AI in Drug Discovery and Development

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

Week-05
Lecture-22

Welcome to the course "AI in Drug Discovery and Development." In this session, we will talk about AI tools for virtual screening. So, by the end of this lecture, you will be able to gain insight into the various tools that harness the power of AI for virtual screening. And also understand the workflow of various AI-based virtual screening tools. So, if you look at how AI is transforming virtual screening, So, in the previous lecture as well, we talked about how we can utilize artificial intelligence tools for screening large ligand libraries for identifying hit compounds. So, we can screen millions or even billions of compounds very easily, especially using the ligand-based virtual screening tools.

We can denoise the differences between inactive and active compounds using the intricacies of AI models. And then we can automatically train those AI tools through ChEMBL data sets. So, we do not need to manually curate the biologically associated biological data. So, we can directly get it from ChEMBL.

And then there are a number of benchmarking metrics to grow model performance that we can evaluate. And then it is highly customizable as well; for example, the RMD algorithm is very specific for ligand-based virtual screening. If we look at this picture, we can see the workflow of the hit identification using virtual screening. And then, if we divide this into these four quadrants based on the structure of the receptor or the ligand, if the receptor structure is available. And the ligand structure is available but not available.

So, we have both the ligand and receptor structures available. So, then what we can use is the combined or hybrid method where we can combine ligand-based virtual screening and structure-based virtual screening. And if we have only the structure of ligands available. So, what we can do is use those ligand-based virtual screening methods, and it starts with the datasets. So, here, for example, we have the data set which we divide into the training data set, test data set, and external test data set to validate those models.

and also the unexplored chemical space that we use for screening purposes. So, what we do here is take this data set and then feed those molecules into the descriptor calculation or feature calculation. So, we can calculate those features, like fingerprints, or they can be molecular descriptors. And then we also have the associated biological activity that we are interested in predicting. So, by using this, you know biological activity as a dependent variable and those features as independent variables.

We make the ML models machine learning-based models or AI-based models using either SVM, RF, KNN, MLP, XGBoost, or CNN/deep learning. By using this model, once we develop it, we validate it and then use it to screen this unexplored chemical space in order to identify the hit compounds. So, this is the workflow of a virtual screening using ligand-based methods. So, if we do not have the structure of the ligands, but we have the structure of the receptor or the target structure available. So, what we do is use the structure-based virtual screening methods, where we use the three-dimensional structure of those biomolecules or target molecules.

So, then we use the structure of the screening libraries to screen the compounds from those libraries. And then we dock those compounds, and we can perform molecular docking using different methods; for example, you can use the classical methods. Or we can use the, you know, AI-based methods like DeepDocking or GNINA, DeepDock, KarmaDock, and then we can score those molecules. And after scoring those molecules, we can identify the hit molecules, and then those will be further experimentally validated and optimized using multi-parametric optimization methods. and identified as a drug candidate, developed as a drug candidate.

So, if we have both the structure of the target and the structure of the ligands available, and here ligand means the three-dimensional structure of the ligands that have associated biological activity. So, it does not mean that we have all those screening libraries where we know the structure of the ligands, you know. It is the structure of ligands for that specific target which are binding to that specific target. Like if I take an example of acetylcholinesterase as it as a target. So, I need to have acetylcholinesterase inhibitors with associated IC_{50} values or biological activity values; only then will I consider them as the ligand structures that are available for that specific target.

And then I can use them for doing ligand-based virtual screening protocols. So, in this case, we take this database as an example: the database of known ligands that have associated biological activities or properties. And then we further divide them into, you know, training, test, and validation datasets, and then we do ligand-based virtual screening. So, here I have shown one method where we are just calculating the fingerprint or the descriptors and developing the model. But here, what we can do is either use the alignment-based, similarity-based screen, or we can use the graph-based, pharmacophore-based methods.

we can do the shape-based screening, or we can do the descriptor-based screening, which will consist of, again, you know, building a model, and then using that model, we can screen the library. And then, after screening the library, we can further screen them using structure-based methods, such as deep docking, Gnina, deep dock, or Karma Dock. And then we can identify the hit compounds, prioritize the hit compounds, and those hits shall be further experimentally validated and multi-parametrically optimized to obtain the drug candidate. So, this is, you know, an overall workflow for identifying hit compounds. But what happens when we do not have the structure of either the target, receptor, or ligands? So, we do not have the structure of the receptor available, and we do not know of any ligand that has activity against that target.

