AI in Drug Discovery and Development

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

Week-03
Lecture-11

Welcome to the course "AI in Drug Discovery and Development." In today's session, we will have a look at the machine learning concepts. So, by the end of this lecture, candidates will be able to understand the basic concepts of machine learning, including features, classification, regression, and model training. Recognize different ML models like KNN, decision trees, SVM, linear or logistic regression, and their uses in drug discovery. Also, learn the working principle of ML models and their use in the drug discovery and development pipeline. So, let us start with why we need machine learning for drug discovery.

So, in earlier sessions, we have also discussed that traditional drug discovery is highly challenging; it is really slow, taking 10 to 15 years, and the cost is very high. So, it takes approximately 2.5 billion US dollars per drug, which is, you know, made available in the clinic for the treatment of diseases. And then it also suffers from a very high failure rate of around 90 percent of those drugs that are in clinical trials, which fail.

And the major reason was the lack of efficacy and safety that we have discussed in earlier sessions as well. So, how ML can help in drug discovery is that it can assist us with hit identification, where we can screen millions and even billions of compounds virtually using various AI-based tools like DeepDock or virtual flow. And then it can help us with lead optimization, where it can predict the molecules' activity, ADMET properties like absorption, distribution, metabolism, excretion, toxicity profiles, etc. And then it can help us with drug repurposing, where it can find a new use for an existing drug. And that we have seen in the case of the treatment for COVID-19, where Remdesivir was, you know, discovered and repurposed as a drug for treating COVID-19.

And then it can help us in discovering personalized medicines by predicting the patient-specific drug responses correctly. So, we can see that overall ML helps a lot and basically it is increasing the speed, reducing the cost, and decreasing the time taken for a drug to reach the clinic from the lab. So, coming to the machine learning. So what it actually is, we have already seen it earlier as well. So just a kind of recap that machine learning is a subset of AI where computers learn from data without being explicitly programmed.

So you don't need to explicitly tell the computer what to do with the data, how to process it, and how to analyze it. So, it automatically analyzes the data on its own and can get the outcome. It can actually come to a result. So, how does it do it? So, it learns the patterns from the data to make predictions and decisions. There are three types of machine learning approaches, such as supervised learning, where the model learns from labeled data.

So, an example could be predicting whether a molecule is active or inactive. Unsupervised learning

is where the model finds hidden patterns in unlabeled data. And the data that is not labeled actually means the algorithm does not know; we are not letting the algorithm know which of the compounds is active, which of the compounds is inactive, or which is soluble or insoluble. So, it will just look at the pattern, look at the features, and then it can identify the hidden features in that unlabeled data. And so it can either cluster them, for example, based on their chemical similarity or based on their features; it can mix different clusters, and that is known as unsupervised learning.

And then we had reinforcement learning as well, where we saw that the model learns, you know, through reward and punishment. So, that is trial and error, actually. So, an example in drug discovery could be the optimization of synthetic pathways. So, when we are making a machine learning model, as I said, we need to convert those molecules into features, whether it is a small molecule or a biomolecule. So, a feature is an individual measurable property or characteristic of the data.

So, it could be a descriptor for a molecule, and in drug discovery, we have, you know, different kinds of features. For example, we have the molecular weight, which tells us how heavy the molecule is. And then we also use the logP, which is the partition coefficient that indicates the lipophilicity of the compound, showing whether the compound is soluble in fat or in water. And then another feature is the hydrogen bond donor or acceptor, which is important for binding interactions. Because whenever a drug goes into the body and engages with the target, it interacts with the target through all those biomolecular interactions.

And then another could be the topological polar surface area, which also tells us about cell permeability and the molecular fingerprint. So, these are encoded bit strings which represent the chemical structure. So, the better the feature, the smarter the model will be. So, that is, you know, a general consideration. An example could be that if we wanted to predict the blood-brain barrier probability of any compound or molecule.

So, features like LogP, TPSA, and rotatable bonds help us determine if a compound can cross the BBB or not because this property depends on these features of the molecules. So, then the features could be of different kinds. So, the first type of feature is called a discrete feature. So, these are categorical data that break down into final values. So, these are features that take on distinct, separate values; usually, these are categories or counts, and they are often encoded as integers like 0, 1, 2, or they can be one-hot encoded as well for the ML models.

