Predictive Modeling for ATME-TOX Properties of Drug Using Machine Learning: A Review

Anjali S. Patel[1], Kiran B. Thakor[2], Megha K. Patel[3]

Department of Computer Engineering, Sankalchand Patel University, Visnagar, Gujarat, India[1,2,3]

patelanjali122001@gmail.com[1], kbthakor.sspc@spu.ac.in[2], mkpatel.sspc@spu.ac.in[3]

Abstract: This survey paper comprehensively explores the landscape of predictive modeling for Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties of drugs through the lens of machine learning (ML) techniques. The review encompasses an extensive analysis of methodologies, data sets, advancements in ML algorithms, and their applications in drug discovery and development. Beginning with an overview of the significance of ADMET properties in drug development, the survey delves into various datasets utilized for modeling, encompassing chemical descriptors, biological activities, physicochemical properties, and toxicity endpoints. It scrutinizes the intricacies of feature engineering, emphasizing the importance of selecting informative features for accurate predictions. The survey critically evaluates an array of ML algorithms employed in predictive modeling, ranging from traditional methods to state-of-the-art deep learning architectures. It highlights the strengths, limitations, and applications of these algorithms in predicting ADMET properties, emphasizing the need for robust experimental design and validation protocols. Challenges such as interpretability, data quality, and integration of domain knowledge are addressed, underscoring the significance of standardized frameworks for ensuring reproducibility and generalizing ability of predictive models. Furthermore, the survey showcases successful applications of ML-based predictive modeling in optimizing drug candidate selection, mitigating toxicity risks, and expediting the drug discovery process.

Keywords: Health care, Machine Learning, Deep Learning, ADMET Properties

I. INTRODUCTION

In recent years, the application of ML techniques in predicting ADMET properties of drug candidates has emerged as a pivotal area in pharmaceutical sciences. The introduction provides an overview of the significance of ADMET properties in drug development, emphasizing the necessity of efficient prediction models to assess these properties early in the drug discovery process. It discusses the complexities and challenges associated with traditional experimental methods in assessing ADMET properties, underscoring the need for computational approaches that expedite the identification of potential drug candidates while reducing costs and laboratory efforts. Predictive modeling for ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties of drugs using machine learning (ML) involves employing computational techniques to forecast how a drug candidate might interact within a biological system. This approach has become integral in pharmaceutical research and development, aiding in the identification and optimization of potential drug candidates while minimizing risks associated with toxicity and inefficacy.

II. METHODOLOGY AND APPROACH

A. ***Objective:*** The primary goal is to predict and assess various crucial properties of a drug candidate:

1. **Absorption:** How the drug is absorbed into the body's blood stream from its administration route.
2. **Distribution:** How the drug spreads throughout t the body's tissues and organs.
3. **Metabolism:** How the drug is chemically altered within the body.
4. **Excretion:** How the body gets rid of the medication and its metabolites.
5. **Toxicity:** Assessing potential adverse effects the drug might induce.

B. ***Data Acquisition:*** Gathering comprehensive data from various sources (e.g., biological assays, chemical databases, research publications) that detail the properties and behaviors of different drug molecules.

https://doi.org/10.63766/spujstmr.24.000007

1. ***Data Preprocessing:*** the data is cleaned by handling missing values, normalizing features, and structuring it for analysis. This step ensures that the data is suitable for ML algorithms.
2. ***Feature Engineering*****:** Identifying relevant features or properties that impact ADMET behaviors. This may involve transforming existing features or creating new ones that enhance predictive power.
3. ***Model Selection*****:** Choosing suitable ML algorithm (e.g., RF, Gradient Boosting, and Neural Networks) based on the nature of the data and the prediction task.
4. ***Model Training:*** Training the selected models on a portion of the dataset to learn patterns and relationships between drug features and ADMET properties.
5. ***Model Evaluation*****:** Assessing model performance using metrics like accuracy, precision, recall, or area under the curve (AUC) for classification tasks, and metrics like RMSE or R-squared for regression tasks.
6. ***Model Validation and Interpretation:*** Ensuring the model generalizes well to new, unseen data. Analyzing feature importance can provide insights into which factors significantly influence the predictions.
7. ***Deployment and Monitoring:*** Deploying the model to predict ADMET properties of new drug candidates. Continuous monitoring and potential model updates as new data becomes available or as performance changes are essential.

