AI in Drug Discovery and Development

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

Week-07
Lecture-33

Welcome to the course "AI in Drug Discovery and Development." In earlier sessions, we talked about how ADMET prediction is an important step in drug discovery and development. So, in today's session, we will talk about conventional methods for ADMET predictions. So, by the end of this lecture, you should be able to define rule-based and statistical ADMET prediction methods and explain Lipinski's, Veber's, and Egan's rules. Recognize structural alerts and toxicophores, identify the effects of logP, logD, pKa, and solubility on ADMET, and compare conventional and AI-based tools for ADMET prediction. So, if we talk about conventional methods in ADMET prediction.

So, the conventional method refers to established, widely used methods for evaluating the pharmacokinetic and toxicological properties of drug candidates. So, this can be broadly divided into experimental methods, in which you know many experimental techniques, such as the cell-based Caco-2 assay and the HepG2 assay. Either HeLa cells or enzyme-based assays, such as cytochrome P450 assays, aldehyde oxidase assays, or animal studies, are used for testing compounds in animals to observe possible adverse effects at the organ or system level. So, these are, you know, experimental methods by which we determine whether the molecules we are developing have optimal pharmacokinetic and toxicity properties.

And then there are, you know, computational methods, which are, for example, rule-based assays; one of the classical examples is Lipinski's Rule of Five. We try to look at some of the properties of molecules, and then, based on those properties, we make some filters; in fact, we create rules. So, we can say that if those molecules do not follow these rules, then they may not be good candidates for development. And then you have physiological descriptor-based methods, such as molecular weight and lipophilicity, because these properties usually correlate well with plasma protein binding. Or you know the permeability, and then you have statistical methods like linear methods, QSPR methods, or regression methods; you also have QSAR methods such as ADMET-QSAR models.

or metabolism prediction models that can predict which kinds of metabolic processes the molecules may undergo. So, why those conventional method they matter because of several advantages like the simplicity and speed. So, these are quick and easy to apply for the initial screening of large compound libraries without requiring complex infrastructure. So, if you wanted to screen for example, billions of molecules. So, at the first step we can use those rule based methods And then we can filter those billions of molecules, which are ultra-large libraries, without any time consumption; they can be filtered in a very short time.

And then, interpretability means that these rule-based decisions are transparent and intuitive, making them easy to understand, justify, and communicate. And then they are, you know, low-resource requirements because they do not require much computational power or infrastructure. So, there is no need for large data sets or high-performance computing, and that is why it is ideal for resource-limited environments or early-stage exploration. And then you have those who are effective in early filtration. So, efficiently eliminates unsuitable compounds before investing in costly computational experimental validation.

And then, regulatory familiarity is another advantage, as these established rules, like Lipinski's, are widely accepted by regulatory bodies and used in preclinical assessments. And then, above all, they are, you know, acting as a foundation for the AI models. So, those conventional descriptors form the input and benchmarking standards for training and validating AI/ML models—advanced models that are currently used to predict pharmacokinetic and toxicity parameters. So, now we will see them one by one. Let us see what those rule-based approaches for absorption prediction are, because absorption is the first step when a person takes a drug.

So, the first thing that happens is absorption. Then there are these six rules, which are like Lipinski's Rule of Five, Veber's Rule, Egan's Rule, the Ghosh Filter, Muegge's Rule, and the MDDR Rule. So, let us take a look at them one by one. The first one is Lipinski's Rule of Five. So, when Lipinski analyzed a large number of orally bioavailable drugs in his database, So, he found that most of those compounds followed these four rules, and these four rules were, you know, fewer than the other ones.

or equal to five hydrogen bond donors, less than or equal to ten hydrogen bond acceptors, a molecular weight of less than or equal to five hundred daltons, and a log P value of less than or equal to five. So, then he realized he found that if a molecule follows most of these, it follows at least three of them. Parameters that the compound is likely to be orally active; thus, he set them as core rules for oral bioavailability. So, he said that the compounds needing to be orally bioavailable had to follow these rules. So, the compounds that violate more than one rule are likely to have poor absorption.