So, some examples of these discrete features or categorical features are the drug class. For example, we assign 0 to the antibiotic, 1 to the antiviral, and then 2 to the anti-cancer. And likewise, these can be the chemical substructures as well, like the presence of a benzene ring; if it is there, then we say yes, and we encode it with 1. And if there is no benzene ring, then the absence will be represented by 0. And then the toxicity classes, like non-toxic, can be assigned 0; low toxicity 1; high toxicity 2.

And likewise, we can classify or featurize those amino acid types as well, such as hydrophobic, polar, or charged, and these can be assigned integer values like 0, 1, and 2. And then we have the

binary features. So those binary features have only two possible values, typically either 0 or 1, or yes or no, or true or false. So, what they do is they encode either presence or absence, any property or any substructure or any activity, or they represent the positive or negative states. So, they often represent the qualitative data in a simplified form and also help models make quick binary decisions.

So, some of the examples of these binary features are Lipinski's rule of 5 compliance if the molecule complies with Lipinski's rule of 5. So, then this will be we will assign it 1, and then otherwise, if it is not complying, we will assign the value of 0 to this feature. And then, flag the toxicity: if it is toxic, use 1; if non-toxic, use 0. Likewise, hydrogen bond donor presence: yes/no; molecular scaffold presence (aromatic ring): yes/no; activity classification: active/inactive (1/0). So, it is basically the same as the discrete features.

And then the next feature type is a continuous feature, which we also call numerical features. So, these are features that can take any value within a range. So, they are measured on a continuous scale. So, you can actually measure their quantity. So, the characteristics are that they can be infinitely divided into smaller values like 5.3, 5.31, 5.314. And they often require scaling or normalization, like we need to normalize them using the min-max value. So, some examples of these continuous features are molecular weight; for example, a molecule can have a molecular weight of 342.3 grams per mole. And logP, which represents lipophilicity, a value of 2.5 for logP indicates hydrophobicity.

And then we have the TPSA where 78.9 angstroms is a kind of continuous value, a numerical value that tells us about the polar surface area and can help us predict the membrane permeability. And then we have the IC50 value, which is the measured value of bioactivity. So, an IC50 value of 12.7 nanomolar measures drug potency, and if it is in nanomolar, depending upon the target, it can be a potent compound.

And then we can also have derived features, which are features that are calculated or engineered from existing data to capture more insight. So this can be continuous or categorical, and they help to enhance the predictive power; they also involve domain knowledge for meaningful transformations. So, some of the examples are ligand efficiency, where the ligand efficiency is calculated by pIC_{50} divided by molecular weight. So it binds, it measures the binding efficiency, how efficient a molecule is in binding to the target and showing the bioactivity. And then we can have the logD, which is a distribution coefficient derived from logP and pKa for drug permeability prediction.

We have the hydrophobic surface area, which is calculated from the molecular structure. We have the drug likeness score, which is a composite of multiple descriptors like molecular weight, logP, hydrogen bond donors, etc. And the polar to non-polar ratio, which is a ratio of polar atoms to non-polar atoms, is useful in predicting the solubility of compounds. So, all these features are derived features, which means we are using two or more features to create this feature, and then we can use them for predictive modeling, making ML models for their use in drug discovery and development. Okay, so now that we have looked at the features, let us have a look at what machine learning is.

So, now we have those features, and then we use them as input to this model, and we get an output, which is a prediction, okay. So, the idea of this model is that, for example, if you wanted to predict the solubility, we input the features of a molecule into this model, which has been trained on labeled data, if we are using supervised learning. And then it will predict the solubility, and we will see how that solubility compares to reality. And then we will say whether the model is good or bad.

Let us see how we can do that. So, before that, let us have a look at supervised learning. So, we have discussed it earlier as well that in supervised learning, it learns from labeled data to predict the outcomes. And there in every day, you see hundreds of examples of supervised learning, like weather forecasting, whether it is going to rain tomorrow or not, what the temperature will be for the next five days, and the speed or wind direction in the next five days or the next week. So, the supervised learning can be classified into two types. So, one classification is where it will just tell us about the presence or absence of something, like whether the compound is typically discussed in drug discovery.

