AI in Drug Discovery and Development

Prof. Rajnish Kumar
Dept. of Pharmaceutical Engineering and Technology
IIT-(BHU), Varanasi

Week-07
Lecture-34

Welcome to the course on AI indirect discovery and development. In today's session, we will talk about openly available resources for ADMET prediction, so by the end of this lecture, you will be able to identify key openly available tools. and databases for predicting ADMET properties recognize how open ADMET tools integrate with machine learning workflow platform Data pipelines in drug discovery and applying appropriate tools for early-stage compound screening and virtual filtering in research projects. So, until now we have seen that there are multiple methods to predict the ADMET properties and how ADMET properties play an important role in drug discovery and development. So, in this session, we will talk about the plenty of open-source tools that are being used for predicting the ADMET properties, and in this session, we will see some of them. And we will talk about what those tools are and how they are predicting those properties, and whether they are good or not.

So, before that, why do we need to have the openly available ADMET tools? So, the first thing is that you need to accelerate the drug screening, which uses AI for early-stage drug candidate selection. The cost reduction is another aspect because it reduces reliance on costly wet lab testing and increases transparency in reproducibility. So, the open-source algorithms are, you know, transparent in nature; they have easy validation and modification as well because anyone can use their codes. And work with them, and then there are, you know, community-driven improvements because these are being constantly updated, and they get support from the scientific community as well.

Accessibility is another advantage, as these tools are freely available to anyone, anywhere. And then leading to increased global collaboration enhances the collective model improvement. And then they also bridge the academy and industry, driving innovation and interactive development. So, before talking about the tools, let us see some data sets, and ChEMBL is, you know, one of the data sets that is being used extensively for making those, you know, predictive servers because, to make any prediction tool, we need to have the training data, right? So, if you wanted to, you know, predict the toxicity. So, we need the toxicity data of a large number of compounds to be used as a training and test set.

So, for that, ChEMBL is actually one of the sources. So, it is an open-access, manually curated database of bioactive compounds providing chemical bioactivity and target protein

data for drug discovery. So, you can see here that this is ChEMBL, and it is also integrated with other databases like PubChem. Which contains the confirmatory assays, and then you know it can contain information about the patents as well, along with the bioactivities, like from the Binding DB or under Explore Target. So, these are other databases that are integrated with it as well, and then the toxicity reference data set from, you know, Tox21, TP Search, Drug Matrix, or FDA.

This is also integrated into it, and then in the clinical research compounds and marketed drugs from drugs at FDA clinicaltrials.gov, and then about the, you know, The black box warning or withdrawals regarding those drug molecules, the therapeutic mechanism indications, and the molecular feature information. So, all these information is, you know, also get obtained from these databases and integrated into the ChEMBL. And then you have the deposited data sets like those from the DNDI, Estrogenica, GSK, all these, you know, proprietary data sets. So, they also have the ChEMBL has also integrated all those datasets into it Mainly, the data is coming from the scientific literature, including publications and patents related to medicinal chemistry, ADME, agrochemicals, reviews, etc.

So, all the data is, you know, integrated into these ChEMBL datasets. So, you can see how rich the ChEMBL database is, where you can find almost every sort of bioactivity for almost every target. And another database is PubChem. So, it is a chemical information repository consisting of three primary datasets: substance, compound, and bioassay. So, the primary database contains the substance that contains the collection of submitted chemical substances.

and then compounds, which are unique chemical structures with properties, and a bioassay that contains biological activity data for the compounds. And then you have the ADMET data, which provides toxicity, metabolism, pharmacokinetics, and safety profiles. So, it is used in drug discovery and helps in identification, lead identification, virtual screening, and risk assessment as openly available. This is freely accessible to academia, biotech, and pharmaceutical research, and it now also supports integration with AI, such as ML models for drug property prediction as well. So, these were the databases, and now let us see what those tools are that are being used extensively or that are openly accessible and used for ADMET prediction.