Then came Veber's rule, which predicted oral bioavailability. So, it states that predicting oral bioavailability based on a 500-dalton molecular weight cutoff alone is not a strong predictor. Therefore, Weber proposed a rule based on two criteria: fewer than or equal to 10 rotatable bones and a polar surface area that is less than 10. or equal to 140 ohm·strom square, or a hydrogen bond donor plus a hydrogen bond acceptor value of less than or equal to. The compounds possessing these features have a high probability of good oral bioavailability, and then came Egan's Rule.

So, it states that for a molecule to be easily absorbed orally, it must possess certain properties. So, it should have a molecular weight in the range of 200 to 600 daltons and a LogP value between -0.4 and 5. The hydrogen bond donor is less than or equal to 6, and the hydrogen bond acceptor is less than or equal to 12. And then the ghose filter came.

So, we say that a drug is actually a drug-likeness filter. Which specifies that a molecule should exhibit a molecular weight in the range of 160 to 480 daltons. The log p value of minus 0.42 plus 5.6, and then the atom count included the molar reflectivity.

So, these two are the new inclusions because molecular weight and log P are what everyone is talking about. So, he included the atom count, which should be in the range of 20 to 70, and the molar reflectivity, which should be in the range of 40 to 130. And then came the MDDR rule, which was published by Tudor I. Opera for evaluating drug-like properties. So, he analyzed a library of drug-like compounds and then formulated these rules for a molecule to be drug-like.

So, it should have a number of rings less than or equal to 3, a number of rigid bones less than or equal to 18, and a number of rotatable bones less than or equal to 6. And then there is the Muegge's rule. So, it changed the properties' ranges and included other parameters to differentiate between drug-like and non-drug-like compounds. So, the drug-like compounds should follow these parameters, such as a molecular weight in the range of 200 to 600 daltons. Log p is in the range of minus 2 to 5; the number of rings is less than or equal to 7, the number of carbon atoms is greater than 4, the number of heteroatoms is greater than 1, and the number of hydrogen bond donors is less than or equal to 5.

Hydrogen bond acceptors of less than or equal to 10, the number of rotatable bonds of less than or equal to 15, and a polar surface area of less than or equal to 150 $Å^2$ are required. So, after absorption, the next step is distribution. There are also some rule-based distribution-prediction methods. So, the rule-based guidelines in distribution are largely based on molecular properties and substructural patterns. So, some of the key rules include BBB probability and blood-brain barrier probabilities.

So, if a drug needs to be CNS-active, it must cross the blood-brain barrier. And then it shall be available in free form in the brain to engage the target, and there is a prediction of plasma protein binding. So, if a molecule has very high plasma protein binding, it will have a very large volume of distribution. A very small amount of the drug is freely available to interact with the target, and we have seen this in the introduction to pharmacokinetics as well. So, there are the volume distribution estimates and CNS activity rules.

So, the BBB probability rule states that the LogP should typically be in the range of 2 to 4 because it is a moderate value of lipophilicity, which helps with BBB crossing. And then the polar surface area shall be less than or equal to 90 angstroms squared to be permeable to the blood-brain barrier, and hydrogen bond donors plus acceptors shall be less than or equal to 8. The molecular weight shall be between 400 and 450 daltons, and there is a prediction of plasma protein binding. So, it says that a high log P has a high propensity to bind to plasma proteins. So, it will show higher protein binding if there are more hydrogen bonds and a higher polar surface area.

So, that will indicate low plasma protein binding. So, on the basis of these rules, we can filter the molecules out, and then we can say which of those molecules has a high propensity to bind to the plasma proteins. And then there is an estimate of the volume of distribution for lipophilic compounds that have a high log P. So, they will have a higher volume of distribution because they will bind more to tissues and plasma proteins. And then, the hydrophilic compounds will have a low log P, specifically those compounds that have a low log P and a high polar surface area.

So, they will have a lower volume of distribution and will remain in the plasma in a free form. So, it means that they will have low-volume distribution and more of the drug will be available in free form to engage the target. And then there are these CNS activity rules; they are a kind of subset of the distribution rules. So, based on Egan's egg or boiled egg model for CNS penetration, which involves the plot of W log P—Wildman-Crippen's Log P versus polar surface area. Where in the plot is CNS penetration depicted in the yolk section? Passive GI absorption is depicted by the white portion of the egg, which indicates that a molecule must be sufficiently permeable to the blood-brain barrier.