So, it will tell us whether the compound is active or inactive, or soluble or insoluble, and then we use algorithms such as support vector machines or logistic regression for classification models. And then we have the regression models where they can, you know, predict the continuous value. So, they can predict quantitative values like they can predict solubility, permeability, or IC50 values. And for that, we use algorithms such as linear regression or random forest. So, if we just look at the difference between classification and regression.

So, the classification can be of multiple types. So, it can be, you know, binary classification where it will predict two possible outcomes: whether the compound is active or inactive, and there we use logistic regression, SVM, or random forest. Or it can be a multi-class classification where it can predict more than two classes. And that can be whether a molecule is an agonist, an antagonist, a neutral ligand, or whether the molecule is partially soluble, completely soluble, or insoluble. And then here we use the algorithms: our decision trees, KNN, and neural networks.

And then we have multilabel classification, where each sample can belong to multiple classes, such as predicting whether a molecule has antimicrobial and anti-inflammatory properties, antimicrobial and antiviral properties, or just antimicrobial properties. So these are, you know, multilabel classifications, meaning each sample belongs to multiple classes, and then we use, you know, multiple algorithms like neural nets or adaptive random forests. And then, in the case of regression, we can have a simple regression where we talk about only one feature. And then, whether that feature is correlated with the property of our interest. So, for example, can we use molecular weight to predict the IC_{50} values? So, if the molecular weight correlates well with the IC_{50} value, then we can use this molecular weight to predict the IC_{50} value of new compounds, and this is called simple regression, for which we use linear regression or SVR.

And then we have the multiple regression, where we use multiple features to determine one output, such as molecular weight, logP, H-bond donors, and all these features to predict the bioavailability of a molecule; there we can use Ridge regression or LASSO regression. And then we have the

polynomial regression which captures the non-linear relationship. So, if the relationship is non-linear, we can use polynomial regression. For example, the dose versus response curve, and there we use polynomial regression models. And then we have another, you know, a type of model or method, which is a log oblique exponential regression, which handles, again, non-linear exponential or decay relationships, like determining drug concentration decay in plasma over time, and there we use the non-linear regression models.

So after talking about these models, just a brief note. So let us see how this machine-learning model works. As I said, we input those features into this model, and then it will predict the outcome. So, let us have a look at how this model learns to predict the outcome or the property, actually. So, before that, we need to have a look at the dataset.

This is data from solubility measurements, representing a measured log solubility, which means solubility in moles per liter of many compounds—somewhere around more than 1,000 compounds. And you can see here, so we have the index here, and then we have the name of the compound or compound ID. And then we have the measured log solubility, and then we have the SMILES, which represent the structure of those molecules. And then, each of their rows represents a different sample in the data set, like we have this compound which has a measured log solubility of minus 1.74, and this is the SMILES structure for this compound.

And then these columns are representing different features that we have calculated from the structure, such as the molecular log P, molecular weight, number of rotatable bonds, and the aromatic proportion. And these features, these properties, we have calculated from the SMILES of the structures of these molecules, which we have in this data set. Okay, so we have the features, and the next thing is the labels, which means these are the activity data, or we can say the property data. So, in this case, this is the solubility data. Each of these molecules has its associated solubility given in this column, actually.

So now what we do is we label it; we call it matrix Y, which is our label matrix, and then we call our feature matrix matrix X. Okay, so now we have got that feature matrix and the label matrix, so let us see how it works and how this model learns to predict the values. So, for example, here when we input these features into this model, this model is making a prediction, and for example, this model has predicted the solubility, which is the y matrix of y minus 1.8. However, the actual value of this property, the solubility, was minus 2.18 for this specific molecule. And this difference between the actual and the predicted is called the loss because the prediction made by this model is that far from reality. Okay, so now taking this loss, this model uses this information in training, and then it trains the model so that the difference between this prediction and the actual value is reduced to the minimum value. And that is in the ideal situation, the prediction will be exactly the same as that of the actual value, and then in that case, we will call it the best model. Okay, we have 20 molecules here; of course, in this data set, we have more than 1,000 molecules, but I have shown only 20 molecules here. So, we cannot use all those 20 molecules for the training purposes in this model because if we use all those 20 molecules, how will we determine whether the model is working fine or not and if it is learning properly or not? So for that, what we do is split this dataset into a training dataset, a validation dataset, which is used to validate the model, and then a test

dataset, which is used to evaluate the model.

