AI in Drug Discovery and Development

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

Week-06
Lecture-27

Welcome to the course "AI in Drug Discovery and Development." In today's session, we will talk about AI tools for drug-target interaction studies. So, by the end of this lecture, you will be able to understand drug-target interactions and their role in drug discovery, as well as the challenges in traditional DTI studies. Also, explore the AI-driven approaches for drug-target interaction prediction and identify the challenges and limitations of AI in DTI prediction and validation. So, let us talk about drug-target interactions. As we have seen, drug discovery is a very complex and challenging task, and the overall objective of drug discovery is to identify molecules that can engage various drug targets.

For example, this is a drug molecule, and when it binds to this enzyme, it is a target for this drug molecule. So, it inhibits this action of this enzyme and then leading to the therapeutic effect which we wanted to have. So, there are ways by which we can determine the drug-target interaction. So, the question is: can we use artificial intelligence tools or ML tools to do this automatically? So, we will explore that in this session.

If we look at the DTI prediction, it is a process for predicting interactions between chemical compounds or drugs. and their corresponding biological targets, which can be proteins, nucleic acids, or enzymes, although enzymes are themselves proteins. So, why is the drug-target interaction important in drug discovery? Because it can enhance our molecular understanding for new drug development and repurposing. So, how can a molecule interact with this new target? Can it be inhibited? Can its function be activated, inhibited, or altered? All of those things can be understood by using this DTI prediction. We can identify therapeutic compounds by targeting specific diseases as well.

We can predict the drug's effects and side effects to identify safer medications. Especially those medicines that have a lot of, you know, side effects—severe side effects like those chemotherapeutic drugs—we can say. And then it can also enable personalized medicine by tailoring therapies to the genetic profiles as well. Because sometimes you know that some people have proteins that are overexpressed or underexpressed, can we use those as drug targets? So, that is called, as you know, personalized medicine. And then it can also help us optimize the multi-drug regimen by managing the drug interactions as well.

We can identify drug-drug interactions as well as drug-target interactions, which can eventually eliminate the ineffective candidates early in drug development. So, if we wanted to see whether the compounds are going to be active or not. So, we can screen them out at the very beginning of preclinical drug discovery. So, as I said, there are multiple methods by which we conduct drug-target interaction studies. And this can be classified into, you know, binding affinity in kinetics-based methods, structural determination methods, thermal stability methods, and gel electrophoresis methods.

So, let us talk about some of the limitations of these methods, and then we will be able to see how AI-based tools can help us overcome those limitations. So, the first method we will talk about is surface plasmon resonance. So, it detects the real-time biomolecular interaction by measuring changes in the refractive index on a sensor surface. The limitations of this instrument are that it is very expensive and requires a lot of purified protein to run the assay. And then the data interpretation is also complex, especially if you want to do interaction studies between a small molecule and a protein, which is actually quite difficult with this method.

So, it is suitable for you to know about biomolecular interactions like protein-protein interactions or protein-nucleic acid interactions, but doing small molecule interaction studies with proteins using SPR is a challenge. And then you have isothermal titration calorimetry (ITC), which is again one of the gold standard methods. which measures the heat change during molecular binding to determine affinity and thermodynamics. However, the limitation is that it requires very high sample consumption. So, again, it needs a lot of protein for, you know, running the assay, and if you do.

We are working on a new target, and with limited resources to produce that purified protein, it can be challenging to use ITC. And then it is also low throughput because you cannot automate it, and you cannot screen hundreds of compounds in one run, and it also requires high protein purity as well. And then you have microscale thermophoresis, which uses temperature-induced motion of biomolecules to quantify the binding affinity. Again, the key limitation of the MTS assay is that it requires fluorescent labeling of those proteins. It is also sensitive to impurities, and it is limited for large molecules because you cannot use it for small molecule protein interactions.

And then you have a fluorescence polarization assay, which measures changes in molecular rotation to determine binding interactions using fluorescence-labeled molecules. So, again, it requires you to know about fluorescent labeling and may suffer from background signal issues. And then you have binding assays like radiolabeled, calorimetric, or fluorescent assays, which measure the binding affinity between a drug and its target using either fluorescence, radioactivity, or calorimetric detection. So, these methods suffer from some limitations, such as low throughput, interference from the sample matrix, and sometimes requiring labeled compounds, especially in the radiolabeled binding assays. And then there are structural determination methods, such as X-ray crystallography.