So, the ADMET predictor by the Swiss Institute of Bioinformatics is one of the popular tools called SwissADME by the Swiss Institute of Bioinformatics. It can predict ADMET properties, including, you know, the pharmacokinetic and drug-likeness, and add in the drug discovery. So, it is free to use and is widely adopted in academia, ADMET prediction, and drug discovery. So, you can see this is, you know, the landing page of the Swiss Admi

where you can draw the structure. Or you can even enter the files of the compounds, and then you can predict the ADMET properties of those molecules.

And then you have super toxic, which is a rich source of toxicological data combining structural, functional, and chemical information along with corresponding toxicity values. So, it is a database of around 60,000 toxic compounds and over 2 million toxicity measurements. So, it allows search by name, CASRN, toxicity values linked to PDB, UniProt, and CAC for target identification as well. It aids in toxicology prediction for drug candidates by providing experimentally validated toxicity data and compound interactions. And then you have ADMETLab 2.

0; now there is another updated version, ADMETLab 3.0, as well. So, it is a free web tool that predicts almost 88 ADMET endpoints using a multi-task graph attention framework that supports batch computation for drug discovery. So, it provides comprehensive ADMET property prediction by evaluating physicochemical, ADME, and toxicity profiles for drug discovery. Again, it is free for academic use and predicts multiple parameters that are related to efficacy, safety, metabolic stability, chemical stability, absorption, and solubility.

So, you can see that it covers all the parameters in the ADMET and is a very good tool to predict the you know ADMET properties. And then here, the ChemXTree is a novel graph-based deep learning model designed to enhance molecular property prediction in drug discovery. So, it is a kind of extensive evaluation on benchmark data sets, including MoleculeNet and eight additional drug databases, demonstrating ChemXTree's superior performance, surpassing or matching the current state-of-the-art models. So, the visualization techniques clearly demonstrate that chemistry significantly improves the separation between the substrate and the known substrate in the latent space for any of the target endpoints. And then you have the PKCSM, a free tool using graph-based signatures to predict ADMET properties, aiding drug development by assessing pharmacokinetics and toxicity.

So, it is an openly accessible open-source tool available for free to everyone to predict the ADMET properties and drug-like molecules. So, you can see here the landing page where, in step one, you can upload the files of your compounds. Or you can just upload the files in a text file or SMI file, or you can also write down these files in this book as well. And then you choose the parameters, like those from ADME, which parameters you want to predict, and then you can predict these properties. And then there is this ADMET print.

So, this tool allows for the prediction of selected physicochemical and ADMET properties of the compounds, such as cardiotoxicity, solubility, genotoxicity, membrane permeability,

and plasma protein binding, and the evaluation of the obtained outcomes using two interpretability approaches. So, it has added, you know, another advantage, I would say. So, that is the applicability domain and also the, you know, explainability of the models. So, it contains the LIME, the local interpretable model-agnostic explanation, so that you know which part of the molecule, for example, is responsible for determining if something is toxic in nature. So, which part of the molecule is showing that toxicity so that this information can be used in the lead optimization? And then it also uses the counterfactual explanation, which is based on the Exmol library.

And then there is, you know, the Chemexone PK prediction tool. So, it is a Marvin sketch by ChemAxon, which is a chemical drawing tool with the PK prediction plugin. So, it calculates PKA values considering protomers and visualize microspecies distribution aiding drug development. So, the ACD Labs PKA DB predicts PKA values using a fragment-based model with a database of over 26,000 compounds. And it visualizes ionizable centers, plots microspecies distribution, and allows customization with experimental data for accurate drug development insights, and then you have some other tools like the You know the open-source QSAR model that predicts the acidic and basic pKa values.

and then GR-pKa, which uses a neural network for efficient prediction of the pKa values, and BCL-XpKa, which is a deep learning-based classifier for local atomistic environments, MolGpKa, which is a graph convolutional model for web-based pKa prediction, and then you have the PypKa server, which combines physics-based and ML-based methods for large systems. And then we have the ProTox 3.0, which incorporates molecular similarity, fragment propensities, and most frequent features. and machine learning models based on a total of 61 models for the prediction of toxicity endpoints such as acute toxicity, organ toxicity, toxicological endpoints, molecular initiating events, metabolism, adverse outcome pathways, and toxicity targets.