And of course, the training data set is used to train the model, and then we can either split it into 60:20:20 or in the ratio of 80:10:10. And now let us see how this validation data set and test data set help in making the model better. So, when we are using a training data set, this model is learning, making predictions, and then we compare the predicted value to the actual value. And then we determine the loss, which is the actual value minus the predicted value. And then this loss is used for training the model, where it makes adjustments and then tries to reduce the loss.

However, the validation data set is used as a reality check because if we are using that training data set, we cannot actually evaluate how good the model is. So, for example, if we have 18 or 20 molecules, we are using 16 molecules for training purposes. So by using only those 16 molecules, we cannot validate that model because those molecules have already been used to train that model. So, we need some molecules that are not a part of the training; then we can try to see how well that model is learning. So that is called the validation data set, where we feed the features of that validation data set into the model.

We make a prediction and then we see how close it is to the actual value or how much loss there is in that model. And then this validation set is used as a reality check during or after the training to ensure the model can handle unseen data. So, for example, we have these four models where the loss for model 1 is 1.3, the loss for model 2 is 1, the loss for model 3 is 0.5, and the loss for model 4 is 0.8. So, we can say that model 3 is the best model, as it has the minimum loss and the predicted values are close to the actual values.

And then we have the test dataset, which is used to check how generalizable the final chosen model is. Because when we are using the validation data set, the feedback from the loss obtained in the validation model helps us in optimizing the model. But the test set is the actual, you know, actual data which is outside the training and validation sets, which is used to check how general the model can be. Means, for example, if we are using 2000 molecules to predict solubility, we are making a solubility prediction model.

So, can we use that model to predict the solubility of any other molecule that is not a part of that data set? So, this is the actual reported performance that we report: okay, this is how the model is performing and how close the prediction for that model is to the externally observed molecules in the test data set. We have two types of loss functions: L1 loss function and L2 loss function. So, the L1 loss function is used to minimize the error, which is the sum of all the absolute differences between the true value and the predicted values, and it is represented by this formula. And then we have the L2 loss function, which is used to minimize the errors, which is the sum of all these squared differences between the true value and the predicted value. And you can see a difference; the blue one is, you know, the L1 loss function, and the red one is the L2 loss function.

So, what you can observe here is that in the case of the L2 loss function, if the difference is minimal, meaning if it is very close to the predicted value, then the penalty is actually minimal. It is far from the predicted value, and the penalty is very high; for example, for the same value of loss, the

difference between the loss or the penalty for the blue one and the red one is actually quite significant. But they are close to each other when the difference between the actual and predicted values is very small. Okay, so let us talk about some algorithms that are, you know, used for supervised learning.

So, we can have it for both classification and regression. So, where the classification predicts categories like soluble or insoluble and the regression predicts continuous values like logP or IC_{50}. And then, these are some algorithms, like KNN (k-nearest neighbors), which can be used for classification as well as regression. We have support vector machines that can be used for both classification and regression; decision trees and random forests can also be used for classification and regression. And then we have the logistic regression, which is used for classification; we have the linear regression, which is used for regression only; and then we have the Naive Bayes, which is used for classification. And then for unsupervised learning tasks, like clustering or dimensionality reduction.

In clustering, we group the data according to the features; for example, the libraries are clustered by their chemical similarity, or in dimensionality reduction, we try to reduce the features, such as using PCA (Principal Component Analysis) on molecular descriptors. And here we have, you know, the algorithms like k-means, which is used for clustering, hierarchical clustering, which is again used for clustering, DBSCAN, which is again used for clustering, and then we have principal component analysis and t-SNE, which are used for dimensionality reduction. In the case of having, for example, 10,000 molecular descriptors or features for a data set, we can reduce them to some meaningful features, which can be done by using PCA and t-SNE. Okay, so let us have a look at some of these algorithms, like the first one, which is the k-nearest neighbor. So it works for both classification and regression, and how it does so is based on those x1 and x2 features; what it does is classify into category 1 and category 2.