Where you determine the atomic structure of a drug-protein complex by analyzing the diffraction patterns from single crystals of the compounds, along with the protein. So, these are, you know, time-consuming, and they require crystallization, which is again a challenge because many proteins do not like to be crystallized, actually. So, single-particle cryo-electron microscopy uses an electron beam to capture high-resolution images of biomolecular complexes in a near-native state. So, this is one of the best structural biology techniques, you know? However, it requires very specialized equipment, which is expensive, and it also requires high computational power. And then, usually, the resolution is lower for small molecules when you work with them.

So, again, this is a challenge for you to identify DTI, direct target interaction, for small molecules. And then you have thermal and stability-based assays, such as differential scanning calorimetry, which measure heat capacity changes to determine protein stability and binding interactions. So, it is a kind of indirect measure of binding that requires highly pure protein samples. And then you have the electrophoresis gel-based method, such as the electrophoretic mobility shift assay, which analyzes protein-DNA or protein-RNA interactions based on shifts in electrophoretic mobility. It is only limited to, you know, nucleic acid-protein interactions, and it is very difficult to quantify; it is also prone to nonspecific binding.

And again, it cannot be used for small molecule inhibitors to determine their interaction with the target. So, these are some of the challenges of traditional DTI studies, and let us see how AI can transform them. So, there are different aspects that we will discuss, including AI-driven predictions and traditional methods. So, on the basis of principles, it uses DL/ML and computational models to predict drug-target interactions based on either descriptors, sequence data, or structural information; these are just the features, in fact. So, you use the features of molecules as well as the features of proteins, or you know your targets, such as RNA and DNA.

And then, by using those features, you try to predict whether those two molecules will interact with each other or not. However, traditional experimental approaches rely on physical experiments, such as binding assays and X-ray crystallography. NMR spectroscopy and high-throughput screening are used to determine direct target interactions. So, if we talk about speeds, the AI-driven approaches are very fast; they can analyze millions of compounds in hours or days. However, the traditional approaches are all experimental-based, so they are a bit time-consuming and resource-intensive as well.

Talking about the cost, these AI-driven predictions are low-cost, and the expenses are mainly associated with the computational processes, which include hardware, software, and cloud services. However, the traditional methods are highly expensive; they require a lot of consumables, equipment, and labor-intensive processes. In the data requirement aspect, the AI-driven approach requires high-quality datasets for training, such as binding affinity data, molecular structure data, interaction network data, and omics data. However, the traditional approach is that we do not need much data because we are generating the data by conducting those experiments. However, we require purified proteins, libraries of drugs, screening libraries, and access to cell culture and assay-specific reagents.

So, talking about scalability, AI-based approaches are highly scalable; they can handle large datasets and multiple targets simultaneously. However, for traditional methods, you know, if you want to scale them up, we need multiple instruments, and that is, again, resource-intensive. Talking about accuracy, in the case of AI/ML modeling, we can achieve high accuracy; but again, that depends on the quality of the dataset, model bias, and generalization ability. So, if we have good-quality data, I think we can generate good models that efficiently predict the target interactions. For the traditional methods, the accuracy is quite high.

However, it may vary from batch to batch and other factors as well. So, the interpretability aspect is that these AI models are considered to be black boxes. However, we can use XAI, or explainable AI, which improves model interpretability. However, the traditional methods they provide give direct mechanistic insights into the binding interactions, and you can also obtain structure-activity relationships. For novel target discovery, we can identify new target interactions using AI trained on diverse biological and chemical data sets.

However, for novel target discovery, it is difficult to use traditional methods because. You know that until we have a developed assay for that target, we cannot use it to screen for the ligand. And then, the personalization potential of AI-based models can be used to enable precision medicine by predicting patient-specific drug responses using AI models trained on genomic and transcriptomic data. However, traditional methods are less adaptable to individual variability and require extensive clinical validation. Finally, the validation requirements, so that you know the AI-based models, ultimately require us to validate the findings of the AI-generated predictions experimentally.

However, traditional methods are being used because we are already conducting experimental work. So, these are, you know, self-validating and obtained from direct measurement and assays. So, these are some of the aspects in which AI-based drug target interaction studies are changing the process of drug discovery and development. Okay, coming to the tools, these can be divided into four categories. So there are feature-based methods, similarity-based approaches, graph neural networks, network-based models, and deep learning and transformer model-based approaches for DTI prediction.

