AI in Drug Discovery and Development

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

Week-07
Lecture-35

Welcome to the course, AI in Drug Discovery and Development. In earlier sessions, we talked about ADMET prediction, where we discussed a lot about how we can use different ML modeling techniques and AI-based approaches for predicting the ADMET properties. And that is very crucial because, as we have seen, the reason for the failure of many drugs is a lack of efficacy and safety. So, if we can predict those molecules beforehand, we will be able to reduce the attrition rate in the clinical trials as well as in the development stage. I will introduce you to the DruMAP platform, which is developed by Professor Kenji Mizuguchi at Osaka University, Japan. So, today we will see what the DruMAP platform is and how we can use it for predicting the ADMET properties.

What you have to do is follow this link. If you go to this link, just open it in your browser; it can be any browser, and it does not need to be Chrome, Safari, or Internet Explorer. So, just open it. And then, we will see how it works.

So, once you open this link. So, what you will see on your browser is this landing page of the DruMAP platform. So, now you can see that this is version 2.0. So, the full name is the drug metabolism and pharmacokinetic analysis platform.

So, DruMAP is a drug metabolism pharmacokinetic analysis platform that consists of a database for DMPK parameters and a program that can predict DMPK parameters from the chemical structures of a compound. So, the DruMAP database contains data on DMPK parameters from curated public data and newly acquired experimental data obtained under unified conditions. It also contains predicted data from our prediction programs. Users can simultaneously predict several DMPK parameters for novel compounds. So, when you land on this page, what you will see here is the button for the compound search, the activity search, or the new prediction.

So, a little bit of detail is given here, such as how many compounds are in this database. So, you can see all the registered compounds. So, there are around two and a half million compounds, and then you have the free base compounds, ignoring the studio structures. And then this is the publication where they published details about this platform. And then if you are using this tool, you can actually cite this publication.

So, let us see what these compound searches do. So, if you click on the compound search, the idea is that this is already a database, actually. So, they have curated millions of compounds with their associated data. So, if we click on compound search, this compound search page will open. And then you can see that this will actually look like this.

So, here you can search a compound of your interest. For example, the idea of ADMET prediction is that you can predict its properties. However, if there are already existing experimental properties available, we can get access to that as well. So, for that, you can search for them with the structure. Or you can search for them by the activity, the property, the name, or the ID.

So, I can say, for example, if I wanted to search for a molecule, we can take the example of a molecule called aniline. So, if I click on the structure, I have to select either to search for it according to the substructure or the similarity. So, if I wanted to search for the exact compound, we shall go for the similarity. So, then I select the similarity, and then when I select only the similarity. So, what it says is that it will ask for the parameters where it will ask for alpha and beta.

So, where alpha is equal to beta, which is equal to 1. So, that will search using the Tanimoto coefficient if alpha and beta are equal to 0.5. So, then it will use the Sorensen-Dice coefficient. So, here we are using the Tanimoto coefficient to search for the compound according to the similarity.

And then we can either load the structure, load the data from this file, or we can draw the structure. So, let us draw the structure. So, when you click on "Draw Structure." So, what you will see here is this JSME editor where you can draw the structure. So, let me draw a molecule that is actually aniline.

So, now I have drawn this molecule called aniline, and then I will apply it, okay? So, now you can see that that structure is directly converted into the SMILES, okay, because the SMILES is being used to calculate the Tanimoto similarity coefficient with the database compounds, okay. And then we can go for the search, and before that, you can also add other rules if you want to add some more, you know, if or and, you know, boolean operators, so you can add other rules as well. So, when I click on search, let us see. So, it says that there is an invalid threshold value. So, because we have to define what threshold we want to give it for the similarity search.

So, for example, I say that I want a molecule that is exactly similar. Then I have to give a threshold of 1, actually. So, if I give a threshold of 1 and then search, let us see. So, if I wanted to search for a molecule that is exactly the same as the query. So, I have to give a

higher threshold value, and you know the tany motor similarity varies from 0 to 1.

So, 1 means exactly the same and 0 means no similarity at all. So, I can give a similarity value of 0.95 here. So, let us see if I search it, what you will find is. Yeah, so now you can see that we have got the molecule aniline.