And then, whenever a new data point comes, how it predicts is that it calculates the distance of this data point from its neighbors, and this distance is called Euclidean distance. So, it determines the Euclidean distance, and based on that Euclidean distance, it classifies it into either category 1 or category 2. And for regression, it can also average the values of the neighbors and then determine, for example, logP values based on the average value of the neighbors in this case. And how can we use it in drug discovery? We can predict whether a molecule is an inhibitor, like in a classification, or we can also do a regression by estimating the molecule's binding affinity by averaging nearby known compounds.

And then we have the support vector machines. This is, again, you know, it can be used for classification and regression. And in this case, it has a hyperplane. And then that best separates data into classes for classification. And for regression, it tries to fit data within a margin of error while keeping the model simple, which leads to less overfitting.

And then it can handle nonlinear data as well using kernels. So how we can use it in drug discovery is that we can classify compounds as actives or inactives based on molecular fingerprints. We can also predict the physicochemical or biological properties, such as solubility or toxicity, by using

SVM. Then we have the decision tree; it also works for, you know, classification and regression. And basically, it's a kind of decision flow chart that splits the data into yes or no decisions based on features, such as whether the logP is greater than two.

Yes or no, so then it will have one branch and another branch. And then it grows the branches until the data is pure; all samples in a leaf belong to the same class or are close in values. However, it is prone to overfitting, but we can use random forest, and the random forest is actually an ensemble of these trees. So, it can be used to reduce overfitting in decision trees, and we can use it for classifying molecules based on structural alerts for toxicity, or we can also use it for predicting the ADMET properties. And then we have the random forest, where we use multiple decision trees; actually, bagging means we are taking the output from all these decision trees, and then we are making a prediction on the basis of that. So it can also be used for both classification and regression, and then it takes a majority vote again through the classification or averages the predictions based on whether we are using it for regression.

Then it reduces overfitting compared to a single decision tree, and it can be used for predicting IC_{50} values of kinase inhibitors or any other inhibitors. And it can also be used to identify active compounds from high-throughput screening data as well. And then we have linear regression, which is used only for regression. And then here you can see that we just have, you know, this dependent and independent variable. We try to fit it into a straight line equation and then it finds the best fit line through data by minimizing the differences between predicted and actual values.

And then it assumes a linear relationship between features and targets like molecular weight versus solubility, where molecular weight is directly correlating with the solubility. and we can use it for predicting logP logD or other physiochemical properties and we can also use it for modeling the dose response curves. And then we have the logistic regression which can also be used only for classification. So despite the name it is for classification for either for binary or multi-class classification. So it uses a sigmoid function to squash predictions between 0 and 1 and also predicts the likelihood of a compound being active inactive based on the molecular descriptors.

So we can we can classify those compounds you know maybe active or inactive or hits or not hits or soluble or insoluble and we can also predict for example whether a molecule is crossing blood brain barrier or not by using this classification method algorithm known as logistic regression. And then we have the principal component analysis. So, this is you know a dimensionality reduction method which is unsupervised and it reduces a high dimensional data set to fewer components while preserving most of the information and mainly it helps in visualizing data and improves model performance by removing the noise. And then in drug discovery, we can reduce thousands of molecular descriptors to 2D, 3D for visualization. And also, we can pre-process large compound data sets for faster training by using PCA.

Coming to the summary, so these machine learning models, they learn from data to recognize pattern, make prediction and classify or classify new information. And then we have features and labels that are essential part of the data set where the features describe data points and labels define the outcome. And then we have the supervised learning which focuses on labeled data like for

classification regression while we have the unsupervised learning which finds hidden pattern in the unlabeled data for example for clustering. And then the model training validation and testing ensure the model generalizes well to new and unseen data because the overall objective of making ML model is to predict the properties of new data set external data set which has not been seen earlier by the model. OK, and then I have some suggestions for you to further enhance your knowledge in this topic by going through this literature, these references.

And then in the end, I have an open question for you. Suppose you build a model that predicts a molecule as active with 95% accuracy, but in the lab, most of the compounds still fail. So why this might happen? So just have a thought on this. And with that, thank you.