So, the WLOP value shall be less than or equal to 5.88, and the polar surface area shall be less than or equal to 90 angstroms squared. Okay then, after the distribution, we will move on to metabolism. So, there are some rules by which you can determine either the site of metabolism or the site of metabolism predictions. So, to identify atoms or bones modified during biotransformation, several factors affect the sites of metabolism.

For example, electron density is one of the factors that affects the site of metabolism. So, if there is more electron density, then that molecule, you know, that site will mainly be metabolized. So, it has a greater chance of being metabolized, and steric accessibility is another factor, which means that there is greater steric hindrance. So, the molecule will be metabolized less because steric hindrance affects the activity of the metabolic enzymes.

And then there is the hydrogen-bonding capacity.

So, if there are more hydrogen bond acceptors or hydrogen bond donors, it will lead to a decrease in metabolic potential. So after the site of metabolism prediction, you can use structural alerts or functional group identification. So, which factors can determine which substructures are prone to metabolism? And then there are, you know, expert-derived biotransformation rules that can be used to predict the analyzed metabolic pathways, as stated in the literature. Mainly for phase 1, which is primarily carried out by the CYP450 enzymes, and the phase 2 reactions. And then, there are some rules for excretion predictions as well.

So, these physicochemical-based guidelines, which influence the excretion of drug molecules, possess the following characteristics: a log P of less than 2. If a log P is less than 2, it favors urinary excretion because the drug is less lipophilic. And then it is mainly soluble in water, so it favors urinary excretion if the molecular weight is less than 400 daltons. So, the molecule is very small, and it will undergo renal clearance if its molecular weight is less than 500 daltons. So, it will undergo biliary clearance because the molecule will not be able to pass through the glomerular filtration barrier.

So, it will preferably go through biliary excretion and then be excreted in the feces. And then, a higher polar surface area, which is greater than 120 angstroms squared, was observed. So, it decreases reabsorption and increases clearance, and a low LogD value at pH 7.4 promotes the faster elimination of both ionized and polar drugs.

So, they have high renal elimination. These are some of the rules that can predict the excretion of the molecules, meaning the route by which the molecules will be excreted. So, there are some rules for toxicity predictions. The toxicity study combines physicochemical thresholds with structural alerts or toxicophores, which are known toxic substructures. So, some of these rule-based features associated with toxicity comply with physicochemical thresholds. So, if the log P value is greater than 5, then the molecule will have a higher chance of bioaccumulation, leading to toxicity, especially if the molecular weight is greater than 500 daltons.

So, then again, there will be decreased clearance because these molecules will not be cleared through renal excretion. So, they will be excreted via biliary secretion, and then they will have decreased clearance. So, they will have high exposure, which means it will lead to high toxicity. There are some toxicophores, such as nitroepoxides and aromatic amines. So, these structures directly lead to toxic risks: a higher polar surface area, too many hydrogen bond donors, and too many hydrogen bond acceptors.

So, they affect the immune response, leading to tissue targeting and the formation of electrophilic groups. So, they lead to DNA-protein binding and can cause genotoxicity. So, looking at some of the toxicophores, for example, the aromatic amine in this drug, procainamide, we have an aniline group. So this, you know, causes agranulocytosis, which is a type of toxicity. And then those nitro groups, like, you know, in nitrofurantoin, are present.

So that leads to hepatotoxicity, and then you have hydrazine and isoniazid, both of which are hepatotoxic by nature. And then the quinones, for example, in doxorubicin, are ROS generators that lead to cardiotoxicity due to the formation of reactive oxygen species (ROS). Okay, then after those rule-based predictions, the next step is the physicochemical descriptors. So, there are some physicochemical descriptors that also affect the ADMET properties and predictions: log P, pKa, log D, and solubility. So, the log P, which is the partition coefficient, is a ratio of the concentrations of compounds in octanol to those in water.