So, what we can do is predict the target structure using multiple methods. Like we can do, initially it was, you know, done by using homology modeling or threading methods, where we were predicting the three-dimensional structure of proteins. But now since AlphaFold has come, it is very powerful in predicting the accurate structure of proteins. So, we can use AlphaFold for predicting the structure. And then we can use that structure to perform the structure-based virtual screening method and then identify the hit compounds, which are further experimentally validated and optimized to obtain the drug candidates.

So, this is kind of a summary of how all these structure-based and ligand-based methods are being used for performing virtual screening to identify hit compounds, which are further optimized into drug candidates. So, if we look at the AI-based ligand-based virtual screening tools, there are similarity-based methods, pharmacophore-based methods, and QSAR-based methods. So, the similarity-based methods, some of the tools that work on the basis of similarity, are ROCS, OpenBabel, ChemMapper, RDKit, Swiss similarity, and SmallWorld. So, where Swiss similarity and SmallWorld are actually web servers. So, where can you draw the structure of your molecule and determine the similarity? Calculate the similarity and identify the compounds.

So, what it does is, as I summarized before, as well. So, it compares the molecular fingerprints. So, we calculate the fingerprints and compare those features with the known compounds. So, what it does is calculate and compare the molecular fingerprints of the library with the features of those known active compounds. In order to identify the compounds that have features similar to the known active compounds.

So, the advantages of these similarity-based methods are that they are fast, computationally efficient, and suitable for finding analogues of known drugs. And then you can see RDKit is a very popular tool that provides extensive cheminformatics support, and it is being largely used for doing all those sorts of things. So, this is a snapshot of the SmallWorld similarity search tool. So, where can you draw the structure of your molecule? So, here, for example, I just draw the structure of benzene. So, you can draw the structure of your molecule.

And then you will immediately see a load of molecules that are similar to your query, and you have the possibility to pick the data set that you wanted to screen. And then you know you have control over the properties as well, like the features that you wanted to control. And then here you will see a different sort of, you know, similarities, and based on that, you can pick up the molecules which you find suitable for your screen. And this is another tool called SwissSimilarity. So, here you can also see that you can enter the structure of your molecule in SMILES format, and once you can also draw the structure using the sketcher.

So, once you draw the structure, you can then select a class of compounds from this drop-down menu, where you can choose which library to screen. Ah, and then you can select the compound library and screening methods. So, you can either use the 2D screen features or you can use the 3D features. So, all these fingerprints and pharmacophore-based features or scaffold-based features, you can pick all those features, and then you can identify compounds that are similar to your query structure. And the basic idea of identifying, you know, similar compounds is the understanding that similar structures will have similar effects.

So, we assume that the compounds which are similar in structure or have similar features will have a similar kind of bioactivity. So, that is the basic idea. And then we have pharmacophore-based methods, such as different tools like Ligand Scout, Pharmer, and Phase, which is actually from Schrodinger. And then what it does is identify essential molecular features required for activity, such as hydrogen bond donors, as well as hydrophobic regions, and it provides interpretable results and visualization. It is highly useful when no three-dimensional structure of the target is available, and it also enables hypothesis-driven compound selection as well.

And then we have QSAR-based methods. So, we have several tools for QSAR based methods like the AutoQSAR, QSAR, DeepChem, PyRMD And then these tools use mathematical models to relate chemical features to biological activity, like using ML-based regression and classification models. So, like the PyRMD, we have been using it a lot. So, in this case, you can take a, you know, from ChEMBL; you can download the dataset of your active compounds. And then you can classify them into active and inactive based on their associated bioactivity, and after classification, you can train the model based on the features.

And then, using that validated model, you can predict the affinity or whether the compound will be active or not by screening a large library. So, some of the advantages of these QSAR-based methods are that they can predict the activity of untested compounds, enable high-throughput screening of large chemical libraries, and enhance model accuracy with machine learning as well. So, then we have the AI-based structure-based virtual screening tools. So, as I said, structure-based virtual screening tools largely consist of molecular docking tools. So, in molecular docking, we can enhance the molecular docking tools based on, you know, these four factors.

For example, we can use AI for docking acceleration, where we can optimize the speed of molecular docking. And, if you remember, the molecular docking was, you know, time-consuming. So, it was not very fast. So, if we can speed up the molecular docking process using these AI-based tools, it will be very good for increasing efficiency. And then we can also have some AI-based scoring functions that improve the binding affinity predictions.