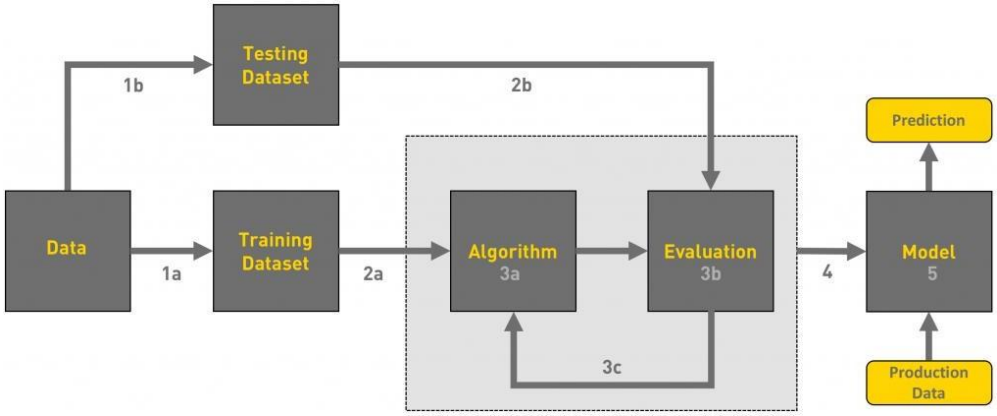

Fig. 1. Work flow of Machine Learning

C. Scope of the Survey:

The survey aims to comprehensively review and analyze various aspects of predictive modeling for ADMET properties using ML techniques. It delves into:

1. ***Data Sources and Preprocessing*****:** Discuss the diverse sources of data used for modeling ADMET properties, data quality challenges, and preprocessing techniques required for ML algorithms.
2. ***Feature Engineering and Selection:*** Highlighting strategies for identifying essential features and engineering approaches to enhance predictive power.
3. ***Machine Learning Models:*** Reviewing a spectrum of ML algorithms (e.g., RF, Neural Networks, and Support Vector Machines) employed in predictive modeling and their applications to specific ADMET properties.
4. ***Model Evaluation and Validation:*** Discussing metrics and techniques used to assess model performance and ensure generalizability and reliability.
5. ***Challenges and Future Directions:*** Addressing challenges such as interpretability, domain knowledge integration, and the need for more comprehensive datasets. Additionally, discussing emerging trends and future directions in the field.

D. Significance:

The introduction emphasizes the significance of this survey in consolidating current knowledge, providing a comprehensive understanding of methodologies, and identifying gaps and opportunities for future research in predictive modeling for ADMET properties using ML. It highlights the potential impact of these models in expediting drug discovery, optimizing candidate selection, and minimizing risks associated with toxicity and inefficacy.

By offering a comprehensive review of existing literature, methodologies, challenges, and potential advancements, this survey aims to serve as a roadmap for researchers, practitioners, and stakeholders in the pharmaceutical industry, fostering advancements in predictive modeling for ADMET properties using ML techniques.