This is now exactly the same molecule as our query structure, and then if I click on this molecule, I will get all the details of this molecule. So here you can see the DruMAP ID is this one, and then the compound name is aniline; the Japanese name is this, and the formula molecular weight. Free base molecular weight and then the exact mass, okay. And then you have the structural details like smiles, InChI, and InChIKey, and then you have the chemical space details where you can plot, you know, using different parameters. Like I wanted to make a plot between the free base molecular weight and log P.

So, this is the overall upload of the compounds, and then you can see here that the red dot represents the annealing action. So, if we can, we can change it to some other properties as well, like we can plot it against the log P and PSA as well. So, you can see here that we have plotted it against the log P, with the log P on the x-axis. On the y-axis, we have the polar surface area, and you can see that our compound, the aniline, lies here. and then you have all the physiological properties as well LogP HBA HBD PSA rotatable bond count aromatic ring, heteroatom, quantitative estimate of drug likeness, HBA Lipinski, HBD Lipinski And then you have the activities as well, in vitro activities, and that activity is taken from the ChEMBL database.

So, here is a kind of integrated data set where it takes information or data from other databases as well, and then we have the activities prediction. These are all predicted activities of this molecule using the machine learning or AI-based models they have actually developed. So, you can see here it has several models like intrinsic clearance class, CIP metabolism probability for CIP 1A2, CIP 2C9, CIP 2D6, and CIP 3A4. So, you have these metabolic liabilities using these enzymes, and then you have the solubility at pH 7.4, and then you have the, you know, the FA class, PAPP, apparent plasma concentration, and then you have the fu.

p class, fu.p regression model, and fu.p class and clearance, and then you have all these models actually. And then you have, it can also give you an estimate of which of these sites are actually metabolically liable. So, it can also provide you with the CYP P450 metabolic sites as well. So, it will highlight the structures where this molecule can be metabolized using the CYP enzymes.

So, this is about the database. Now let us go back home, and then we will go to the

prediction, actually. So, that was about the compound search. So, if we can search, we wanted to predict these ADMET properties of a new compound. So, what we can do is always first search the database to see whether our compound already exists in the database, and if it does not exist there, then we can make a new prediction.

So, let us click on the new prediction. When we click on the prediction, we will see that we will land on this page. So, on this prediction page, what we will see is that we will have the possibility to draw a structure like we did for the compound search. So, what we will do here is we can just again draw the structure of aniline here. This structure drawing tool is an editor, so now we have added the structure of aniline, and then we can add it to the list.

Okay, once we add it to the list, you will see compound one. And then you will see the smiles of the compound, and then you can also add some parameters like apKa, bpKa, and LogP as well. So, once we have added it, the next step is to select the models, actually. So, you can see they have developed plenty of models for prediction of different ADMET properties, and then you can see that the first one is the solubility at pH 7. So, this is a kind of regression model, where you will get the output as values.

And then if you look at the statistics of this developed model, you can see the MSE is 0.533. The R square value is 0.651, and the training set they used is from the ChEMBL database, which contains around 22,000 compounds. They used the GCN method and the GIN convolution method.

And then this was updated in February 2025. So, you can see. So, this platform is regularly updated, and they are trying to add even more parameters as well as more training data points. So, this is for solubility, and then you have solubility pH at pH 6, solubility at pH 1; again, these are regression modeling, and then you have fu, the fraction of drug unbound in plasma. And then this is again a regression model, which means you will get a.

You will get a value in the end, actually, and then you have fu.P Again, this was for humans; you have got it for rats as well, and then you have the fu.P for mice as well, okay? And then you have FU brain for mammals. And then that is again a regression model, and then we have the statistic R MSE equal to 1.

48. An R-squared value of 0.58; the model that they have developed is based on the gradient boosting technique. And then what they have used here is the Mordred and J compound mapper for calculating the descriptors of the compounds in the training data set. And then they updated it in 2021, and you can find the older version as well. Likewise, they have the model for CYP probability for humans, and this is actually giving you the probability for

different enzymes like CYP1A2, CYP2C9, CYP2D6, and CYP3A4; they have provided the accuracy for different models, okay. And here they are using the random forest model, and the molecular descriptors they have used are calculated using MordRed and RDKit.

And then you have PAPP and the PGPNER; this is typically relevant to the blood-brain barrier probability because P-glycoprotein is one of the proteins responsible for expelling molecules from the brain. So, here you can see that it is a classification model that provides a three-class classification, giving you a low, medium, and high probability of a molecule having this kind of endpoint. And then in this case, the statistic says that they have a kappa value of 0.45, and again they have used a gradient boosting method here, using the descriptors from mood red and j compound mapper. Okay, and then they have models for clearance as well.