First, we will look at the general workflow, and then we will discuss these methods one by one. So the general workflow is as follows: you take the target structures, and then you take the screening library structures or the drug structures. And then you extract the features, select those features, make a drug-target pair, and use them as a training set and test set. And then you develop a classifier, and this classifier can classify a new molecule based on whether it interacts with this target or not. So, this is a very generic workflow for working with those, you know, DDI prediction methods.

So, let us talk one by one. First, we will discuss the feature-based methods. So what we do here in these feature-based methods is convert molecules and targets into numerical features to create ML models. Some of the key features that we use are molecules that can be converted into molecular descriptors. So we discussed earlier that these molecular descriptors represent features of molecules in numbers, such as molecular weight, log P, H-bond donors, acceptors, etc. We can also generate chemical fingerprints, which are fixed-size bit strings that represent molecular substructures, such as eccentric connectivity fingerprints or MACCS keys.

For the target, these two features belong to the drugs, the small molecules, and the target can be a protein. And then we can also generate the chemical fingerprints, which are fixed-size bit strings that represent the molecular substructures; these can be eccentric connectivity fingerprints or MACCS keys. And then for the target, these two features belong to the drugs: the small molecule and the target, which can be a protein. So, usually, we have the amino acid sequence, the amino acid composition as features, and the binding site features if we know the binding site. Which amino acid residues participate in the binding of an inhibitor and activator in the binding site? So, we can use that information as a feature.

So, some of the models which we can use like Random Forest, Support Vector Machine, XGBoost. So, these are well suited for structured feature data. And then you have DeepChem or Mol2Vec, which can be used to generate features for either the ligands or the target as well. And then these are some of the methods, such as Krone RLS. So, it is unclear where the input drug representation is the molecular fingerprints and the input protein representation is the protein sequence.

The algorithm used for the drug feature learning method is kernel-based regression for both the drug features and the protein features. And then, it uses regularized least squares for DTI predictions. And then we have SIM Boost, which again uses molecular descriptors as the drug representation and protein descriptors as the input protein representation. And then it uses gradient boosting trees for drug feature learning and protein feature learning, followed by feature engineering-based learning for DTI prediction. And then we have the deep DTA, which uses SMILES as an input representation and the protein sequence as the input protein representation; it uses CNN-based feature extraction for both the drug feature and the protein feature.

And then it uses a convolutional neural network to learn molecular and protein features and creates a model to predict whether a molecule will bind to the target. So, for example, you can see the deep DTI here: you have the smiles of the ligand, and then you can use a convolutional neural network. So, it converts them into features, and then you have the protein sequence, and then you use the CNN again. So, it convert them into sequence representation and by using these features it generates you know a model which can then predict the DTI for a new molecule. So, it uses the sequences embedded in the process by the CNN to detect substructure patterns that lead to binding affinity predictions.

and it is usually more accurate than classical methods like KronRLS or SIMBOOST achieving a higher concordance index and lower MSC on benchmark data set like Davis and Kiba. So, these are benchmark datasets to benchmark those DTI prediction tools. Okay, coming to the next type of tools, which is similarity-based approaches. So, what they assume is that the structurally similar compounds interact with similar targets; we have already seen that if two of the molecules are structurally quite similar, they do interact. So, they may bind to the same target; that is the idea.

So, again in this case, the similarity methods are being used. So, the chemical similarity Tanimoto coefficient is a fingerprint-based method. So, first calculate the fingerprints of all these molecules, and then calculate the Tanimoto coefficients. If it is 1, then it means that those molecules are completely the same, and if it is close to 0, then they are completely different from one another. And then you use a sequence similarity search, where you can use the BLAST (Basic Local Alignment Search Tool).

You can use the sequence alignment-based target similarity calculation, and for structural similarity, you can also use molecular docking-based binding site comparisons. So, some of the challenges associated with similarity-based approaches are that they struggle with novel scaffolds. So the basic principle of these tools is that if two compounds are similar in nature, they will act on the same target. So if we have novel compounds, which have not been earlier reported to bind to a known target, so how they will be identified by using this approach.

So that is one limitation. However, we can combine it with deep learning for better generalization, or it can also be used for novel scaffold identification. So, some of the features of the chemical similarity-based tools are the BLM and the NII. So, where the input representation for the drug is a chemical similarity matrix and for the protein is a protein similarity matrix, it uses Bayesian learning for drug feature learning and protein feature learning. And it utilizes known interactions and similarity scores. And you have the NRLMF, which again uses a chemical similarity matrix and a protein similarity matrix for the input drug representation and the input protein representation, respectively.

