (MMP) analysis assesses the impact of localized changes in drug candidates on their properties and bioactivity. This method, often used in quantitative structure–activity relationship (QSAR) studies, generates MMPs through retrosynthesis rules. Machine learning methods, including random forest, gradient boosting machines, and DNNs, are then applied to predict new transformations and modifications. Studies have shown that DNNs generally outperform other methods in predicting compound activity. With the expansion of public databases like ChEMBL and PubChem, MMP analysis has been extended to predict various bioactivity properties, including oral exposure, distribution coefficient (logD), and absorption, distribution, metabolism, and excretion (ADME).

Recent advancements also include using graph convolutional networks to extract drug target site signatures, allowing predictions based on continuous latent vector spaces and differentiable models of binding affinity.

- Prediction of Toxicity

Accurately predicting a compound's toxicity is a critical and often resource-intensive task in drug development. The DeepTox algorithm, a machine learning-based approach, has excelled in toxicity prediction challenges, such as the Tox21 Data Challenge. This algorithm processes chemical representations to compute numerous descriptors, both static and dynamic, to predict toxic effects.

Despite the complexity and variety of potential dynamic features, DeepTox maintains manageable dataset sizes and demonstrates strong accuracy in predicting compound toxicity.

- Planning Chemical Synthesis with AI: Retrosynthesis Pathway Prediction

Retrosynthesis is a complex method for designing organic synthesis, significantly enhanced by advancements in AI. After a molecule has been virtually screened for bioactivity and toxicology, finding an optimal chemical synthesis pathway begins. This step is often challenging and inefficient. Despite extensive knowledge of transformation steps, novel molecules with unique structural features or conflicting reactivities may not be easily synthesized.

Retrosynthesis analysis involves recursively searching for 'backward' reaction pathways until simpler, available precursors are identified. Monte Carlo tree search (MCTS) is particularly suited for this process, as it performs random search steps without branching until an optimal solution is found. Previous algorithms for computer-assisted synthesis planning (CASP) have not gained widespread popularity, as they relied heavily on manually encoded knowledge, which does not scale with exponentially growing chemical knowledge and often lacked chemical insight.

Machine learning (ML) approaches trained on empirical data now offer improved methods: they predict the likelihood of a transformation at specific branching points and guide the selection of random steps. AI algorithms can be trained on literature regarding yields and costs of transformation rules to predict the most feasible retrosynthesis pathway for a given molecule.

The 3N-MCTS method combines three neural networks with MCTS to create a workflow for CASP. Each network handles a distinct task: an expansion node explores new transformation possibilities, an update node evaluates pathways, and a rollout node uses frequently reported transformation rules for efficient search and evaluation. The 3N-MCTS method has demonstrated superior performance, solving 80% of retrosynthesis problems within a 5-second time limit and over 90% within 60 seconds. It operates 20 times faster than traditional Monte Carlo methods.

- Reaction Yield Prediction and Insights into Reaction Mechanism

AI algorithms not only design synthesis routes but also predict the products and yields of organic reactions based on molecular properties. Historically, predicting complex chemical reaction outcomes has been challenging. Quantum chemistry methods such as Hartree–Fock, semi-empirical methods (e.g., AM1, PM3), and density functional theory can model experimental outcomes in silico effectively.

Recent studies have utilized AI to automate and enhance yield prediction. For instance, Doyle and Dreher demonstrated that ML could predict yields for a Buchwald–Hartwig coupling reaction—a key process for synthesizing carbon–nitrogen bonds in pharmaceuticals. By using quantum chemistry-derived descriptors and high-throughput experimental data, machine learning approaches like Random Forest (RF) have successfully explored relationships between these descriptors and product yields, achieving promising accuracy in predicting yields for various reactant variants.