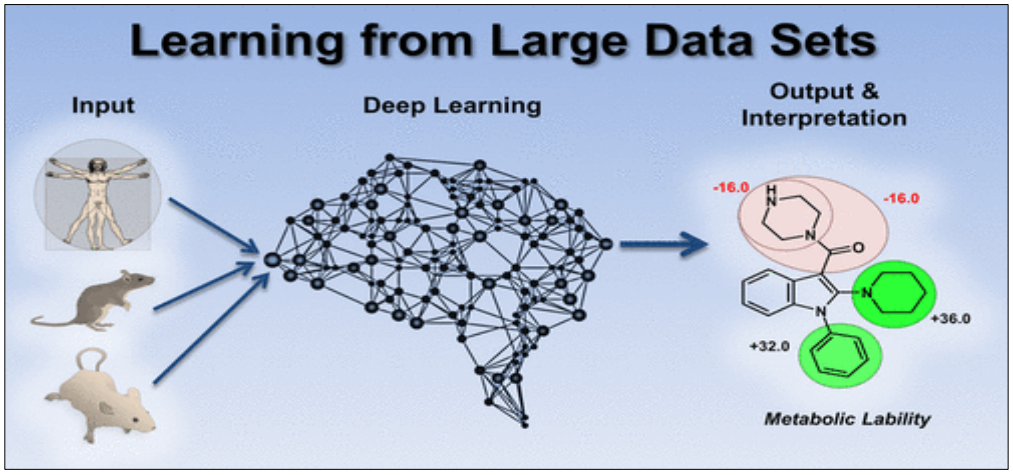

Fig. 2. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties

III. LITERATURE SURVEY

A. Using Tox21 and Machine Learning to Determine the Protein Features and Pathways That Cause Toxicity: Consequences for Predictive Toxicology [13]

For assessing drug toxicity, this is the first computational pipeline that uses protein descriptors to extract important information from twelve toxicity endpoints in the Tox21 dataset. Our strategy combines several protocols that are part of the CANDO drug discovery platform. Compound-proteome interaction signatures, data balance, feature selection, and enrichment analysis are all part of these techniques. Understanding chemical toxicity patterns at the protein pathway level is the goal of this coordinated study. We hope that this new computer pipeline will provide a new way to evaluate environmental chemicals. Additionally, it offers the pharmaceutical industry and researchers a chance to investigate the underlying proteome mechanisms that cause toxicity and may even help create new treatments that target pathways linked to toxicity.

B. ML in drug design: Investigating the link between chemical structure and biological function using artificial intelligence [14]

An extensive review of the application of artificial intelligence (AI) systems in drug design is provided in this work. One of the AI methods used to find chemical compounds with potential medical applications is a neural network. The comeback and long-term effects of AI in medicine were addressed in the 2019 Nature Machine Intelligence article. It validated the increasing contribution of computational techniques and computer developments to drug design. While several neural network topologies, such as CNN, capsule, or GAN, are used in drug creation, the review stressed that no single network is unquestionably the best technique. But deep learning (DL) solutions are becoming more and more well-liked because they can replicate complex human thought processes and independently determine design significantly.

C. A review on machine learning approaches and trends in drug discovery [15]

A collaborative effort is imperative to seek and implement standardized frameworks. This effort stands as a crucial factor in swiftly transitioning academic findings into industrial applications. Lack of standardization in processes and methodologies poses a challenge, preventing the extension of research outcomes to practical clinical tasks. Hence, when employing machine learning

techniques, it becomes essential to design experiments robustly to ensure reproducibility across diverse researchers. Throughout this review, inconsistencies in this aspect were evident across various articles analyzed. To arrive at definitive conclusions, addressing this issue deserves significant attention. Nonetheless, the potential and benefits offered by machine learning techniques remain vast, particularly within the realms of precision medicine and drug discovery.

IV. DATASET

Several datasets are available for predictive modeling of ADMET properties using machine learning in drug discovery. Some popular datasets include:

1. ***Tox21Dataset:*** This data set, developed by the National Institutes of Health (NIH), contains results from high-throughput screening assays measuring toxicity-related properties for thousands of chemical compounds.
2. ***Drug Bank:*** Drug Bank is a comprehensive database that includes information on drugs, their targets, chemical structures, and ADMET properties. It's a valuable resource for predictive modeling in drug discovery.
3. ***ChEMBL:*** ChEMBL is a large database that provides bioactivity data, including ADMET properties, for a wide range of compounds. It's frequently used in drug discovery research.
4. ***PDBbind:*** This dataset focuses on protein-ligand binding affinity and contains information about protein structures, ligands, and their binding affinities, which can be relevant for drug design and ADMET prediction.
5. ***PubChem BioAssay Database:*** PubChem offers an adverse collection of bioassay data, including ADMET-related assays, which can be used for modeling toxicity properties.
6. ***TCRD (Therapeutic Target Database*)**: TCRD provides information on drug targets, including protein interactions, pathways, and associations with diseases, which can be used in predictive modeling.

When using these datasets for predictive modeling, it's crucial to preprocess the data, handle missing values, perform feature engineering, and split the dataset into training and testing sets for model development and validation.