So, Protox 3.0 is a free web tool for in silico toxicity prediction using ML models. It assesses toxicity endpoints like hepatotoxicity, carcinogenicity, and mutagenicity, aiding early drug development, and you can see, you know, the landing page of Protox 3.0, where you can draw the structure in this structure drawing tool, or you can upload the SMILES as well, or you can draw these SMILES, and then you can use it for predicting the toxicity endpoints. So, let us talk about this case study involving a group of authors. So, they studied they tested 18 different ADMET prediction tools on 24 FD approved tyrosine kinase inhibitors.

So, highlighting their accessibility for research, but variable accuracy. So, they concluded that using multiple tools is recommended for a comprehensive ADMET profile in drug

development. And then an example is predicting hepatotoxicity using PK, CSM, and ProTox-II. So, they used PK, CSM, and ProTox-II to predict the hepatotoxicity of a novel drug candidate. So, pKCSM analyzed metabolism-related toxicity, while ProTox-II classified compounds based on toxicity risks.

and the combined approach improved the early toxicity screening, aiding in the safer drug design. Then we have the DeepTox, which is an ML-based tool for toxicity prediction using deep neural networks to analyze chemical structures and assess potential toxic effects, aiding in early drug discovery. So, it uses deep learning to predict chemical toxicity, improving early risk assessment in drug discovery and utilizing a public data set for training deep learning models to predict chemical toxicity. So, you can see a representation of a toxicophore by hierarchically related features. So, simple features share chemical properties coded as reactive centers, and combining reactive centers leads to toxicophores that represent the specific toxicological effects.

So, for example, you can see here that we have the simple features, and then when they are combined, they lead to the reactive centers, which are important or responsible for the toxicity. And those are called toxicophores. And then we have DeepChem. So, DeepChem is an open-source AI library for deep learning in drug discovery, enabling molecular property prediction, virtual screening, and ADMET modeling. It is an open-source AI framework for chemical modeling, enabling deep learning-based molecular property prediction and virtual screening.

So, it supports one shot learning and graph neural network for efficient drug discovery as well. So, it enables custom ADMET model training using deep learning and molecular data as well. So, you can train your own model based on your own input data and training data by using the deep learning libraries. And then there is this DruMap, which is a drug metabolism and pharmacokinetic analysis platform developed by Professor Kenji Mizuguchi at Osaka University, Japan. So, it consists of a database of DMPK parameters and a program that can predict many DMPK parameters from the chemical structures of a compound.

So, the database contains data on the MPK parameters from curated public data and newly acquired experimental data obtained under unified conditions. Because one of the problems with the public data is that the data is not highly reliable in nature. So, what they did was collect the data themselves as well as in collaboration with the pharmaceutical companies. And then they use this; they put the data into this database, as well as they made this prediction server where anyone can predict the properties of those molecules.

So it contains data for over 2.4 million compounds, including both experimental and

predicted DMPK parameters. And this is the prediction server where you can draw the structure, or you can upload the SDA files, and then you can predict those properties and parameters by using this server. And then you have the ADMET AI. So, it is an ML-based platform that offers rapid and accurate predictions of ADMET properties. It is accessible both as a web interface and as a Python package, facilitating large-scale chemical library evaluation.

So, it uses GNN models and utilizes Chemprop-RDKit models trained on 41 datasets from the therapeutics data commons. And it has a very high performance, achieving the highest average rank on a TDC and ADMET benchmark group leaderboard. And then it is also showing a very high speed, recognized as one of the fastest web-based ADMET predictors, reducing prediction time by 45% compared to other web servers. Then you have the Helix ADMET, a robust and extensible ADMET prediction system that incorporates self-supervised learning to enhance prediction accuracy. So, you can see here it has these three steps.