- Automation of Chemical Synthesis with AI

- Digitization and Standardization of Synthesis

There are significant initiatives aimed at leveraging AI to automate chemical synthesis with minimal manual intervention. Established technologies, such as the ‘solid phase’ method—where the polymer chain is attached to an insoluble matrix—have already automated the synthesis of various compounds, including peptides and oligonucleotides. However, these methods rely on distinct protocols due to the absence of standardized digital automation methods for computer control of chemical reactions. Currently, there is no universal programming language for computational control of chemical operations.

The Chemputer platform represents a significant advancement in this field. It offers a generalized standard by integrating codified standard recipes, or chemical codes, for molecular synthesis. Operated by the Chempiler program, the Chemputer accepts codified synthesis procedures from the Chemical Assembly (ChASM) scripting language and manages specific low-level instructions for the robotic platform's modules. ChASM employs a chemical descriptive language (XDL) to systematically compile all necessary information for a synthesis procedure. The physical modules and their connections are represented as a directed graph using an open-source markup language, GraphML. This allows the Chempiler to control robotic operations,

enabling users to execute chemical syntheses without manual reconfiguration. The system has been validated by successfully synthesizing three pharmaceutical compounds—diphenhydramine hydrochloride, rufinamide, and sildenafil—without human intervention, achieving product yields and purities comparable to or better than manual methods. This development marks a step towards fully automating bench-scale chemistry, enhancing reproducibility, safety, and accessibility of complex molecules.

- Automated Sampling of Reaction Space with AI

AI-driven synthesis robots can also explore unknown reaction spaces. Recently, Leroy Cronin and his team employed a synthesis robot to conduct reactions with random substrates, using a vector presentation of substrate selection as input for a Support Vector Machine (SVM) model. Automated reaction analysis using infrared (IR) and NMR spectroscopy enabled the model to classify substrate reactivity. This information was used to update the reaction database, and a Linear Discriminant Analysis (LDA) model was trained to predict the probability of remaining reactions. LDA identifies a linear combination of chemical features to determine reaction likelihood. This iterative approach accurately predicted the reactivity of approximately 1000 reaction combinations with over 80% accuracy based on real-time data from a limited number of experiments.

Further applying this 'self-driving' approach to Suzuki–Miyaura reactions, predicted reactive combinations were manually verified by chemists, resulting in the discovery of four previously unknown reactions. Comparison with millions of reactions showed that these new reactions had Tanimoto similarity scores in the top 10 percentile, indicating their uniqueness. This method represents a significant advancement in the digitization of chemistry, potentially enabling real-time exploration of chemical spaces and facilitating the discovery of new drug candidates in a more efficient and cost-effective manner.

#Summary and Conclusions

Summary:

Artificial Intelligence (AI) has made a substantial impact on drug discovery and development, offering transformative potential across various facets of the pharmaceutical industry. Key areas of AI application include drug discovery, drug repurposing, productivity enhancement, and clinical trials. The integration of AI has led to reduced manual workload and accelerated progress in pharmaceutical research.

Recent advancements in AI have improved the prediction of protein structures and interactions, essential for drug design. Deep learning techniques have significantly enhanced the prediction of protein folding and interactions, leading to better identification of druggable sites and new drug targets. In drug repurposing, AI facilitates the discovery of new uses for existing drugs, leveraging network analyses to identify potential new indications.

Virtual screening and activity scoring have also benefited from AI, with machine learning and deep learning methods improving the efficiency and accuracy of identifying bioactive molecules and predicting binding affinities. AI methods, including convolutional neural networks and other deep learning techniques, have advanced the evaluation of ADME/T (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties, crucial for drug development.

Conclusions:

1. AI Integration in Drug Discovery: AI has revolutionized drug discovery by enhancing the prediction of protein structures and interactions, streamlining virtual screening processes, and improving activity scoring. This integration accelerates drug development and increases precision in identifying therapeutic targets.

2. Drug Repurposing: AI supports drug repurposing efforts by uncovering new applications for existing drugs, thereby reducing development time and risk. Network-based analyses have proven effective in predicting new drug-disease interactions.

3. Virtual Screening and Activity Scoring: AI-driven virtual screening and scoring methods have shown significant improvements in identifying and evaluating potential drug candidates. Machine learning and deep learning approaches offer enhanced accuracy and efficiency over traditional methods.