So, the scoring functions which are being used in classical molecular docking they have several limitations. So, if we can come up with some advanced scoring functions that use AI or machine learning and that can accurately predict the docking score or binding affinity of the ligands towards their receptor binding pocket or target binding pocket. So, that will be very good for increasing the accuracy and efficiency of these screening methods. And then finally, we can use the generative AI-based docking tools; we have the flexible and novel docking methods, which are, you know, working on various principles. And then we can have AI-powered docking frameworks where we have the learning-based docking app that is currently in the development phase.

So, if we talk about the AI tools for docking and acceleration, where we are trying to optimize the speed and efficiency. So, there is a tool called AutoDock GPU (AI enhanced). So, it uses GPU acceleration combined with AI driven optimization. So, it significantly reduces docking time. And then we have the delta dock, which employs AI to predict docking poses faster, optimizing docking efficiency and screening large chemical libraries very rapidly.

And then we have Gigascreen, which uses AI to accelerate ultra-fast docking, significantly improving screening speed. And then we have deep docking, which uses QSAR to predict the docking score of the screening library. So, it is basically relying on the deep learning-based QSAR model, which can predict the docking score of the molecules before screening them using molecular docking. So, some of the advantages are that they speed up the virtual screening for large data sets if you want to screen maybe like 1 billion compounds. So, it is very difficult to do that using classical docking, but these AI-based docking acceleration tools can efficiently increase the speed of the process.

Make it possible to screen billions of molecules, reduce the computational cost, and make it suitable for high-throughput applications as well. So, deep docking is, as I said, one of the examples that utilize QSAR deep models trained on docking scores of a subset of the chemical library to approximate the docking outcomes for yet unprocessed entries. and therefore, to remove unfavourable molecules in in iterative manner. So, we have the validation test and initial training sets, which are randomly sampled from the entire docking library at the first dd pass. From the second iteration, the training set is iteratively augmented with random batches of molecules classified as virtual hits in the inference stage of the previous iteration.

So, we will talk about this deep docking in the later sessions. And then we have AI-based scoring functions that improve the binding affinity prediction. So, one of them is like G-NINA, which uses a deep convolutional neural network (CNN) to refine the scoring function, which improves the docking accuracy. And then we have ONIONNET, which is an AI-powered scoring function that evaluates protein-ligand interactions based on 3D spatial features. So, some of the advantages which those AI based coding functions they pose are they can provide more accurate binding affinity predictions and that is how they can reduce the false positives in virtual screening and also enhances the precision in lead identification.

So, it is a little bit more about the Gnina, which is a docking tool able to outperform Autodock Vina. However, it is built on AutoDock Vina; only pose generation and ranking are done by AutoDock Vina. the CNN models which are trained on a large number of compounds. So, they are used to recalculate the docking score of the molecules, and then it generates another score, which is the CNN score. And then another method is generative AI-based docking, where we use the flexible and novel docking methods.

So, like diff dock which leverages the diffusion models to predict ligand binding poses with high flexibility overcoming traditional docking limitations. And then we have DeepDock, which uses deep learning to model molecular interactions and docking poses more efficiently. So, the advantages of these methods are that they improve docking for flexible ligands and provide better generalization to unseen protein targets as well. and then they are the the AI driven molecular interaction modeling. And then we have AI-powered docking frameworks like these, which are learning-based docking approaches.

such as DeepVS which is a CNN based docking system that learns protein ligand interactions from large data data sets improving docking prediction. So, the advantage of this DeepVS is that it is an AI driven approach which enhances prediction accuracy and it adapts to new datasets using machine learning and it can be combined with scoring functions for better results better results. So, we can combine it with the classical scoring functions to get a kind of hybrid method to obtain the hits that have a high potential for having potent activity. So, if we summarize this session, So, the AI accelerates docking and scoring improving speed and accuracy in virtual screening and then we have generative models that enable us to design compounds with optimized properties. And we have hybrid approaches that combine structure and ligand-based methods for better prediction.

The beautiful thing is that cloud and distributed computing enhance scalability for large-scale screening. So now it is possible to screen even billions of compounds within a very short period of time using all these AI-based virtual screening methods. So, then I have an open question for you: how can the integration of AI-driven docking like G-NINA and AI-powered ligand-based screening like PyRMD enhance the efficiency and accuracy of the virtual screening pipeline in real-world drug discovery projects? So, these are some of the important publications that you can go through to get more details about this topic, and these will be very helpful for you to gain knowledge about this area. And with that, thank you.