So the task is divided into pre-training, final tuning, and using that finally tuned model for prediction. And then talking about some of the BBB prediction servers, BBB predictor is one of the web servers that can be used for predicting BBB permeability. So, you put your, you know, molecule either in the text format, like in the SMILES, or even in the SDF text format as well, and then you can predict the BBB permeability. And CbLigand-BBB pred is another online BBB probability prediction tool where you can, you know, draw the structure or upload those molecules into the server, and then you can choose the algorithm as well, like which algorithm you want to use, such as AdaBoost or SVM, and you can also specify which fingerprints you are going to use to predict the.

BPP probability of your molecules is. It has the possibility to use Max fingerprint, OpenBabel FP2, Molprint 2D, and PubChem fingerprints. And then we have the AdmetSAR. So, again, admetSAR is another web server that is extensively used for predicting ADMET properties. Basically, it is also based on quantitative structure-activity relationship modeling and predictive analytics, where the models that are already being developed are used to predict the properties of new molecules. And then you have the LightBBB, which is again a blood-brain barrier permeability predictor.

So, again, you can input the molecule structure in files, and you can upload the files as well. And then you have the ADMET lab as well as a similar platform which is being extensively used for predicting the properties, and you can see here that you can predict all these properties, such as the physicochemical properties, absorption, distribution, and metabolism. So, all these properties can be predicted. By using this model as well. And then you have BBP, the BBP predictor for blood-brain barrier peptides.

So, you know it can predict whether a peptide will close the blood-brain barrier or not. So, that is specifically for, you know, peptide sequences, actually. And then coming to another popular, you know, BBB probability predictor. So, it is integrated into the Swiss ADME, and it works on the basis of this boiled egg approach. So, you can see here, for example, if a molecule is falling between this yellow area and the yolk.

So, the molecule is expected to be BBB permeable, and if it falls outside this yolk yellow area, then the molecule will not be BBB permeable. So, coming to the comparison between those open-source and commercial ADMET tools. So, if we talk about the cost, these commercial tools are expensive and they are licensed. So, in this session, we did not talk about commercial tools, but there are plenty of commercial tools available for, you know, ADMET prediction as well. Like you have the Quick Pro from Schrodinger, and Gastro Plus is there, which is also a very popular tool for predicting pharmacokinetic parameters.

Comparing the costs, these commercial tools are expensive because you need to buy a license, while those open-source tools are free and open access, and talking about the accuracy. So, the commercial tools are highly accurate because they sometimes use their own proprietary data set to make those models; however, the open-source tools do not. So, their accuracy is moderate; sometimes it is good to have high accuracy. However, it actually depends on the data quality, specifically on what kind of data they have used.

For example, DruMap is a very good platform. So, which have used you know a well curated data set for making those models and then however whenever you are using the suggestion is that you compare like multiple those prediction servers. And then you try to validate them by using known, you know, permeable compounds. For example, you take one BBB-permeable drug. and one BBB impermeable drug and try to predict its BBB permeability on different servers and then go for the one that is giving you the accurate results for the for this labeled compound, for this known compound.

And then talk about the customization. So, these commercial models use advanced models, and the open-source models have some limited options; we cannot really customize them. And then the data source for commercial is proprietary. However, the open-source tools are basically developed on the public datasets available through, you know, the publications, research publications, or patent data. And then the user base for commercial ones is like the industry and pharma. However, the open-source tools are being extensively used in academia, and they are more popular in startups as well.

And talking about the ease of use, the commercial tools are complex and need training. However, the open-source tools are user-friendly, and these are usually web-based. And

ah, talking about scalability, the higher they are, the commercial tools are ah, they have a very high throughput. So, they can, you know, run on multiple, ah, nodes in clusters as well, and then they can screen hundreds of thousands of compounds very quickly. However, the open-source tools are often implemented as web servers, and you cannot screen thousands of compounds because of the limitations in computational capacity.

And then, talking about transparency, the open-source tools are, you know, the winners because they are using open algorithms, and everything is, you know, available to review and also to modify. However, those commercial tools have kind of black box models where you do not know how they were developed, and the data is also not available. Okay, tell me about the current limitations of ADMET prediction tools because, as I said, everything comes with a limitation. So, the general limitations of those tools are that the first one is data dependency because the accuracy of any model heavily depends on the quality and quantity of training data. The open-source tools have limited performance on rare or underrepresented chemical classes.