And it uses matrix factorization to learn drug features and protein features. And then it predicts interactions using neighborhood regularization and regularized learning. And then you have CMF DTI, which uses molecular similarity and protein similarity as inputs for drug representation and protein representation. So, we have the basic models, the key GNN-based models like molecular graph representation, where you represent the atoms as nodes and the bonds as edges. And then, in the case of the protein interaction network, it uses the protein-protein interaction network to infer drug-target relationships.

So, some of the challenges associated with these GNN models are that they are highly computationally expensive. So, we require a lot of computational power as well as large datasets to improve scalability and interpretability. These are methods like Graph DTA, Graph CPI, mGraph DTA, and GAN DTI, including Graph DTI. All three of these models work with molecular graphs as input for drug representation and with protein sequences as input for protein representation. And then they extract the features or learn the drug features using a graph convolutional network.

Then, they use a CNN-based sequence embedding to learn the protein features, and they particularly like Graph DTA, which learns molecular interactions through graphs. And graph CPI uses multiple GNN architectures for better prediction, while M graph DTA uses multi-scale learning for drug-target interactions. And GAN DTI uses adversarial training for better generalization because it employs an attention module to learn the protein features and a residual GNN to learn the drug features. So, this is, for example, mGraph DTA, which integrates a multi-scale graph neural network for drug encoding, a multiscale convolutional neural network for target encoding, and a multilayer perceptron for final binding affinity prediction. So, it introduces a novel gradient-weighted affinity activation mapping for visual interpretability that leverages gradient information from the last graph convolutional layer to highlight important molecular substructures for binding predictions.

So, you use the input drug graph, and then you use the input protein sequence with the MCNN and MGNN. So, it concatenates all that information using an MLP to predict the affinity, as well as to identify the fragment of the molecule that will be important for binding to the protein. And then we have graph DTA, which is a graph neural network-based model for target affinity prediction; it uses molecular graphs and protein sequence embeddings for interaction learning. So the key features are that GNNs for molecules capture graph-based structural properties, and CNNs for proteins learn from sequence embeddings. And it is trained on large-scale bioactivity datasets such as Kiba and Davis.

And some of the strengths are that it captures complex molecular interactions more effectively than the simple ML models do. However, the limitation is that it is computationally expensive and requires a GPU for training. And we can use it for binding affinity prediction, which is, again, a direct target interaction prediction. Or we can use it for AI-powered virtual screening and multitasking discoveries. And you can go through this, you know, the GitHub link that I provided.

And then there are deep learning and transformer-based models. So they use these deep learning architectures like CNN's RNN transformers to extracthidden patterns from large-scale datasets. So, it uses CNNs and RNNs to learn the spatial and sequential patterns in molecular sequences, and it uses transformer models for the self-attention mechanism to achieve better feature learning. Some of the challenges are that we require high-quality data sets, and it also suffers from the black-box nature of its interpretation. The solution is that we shall focus on the explainable AI tools. There are methods such as Transformers, CPI, BioBERT, and ProTraS.

We have DTI Voodoo, MOL-TARC, and DUPURPOSE. So mostly, the input drug representation is either molecular features, molecular descriptors, SMILES, or text-based drug representations, especially for BioBERT and ProTrans. And the input protein representation is usually the amino acid sequence or the text-based protein representations. And then the key features of TransformersCPI are that it uses Transformers to model the interactions. BioBERT and ProTrans leverage natural language processing models for understanding molecules and proteins. And DTI Voodoo uses a self-attention mechanism for DTI predictions.

And the MOL-TARC learns shared features across multiple tasks. And Deep Purpose uses an open-source deep learning toolkit for DTI. Okay, so this is a little bit more detailed information about Deep Purpose, which is an open-source deep learning framework for DTI prediction. Provide pre-trained models and a flexible architecture for easy integration. Okay, this is something we have already talked about. So, the strengths are that it is easy to use, scalable, and adaptable.

However, the limitation is that it requires high-quality labeled data for accurate prediction because the data is, you know, the key here; if we have good quality data, then only can we obtain a better model. So, it can be used for virtual screening by activity prediction or drug repurposing, and this is the GitHub link for the deep purpose. And then we have the DTI Voodoo as well. So, it is a transformer-based deep learning framework for drug-target interaction prediction that uses a self-attention mechanism to learn from molecular and protein sequences. So, what it does is incorporate transformer models to capture complex drug-target interactions, and it learns from sequence embeddings like SMILES and FASTA without handcrafted features.