It indicates whether a molecule is lipophilic or hydrophilic. So, if the log p is moderate, it means it is between 1 and 3. So, it favors passive membrane diffusion, and if the log P is high, it means it is greater than 5. So, it leaves it showing increased lipophilicity; thus, it will have increased toxicity and bioaccumulation. And if the load is very low, less than one, it will have poor permeability.

And if the permeability is poor, it can promote excretion. And the LogD, at pH 7.4, is the pH-dependent version of LogP, which accounts for both ionized and un-ionized forms. So, the effect on ADMET reflects realistic tissue distribution under physiological conditions. So, the drugs with LogD in the range of 1 to 3 are optimal for oral absorption, and if a drug has a low LogD, it will have increased renal excretion.

Talking about the pKa, which is the pH at which 50% of the molecules are ionized, is important. So, it has a lot of effect on ADMET prediction, as it determines the ionization states, which affect permeability since only the unionized form can cross the membranes under acidic conditions. So, those weak acids tend to be ionized, enhancing absorption, and under basic conditions, weak bases are more likely to be ionized, which improves absorption. And it also affects renal clearance through ion trapping. And then, solubility is another important parameter; it is, you know, the ability to dissolve in aqueous environments.

So, the effect it has on the ADMET prediction is that if the molecular weight is very high, it will have low solubility. Because solubility is indirectly related to bioavailability, if a drug is not soluble, the free drug will not be available for absorption. So, another parameter is the TPSA: total polar surface area. So, if a higher TPSA means better solubility, then the rule of thumb states that if the solubility is poor, there will be poor absorption, impacting bioavailability. So, there are statistical methods for ADMET prediction as well.

The statistical methods are used to determine patterns. Mathematical models are used with molecular descriptors, such as log P, molecular weight, hydrogen bond counts, and ADMET properties. And this we have seen in, you know, earlier sessions as well, where linear regression can be used to predict continuous properties like log P and clearance. So, these are, you know, simple and interpretable models where it can be seen that, for example, they find a correlation, and log P is really correlated with solubility, with a higher log P meaning lower solubility. And then, multiple linear regression models can also be used, where they model ADMET as a function of multiple descriptors, which can easily handle small datasets and provide straightforward analyses. And then you have the logistic regression, which classifies the outcomes, toxic versus non-toxic, and is used for binary toxicity predictions.

So we look at the workflow. So we first collect the molecular descriptors, such as size, shape, and polarity, and then use a training set of our drug with known ADMET data. Fit the data into a statistical model, such as regression or classification, and predict the ADMET properties of new or unknown molecules. And then, coming to QSAR or QSPR-based antimatter prediction, it is similar to what we saw on the previous slide. So, it uses a mathematical model to predict the properties of chemical compounds based on their molecular structures. So, QSAR, which we have already seen, is a quantitative structure-activity relationship that predicts biological activity, such as toxicity or enzyme inhibition, from chemical structures.

And QSPR is, you know, a quantitative structure–property relationship in which we can predict physicochemical properties, such as low solubility and pKa, from the chemical structure. So, this methodology we have already seen in earlier sessions as well, where we collect the compounds with known ADMET properties for the training set. Then we use the molecular descriptors, select the features that are most relevant, and build the model using machine learning or statistical methods. And then, during training and validation, we evaluate the models using metrics such as R-squared, RMSE, accuracy, AUC, etc. Okay, then comes the PBPK modeling, which is called physiologically based pharmacokinetic modeling and was first introduced by Thorell in 1937.

So, it estimates a compound's pharmacokinetic profile using ADME data, predicting target organ exposure by accounting for absorption, distribution, and local metabolism. So how it is done is that you define the objectives and scope, and then you assemble the physiological backbone, where you collect the data, including organ tissue volumes, blood flow rates, and tissue composition. And then you gather the drug-specific inputs, which include physicochemical data such as log P, distribution factors, binding, and elimination parameters. And then, mathematical calculations are performed that involve equations to evaluate absorption and elimination kinetics. And then parameterization is done for physiological and drug parameters; IVIVE is performed for extrapolation, and in vitro to in vivo extrapolation is conducted to extrapolate to the whole-organ level study.