So, limited interpretability in many models makes them act as black boxes with limited mechanistic insight into the prediction, and it is hard to understand why a compound fails or possesses specific ADMET properties. So, in that case, we talked about that activity cliff as well, like how a very small change in a molecule leads to a very big change in the activity, or you know, all those endpoints like toxicity or permeability. So, that can also that is also you know leading to the limited interpretability of those models. And then, applicability domain issues, like predictions, may be unreliable for compounds outside the chemical space of the training set. So, like novel scaffolds or large biomolecules, they often yield poor results.

If you are building a model on 100 molecules, we cannot implement that. To predict the properties of millions of molecules, it has not covered, you know, that chemical space so that it can be applied to millions of molecules. So, for that, we need to do the applicability domain analysis to ensure that the model can be used for which kind of predictions. And then there are model-related limitations, like the risk of overfitting, where complex models may overfit to training data, reducing generalization to new molecules. And the endpoint specificity as well; some tools may not cover all relevant ADMET endpoints, like the transporter interaction and immunogenicity, and different tools provide varying levels of accuracy for different endpoints.

And then there are problems with the technical aspects and practicality, so there is a lack of experimental feedback. So, limited integration of experimental data for iterative model refinement and tool-specific biases means that different AI tools like ADMETLAB, pKCSM, and SwissADME often produce inconsistent results for the same molecule. So,

you take the same molecule and predict it using about 10 different ADMET servers, and you will probably get different results. So, that is, you know, a problem with the tool-specific bias and then regulatory acceptance. So, AI-based predictions are not yet fully accepted by regulatory agencies as standalone evidence.

So, the agency says that you should use it first, but then you have to verify that by using some wet lab methods. So, let's come to the advantages of ADMET prediction tools. So, let's talk about the advantages of both open source and commercial tools. So, the first advantage is that it is AI-powered prediction. So, both uses advanced ML DL algorithms for high throughput ADMET property prediction.

So, which can you know identify those minute features which cannot be identified by using simple methods and then support for the drug discovery pipelines. So, both can be integrated into virtual screening workflows and support the drug discovery pipelines. And then improved speed so they are faster than wet lab testing, enabling early-stage filtering as well as scalable predictions, so they can handle large data sets and compound libraries, and coming to the advantages of commercial tools. Sometimes they provide the regulatory-oriented outputs. So, they often designed to align with regulatory standards with validated models and comprehensive documentation and access to proprietary data sets.

So, these are trained on high-quality curated and often proprietary experimental data sets that are not available in the public domain. And the broad endpoint coverage includes rare or advanced ADMET endpoints like cardiotoxicity, immunotoxicity, and harsh inhibition, etc. In addition to professional technical support, they provide dedicated customer service, onboarding assistance, bug fixes, and regular updates. Integrated workflow suites are bundled with related tools like QSR modeling, metabolism simulators, docking, and PK modeling. Advanced visualization reporting provides detailed visual summaries, graphs, and exportable reports tailored for R&D decisions.

Coming to the summary, the open ADMET tools and databases provide cost-effective solutions for predicting drug properties, reducing experimental costs, and improving efficiency in academia and industry. The AI-driven models enhance prediction accuracy, fostering innovation in drug discovery. So, the key takeaways are that open-source tools offer free access to ADMET predictions, ML improves accuracy and efficiency, and the public databases support data-driven drug development. and combining computational experimental methods they ensure reliability and the future advancements will further integrate AI with large scale data sets. And I have given some of the papers that you can refer to in order to enhance your understanding in this area.

And in the end, I have an activity for you. So, you shall download the structures of any 5

CNS drugs, meaning the drugs that are CNS active. and predict their BBB probability using different pharmacokinetic prediction web servers and compare the results. So, you will probably see that all those different servers are giving different probability outcomes; just compare it with reality, and then you will see how those servers work. And with that, thank you.