4. Predictive Modeling for ADME/T: AI models have advanced the prediction of ADME/T properties, providing critical insights into drug behavior and safety profiles. Techniques like deep learning offer improved performance in predicting solubility, permeability, and toxicity.

5. Challenges and Future Directions: Despite the progress, challenges remain, including the need for high-quality data, computational resources, and interpretability of AI models. Future advancements will depend on overcoming these challenges and further integrating AI with drug discovery workflows to optimize pharmaceutical research and development.

Reference:

1. Debleena paul , Gaurav sanap , Snehal shenoy Kalyani, nyaneshwar Kaliya Kiran Kalia , and Rakesh k. Tekade, artificial intelligence in drug discovery and development (2020) 5135 96446 page number 01 (Abstract)

2. Debleena paul , Gaurav sanap , Snehal shenoy Kalyani, nyaneshwar Kaliya Kiran Kalia , and Rakesh k. Tekade, artificial intelligence in drug discovery and development (2020) 5135 96446 page number 01 (Introduction)

3. Feishsheng zhong, jing xing , xutong li, xiachhobg liu , zunyun fu , zhaoping xiong , dong lu , xiaolong wu , artificial intelligence interact designs (2018) 10.1007. page number 4 (Prediction of protein sequence)

4. Feishsheng zhong, jing xing , xutong li, xiachhobg liu , zunyun fu , zhaoping xiong , dong lu , xiaolong wu , artificial intelligence interact designs (2018) 10.1007. page number 9-10 (machine learning strategies and programs for drug design)

5. Feishsheng zhong, jing xing , xutong li, xiachhobg liu , zunyun fu , zhaoping xiong , dong lu , xiaolong wu , artificial intelligence interact designs (2018) 10.1007. page number 10 (transfer learning for low data)

6. Feishsheng zhong, jing xing , xutong li, xiachhobg liu , zunyun fu , zhaoping xiong , dong lu , xiaolong wu , artificial intelligence interact designs (2018) 10.1007. page number 4(prediction of protein protein intraction)

7. Feishsheng zhong, jing xing , xutong li, xiachhobg liu , zunyun fu , zhaoping xiong , dong lu , xiaolong wu , artificial intelligence interact designs (2018) 10.1007. page number 5 (Drug repurpose)

8. Feishsheng zhong, jing xing , xutong li, xiachhobg liu , zunyun fu , zhaoping xiong , dong lu , xiaolong wu , artificial intelligence interact designs (2018) 10.1007. page number 10 (the cross validation method)

9. HC Stephen chain, hannabin shin, thomani dahoun , horst Vogel , and shaguang yaun , advanced drug discovery by artificial intelligence (2019) 06.004 page number 3-4 (Primary and secondary screening by Ai)

10. HC Stephen chain, hannabin shin, thomani dahoun , horst Vogel , and shaguang yaun , advanced drug discovery by artificial intelligence (2019) 06.004 page number 7-9 (Planning chemical ssynthesis with ai)

11. HC Stephen chain, hannabin shin, thomani dahoun , horst Vogel , and shaguang yaun , advanced drug discovery by artificial intelligence (2019) 06.004 page number 9 (Automation of chemical synthesis with ai)

12. Jose Jimenez - Luna , francasca grisoni , nilsweskamp and gisbert Schneider , artificial intelligence in drug discovery : recent advances and future perspectives (2021) 1909567 page number 2-5 (QSAR)

13. Jose Jimenez - Luna , francasca grisoni , nilsweskamp and gisbert Schneider , artificial intelligence in drug discovery : recent advances and future perspectives (2021) 1909567 page number 6-7 (De no vo desing)

14. Jose Jimenez - Luna , francasca grisoni , nilsweskamp and gisbert Schneider , artificial intelligence in drug discovery : recent advances and future perspectives (2021) 1909567 page number 6-7 (Automated synthasis planning with Artificial intelligence