And then calibration is done, which involves the comparison of in vitro and in vivo PK profiles. Finally, verification, validation, uncertainty analysis, and model refinement are conducted. So, after those rule-based methods, let us see what the AI-based ADMET prediction methods are. So, with the current advancements AI based models utilizing ML or DL techniques have provided the platform for rapid and accurate screening. Many AI models have been developed for drug-like properties in terms of optimal ADMET characteristics.

Some of those tools, such as ADMET AI, DeepChem, and HitGen, are useful. So, there are plenty of them; actually, there are thousands of them—no, not thousands, but hundreds of them. So, which one can we actually use? And those are basically based on utilizing machine learning or deep learning algorithms to find correlations between the available data and the labeled data to predict the ADMET properties. So, we talked about whether we should compare conventional methods to AI-based methods. So, in methodology so, the conventional methods they use predefined rules heuristics or SAR QSAR and heavily expert driven.

However, those AI-based methods use MLDL models that are trained on large datasets to learn patterns. So, the data input in conventional methods relies on known physiological structures, physicochemical features, and curated rules, while in AI-based methods, it ingests large, diverse data sets that can come from ChEMBL, Tox21, ZINC, omics, or text, and the data needs are lower in the case of rule-based methods. However, you need a huge amount of data for AI-based methods because performance improves with the size and quality of the annotated data. Talking about speed and scalability, rule-based methods are fast for known scaffolds, but they have poor scalability, and manual updates are required. However, those AI-based methods are high-throughput; you can screen millions of compounds efficiently and rapidly, and the model generalizes and scales effectively as well.

The accuracy of the rule-based methods, or conventional methods, is moderately limited by rule coverage and descriptor usage. However, the accuracy of the AI-based tools is high because they also capture nonlinear and hidden relationships. The rule-based methods are, you know, limited to descriptors such as log P, molecular weight, and hydrogen bonding.

So, these are the kinds of simple ADMET filters. However, the AI-based methods predict activity toxicity or ADMET property interactions, and they can also be used for de novo drug design.

So, the interpretability of rule-based methods or conventional methods is very high because they are, you know, transparent by nature. However, AI-based methods have low transparency because these AI models are sometimes considered black boxes. However, they require some additional explanatory tools. Regarding generalization, these conventional methods are inadequate for novel scaffolds and mechanisms. However, the generalizability of these AI-based methods is good, especially with transfer learning or pre-trained models, but it depends on biases and limitations.

So, conventional methods are biased toward known chemistry and are not suitable for innovation. The AI-based methods are sensitive to training data bias but require validation and careful curation; if the input data are not good, the outcome will not be reliable. And then, regarding utilization, conventional methods are used for early filtration. SAR and regulatory checks, along with AI-based methods, are used for hit-to-lead, ADMET prediction, target identification, virtual screening, and de novo design.

In clinical trials, we cover all the materials in this course. So, looking at the examples, you have the Lipinski, Ghosh, Veber, and rule-based QSR toxicity filters. However, there are all those methods, such as SVMs, random forests, CNNs, DNNs, and GANs. And then, human involvement is very high in conventional methods. However, in AI-based methods, it is often moderate and sometimes low as well because those models can automatically learn features and make predictions. And then, cost-wise, those conventional methods have a low computational cost, but the AI-based methods are computationally intensive; they require a lot of computational power, especially to train the model on a very large dataset.

And then, regarding adaptability, the adaptability of these conventional methods is low; they require rule revisions for new endpoint source scaffolds. However, those AI-based models have very high adaptability, and they can be quickly retrained or adapted to new tasks. So, after this comparison, let us summarize the session. So, beyond rule based approaches classical QSAR physicochemical models and expert systems are widely used. So, these models they rely on established descriptors mathematical models and curated knowledge.

They offer interpretable results and are often accepted by regulatory bodies. However, they may struggle with novel chemical spaces and lack adaptability, and the AI-based tools complement this by enabling high-throughput, data-driven predictions. So, I have some suggestions for further reading that you can go through if you want to learn more about this topic. And in the end, I have an open question for you regarding the race from lab to clinic: Is ADMET the real bottleneck? So just ponder it, and with that, I thank you.