It can handle zero-shot learning for novel drugs and targets as well. However, the limitation is that it requires large-scale pre-training to avoid overfitting, and the good thing is that it has high accuracy and generalization for unseen drug-target pairs as well. And we can use it for virtual screening, AI-driven drug repurposing, and target-based predictions as well. And then we have BioBERT and ProTrans. So, this is an NLP-based protein-ligand interaction model.

So, the strength of which we have already discussed is. So, the strength of BioBERT and ProTrans is that it leverages textual and sequential data for enhanced prediction. However, the limitation is that it requires fine-tuning on specific DTI datasets for optimal performance. And here is the GitHub link for BioBERT, which can be explored further. Okay, let's talk about multimodal AI for DTI prediction. So, instead of using, you know, a single kind of representation for either drugs or the targets, can we use multiple representations and fuse the data? So, it leverages various types of data—specifically structural, sequential, and chemical data—to enhance the accuracy and effectiveness of its predictions.

So, we can use the chemical data, which include the chemical properties such as descriptors and physicochemical parameters. We can use the sequence data, which are either the SMILES for the small molecules or the protein sequences for the protein molecules. Or we can use the structured data that represents the three-dimensional structure of the molecules. So, this is multimodal feature fusion and domain generation: MMDZDTI. So, it is one of the examples that uses, you know, multiple kinds of features, like a textual feature extractor.

So, for example, it uses ProdBERT to extract features from the protein structure. Smile bar for extracting features from the ligands; it can also take the amino acid sequence, the graph structure representation, and finally use a kind of fusion DTI. By using multiple input features and multiple algorithms, it can predict the drug-target interaction between molecules. Okay, so this is one of the popular tools that is used for, you know, predicting the activity spectra of substances or predicting drug target interactions.

So, you can see here that it contains around 1.6 million substances and has approximately 143,000 descriptors. And it contains around 10,000 types of activities; this means that if you just draw your structure into this tool, So, it can predict which targets this molecule can work on out of 10,000 different activities. However, there are some challenges, and they are mostly related to data and model validation. Because we understand that we need high-quality data sets for making those AI-driven DTI models and for the successful prediction of direct target interactions. So, because it depends on well-curated data sets for accurate predictions, high-quality data ensures reliability and reproducibility.

Some of the challenges in data quality are that the data are inconsistent in nature because there is a lot of variability in experimental conditions, which affects comparability. And sometimes it is not reliable either because we usually get the data from public sources. And then, due to sparse negative samples, there is a lack of confirmed non-interacting pairs, which skews the training because we know that this ligand interacts with this protein. Experimentally, we have that data, but another ligand is not interacting with the protein; that information is what we are missing. So, that is also somehow leading to biased training, and class imbalance is also present, where the overrepresentation of well-studied target pairs mostly dominates the world of DTI.

So, these are some of the data sources for DTI modeling, where you can see the targets, the number of drugs, and the number of interactions. You can see that ChEMBL is the leader, where a lot of DTIs are available, and then you have the CAG, the DGI, the DB, SuperTarget, and the BindingDB as well. And finally, even if we are making a prediction using, you know, these tools. So, we need to validate those targets experimentally. Those AI predictions must be validated through laboratory experiments to confirm real-world relevance.

So, we have discussed those methods earlier, at the beginning, that we can use the in vitro validation methods either using the SPR, fluorescent radiolabeled binding assays, or cell-based assays. Or we can perform the in vivo validation using animal models, where we can measure the drug-target interaction and the corresponding effects through the assessment of pharmacokinetic efficacy and toxicity. For the CRISPR knockout studies, we can confirm the target specificity in biological systems. So, coming to the summary, AI is transforming DTI prediction by leveraging ML, DL, and graph neural networks, and it is not limited to that. And then there are supervised learning methods, including similarity-based and feature-based approaches, that enhance accuracy in DTI modeling.

And then there are advanced models like transformers and multimodal AI that improve feature extraction and prediction reliability. However, despite advancements, challenges such as data quality, model interpretability, and validation remain key concerns in AI-powered demand prediction. And in the end, I have an open question for you: if AI predicts a strong drug-target interaction, but experimental validation fails to confirm it, how would you resolve this issue? How would you approach this? What kind of changes will you make to your model to solve this problem? I have some suggestions for further reading that you can go through to get more information about this topic. And with that, thank you.