Okay, and then they have a model for the KP brain as well. For example, this KP Brain AH model is developed again using the gradient boost method and utilizing ModRad, JCompoundMap, and CAM ChemAxon descriptors. Okay, so the idea is once you draw the structure, I have drawn the structure here, so the next step is to add it to the list. And once I have added it to the list, what I will do is select the properties that I want to calculate, okay? And then we click on the calculate button. So now, when I click on the calculate button, what this program is doing is that this server is taking the structure we have given to it.

And then, if you remember, when we talked about building this QSR model. So, we were using the training dataset, okay? And then we were converting the training dataset into the descriptors, right? So, those descriptors were like independent variables, and then we had the activity; in this case, maybe we can say solubility. So, solubility is our dependent variable. So, which depends on the features actually and the features we have calculated using molecular descriptors. So, we convert those molecules into molecular descriptors, and then we have the activity data, which is a dependent variable.

And now we see a correlation between the dependent variable and the independent variable, which are features here. And to create a model, we can use different ML techniques that we discussed earlier. So, now what this tool is doing is this web server is doing is. So, they have developed all those models, and all those models are at the back end, actually. Now, this web app is taking the input structure and converting it into the features.

And then those features are the converted features for these new molecules; it is putting all those features into those models, actually, and then it is predicting the property. Okay, so now you can see here. Now it has predicted all those properties. So, you can see the

parameter, and then the organism or cell, output, unit, and the compound. So, we use compound 1 here, and for solubility, it is giving a value of 47,711.

776. For solubility at pH 6 and pH 1, you can see that the values are given. And then fu.P of humans, so all these are actually regression models. So, it is giving all those values for properties, these parameters you know. So, you can see for clearance, it is giving the values in the unit of liters per liter per kilogram.

For different parameters, you will see all these values and properties, okay. All four of these different parameters. So, after the predictions, we can see that all these parameters have been predicted with different values. So, this is the idea of how you can use this tool to predict all these parameters, and this tool is highly rigorous. So, they have curated the data very carefully, and they have even generated data by themselves in the lab using a uniform assay as well.

So, that is, you know, the biggest advantage of this tool, and if you go back. So, then you can see we used the structure drawing tool to draw the structure and then predicted. Otherwise, what you can do is load the SDF files or load the SMILES files. If I click on load SDF, Then, if I have downloaded any SDF, you can see that I had one SDF file, which I downloaded from PubChem, for example, or you can even draw your structures in any structure drawing tool.

And then save them as SDF, and then you can use them. So, I downloaded this structure from PubChem, and then you can import that structure. And then you can, you know, use it directly as well, or what you can do is you can even. Even take the smiles as well, so you can click on the load smiles and then import the smiles you can see, like compound one. And then you have the smile structure, and you can use it to calculate the parameters.

This web server is highly flexible and highly useful. So, there are other platforms as well, like you might have heard about SwissADME. So, there are other tools as well that can be used, not only SwissADME; there are other, you know, ADMET predicting tools as well. But the good thing about this tool is that the models they have developed are highly reliable, and the data they have used is also highly reliable for predicting the ADMET properties. So, they are continuously developing it, evolving it, and maybe in the future they will also keep adding new parameters that are mostly relevant to preclinical drug discovery and development. So, you can see again that it is, you know, predicting the properties, and out of 14 properties, it has already predicted 10 properties.

And then soon we will see the outcome from this prediction. So, in this case, what we did this time was use the smile structure. And again, you can also draw the structures in any

tool and then convert them into SMILES or SDF, and then you can use those SDF files or SMILES. For predicting your molecules' properties, you do not need to do it one by one. So, you can also add multiple molecules at a time. So, you can add multiple molecules into a SMILES file or, you know, an SDF file, and then you can import them and work on those molecules.

So now the prediction is complete. So, you can see, this is our compound, and then you can see these solubility values. So, this is furosemide; actually, furosemide is a diuretic that is used to treat several diseases, one of which is hypertension. So, at pH 6, pH 7, pH 1, and all those parameters you know, you can see that. This is how we can use the DruMAP platform. So, if you have any questions or queries, we can discuss those, and with that, thank you.