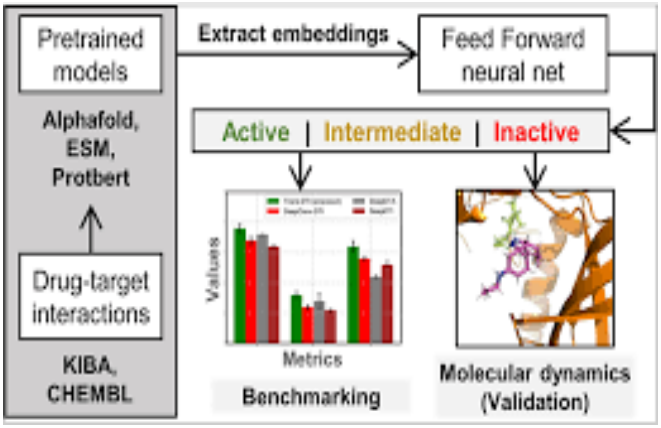

Fig.3. Transformer-Based Language Models for Estimating DTI and Building a Drug Recommendation Workflow [16]

A. *Attributes of Dataset*: Attributes in a dataset for predictive modeling of ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties of drugs using machine learning typically include various features describing chemical compounds and their biological properties. Here are common attributes or features found in such datasets:

1. ***Chemical Descriptors:*** Molecular descriptors representing structural features of compounds, such as molecular weight, chemical formulas, atom counts, bond types, etc.
2. ***Biological Activity:*** Information about the interaction of compounds with biological targets, including binding affinity, enzymatic activity, or cellular responses.
3. ***Physicochemical Properties:*** Properties like solubility, lipophilicity, polar surface area, and hydrogen bonding capacity influence a compound's behavior in biological systems.
4. ***Toxicity Endpoints:*** Measurements or predictions of toxicity-related properties, including cytotoxicity, mutagenicity, genotoxicity, carcinogenicity, hepatotoxicity, cardiotoxicity, etc.

5. ***ADMET Parameters*:** Attributes describing the Absorption, Distribution, Metabolism, Excretion, and Transport properties of compounds, such as bioavailability, permeability, metabolic stability, plasma protein binding, etc.
6. ***Biological Pathways/Targets:*** Information about biological pathways affected by the compounds, target proteins, gene expression changes, and pathway interactions.
7. ***Experimental Conditions:*** Conditions under which the data was collected, including concentrations, assay types, cell lines, organisms, and experimental protocols.
8. ***Metadata:*** Additional information like compound IDs, assay IDs, sources of data, assay descriptions, and any other relevant contextual information.
9. ***Outcome/Label:*** The target variable representing the toxicity or ADMET property being predicted or classified.

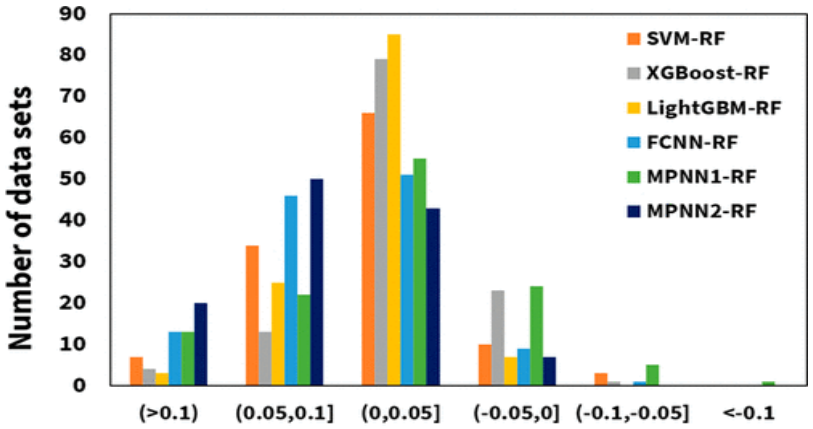

Fig. 4. Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion[17]

These attributes collectively provide a comprehensive profile of compounds and their behaviors, facilitating the development of predictive models to estimate ADMET properties and predict the potential toxicity or efficacy of drug candidates. The choice and relevance of attributes often depend on the specific research question, the nature of the compounds, and the goals of the predictive modeling task.

V. CONCLUSION

This survey comprehensively explores the landscape of predictive modeling for Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties of drugs using machine learning (ML) techniques. The review underscores the transformative potential of ML in revolutionizing drug discovery and development processes. Through an in-depth analysis of various datasets, methodologies, and advancements in ML algorithms, it is evident that predictive modeling holds promise in efficiently evaluating ADMET properties, aiding in the identification of drug candidates while mitigating risks associated with toxicity and inefficacy. The review highlights the significance of interdisciplinary collaboration, emphasizing the need for standardized frameworks and robust experimental designs to ensure the reproducibility and generalizability of predictive models. Challenges such as interpretability, data quality, and the integration of domain knowledge remain crucial areas for further exploration and refinement. Moreover, the versatility of ML algorithms, coupled with their ability to decipher complex biological interactions, opens avenues for precision medicine and targeted drug design. However, it's imperative to acknowledge that while ML techniques offer tremendous potential, they complement rather than replace human expertise in the decision-making process.

The future of predictive modeling in ADMET properties using ML appears promising, paving the way for accelerated drug discovery, enhanced drug safety, and personalized therapeutics. As this field continues to evolve, it is crucial to foster collaborative research efforts, leverage emerging technologies, and address existing challenges to realize the full potential of predictive modeling in optimizing drug development."This conclusion serves to summarize the key takeaways from the survey paper, emphasizing the opportunities, challenges, and future prospects in the realm of predictive modeling for ADMET properties using machine learning techniques.

REFERENCES

[1] Cheng, F., et al. "*Prediction of drug-target interactions and drug repositioning via network-based inference*". PLOS Computational Biology, 8(5), e1002503, 2012.

[2] Chen, H., et al. "*Deep learning in label-free cell classification"*, Genome Biology, 2018, 19(1), pp. 1-10.

[3] Mayr, A., et al. " *DeepTox: Toxicity prediction using deep learning*" Frontiers in Environmental Science, .2016, pp.80-85.

[4] Xu, Y., et al. "*Deep learning for drug- induced liver injury*" Journal of Chemical Information and Modeling, 2018 58(3), pp. 487-492.

[5] Wallach, I., & Heifets, A. "*Mostlig and-based classification benchmarks reward memorization rather than generalization*" Journal of Chemical Information and Modeling, 2018, 58(5), 916-932.

[6] Yang, K.,& Swanson, K."*Comparative analyses of deep learning and machine learning models for prediction of chemical toxicity*" Journal of Chemical Information and Modeling, 2018, 58(8), 1553-1562.

[7] Zhang, L.,Tan,J.,Han, D.,Zhu,H., & Fromm, M. F. (2016). " *Prediction of drug-drug interactions through drug structural similarities and interaction networks incorporating pharmaco kinetics and pharmaco dynamics knowledge*" Journal of Chemical Information and Modeling, 56(12), 2446-2456.

[8] Ekins, S., et al. (2015). "*Progress in predicting human ADME parameters in silico*" Journal of Pharmacological and Toxicological Methods, 71, 126-145.

[9] Fourches, D., et al. "*Chem informatics analysis of assertions mined from the literature that describe the drug-induced liver injury in different species*" Chemical Research in Toxicology, 2010, 23(1), 171-183.

[10] Hughes, J. D., et al. (2008). "*Principles of early drug discovery*", British Journal of Pharmacology, 152(1), 38-44.

[11] https://towards data science.com/ workflow-of-a-machine-learning-project-ec1dba419b94

[12] Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets]https://pubs.acs.org/doi/abs/10.1021/acs.jcim.8b00785

[13] https://www.mdpi.com/1420-3049/27/9/3021

[14] https://wires.onlinelibrary.wiley.com/doi/epdf/10.1002/wcms.1568

[15] https://www.sciencedirect.com/science/article/pii/S2001037021003421

[16] https://pubs.acs.org/doi/10.1021/acsomega.1c05203

[17] https://pubs.acs.org/doi/abs/10.1021/acs.jcim.3c00160