AI in Drug Discovery and Development

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

Week-05
Lecture-24

Welcome to the course "AI in Drug Discovery and Development." Today, we will discuss the workflow of high-throughput virtual screening. So, in this session, you will be able to understand the role of chemical libraries and databases in drug discovery and how to curate a data set for virtual screening. Learn the step-by-step workflow of high-throughput virtual screening, either structure-based or ligand-based. Also, explore the key tools and techniques used in high-throughput virtual screening to identify potential drug candidates efficiently. So, we have seen earlier that virtual screening is becoming one of the popular methods to identify hit compounds.

And those hit compounds are essential to identify molecules that can engage the targets, and those hit molecules can further be converted into lead molecules. And then those leads are developed into drug candidates. So, the first thing that we need for virtual screening is actually the chemical space. And chemical space is, you know, the commercially available chemical compounds, the libraries which one can screen virtually.

And then can purchase compounds that can be screened in physical assays. If you look at the size of the chemical libraries, they are growing exponentially. So, if you look at, for example, the approved drug database, which is in the range of, you know, a couple of thousand compounds, but if you look, there are libraries like Wuxi Virtual, MQL Ultimate, LHMS in stock, E-Molecules Plus, Scooby Doo, or PubChem. And all those libraries are in the size of millions of compounds. And then you have, for example, zinc-22.

And then we have all these, you know, Millipore-Sigma SA-Space, Voxigalaxy, Enamin-Real Space, Otawa, Ally-Lele, and then you have Ambinter-Ambrosia. So, you have e-molecules to explore. So, these are billions of molecules. These libraries contain billions of molecules. And you can compare it with the stars in the Milky Way galaxy; there has been a search in the size of the library.

So, for example, there are now libraries like a GSK 2020 double Excel. So, which contains even, you know, molecules in the range of 10 raised to the power of 26. So, you can just think about it like how big the chemical space is and how many compounds there are that are virtually or even not only virtually. but which are physically present in this world, actually. So, those compounds are one of the sources of hit molecules for any new drug target.

And in virtual screening, this is one of the major assets where we look to identify the hit molecules. So, once we have identified a library like the one, we wanted to screen. So, the next step is the curation of the library or the curation of the molecules in that library. Usually, all those commercial libraries are very well prepared, taking care of all these, you know, problems that a molecule can have. So, if you wanted to prepare your own library for virtual screening.

So, then we usually need to fix some of the problems with those molecules. Those can be, for example, the list of smiles that is the structure of the molecule. So, we have to remove the mixtures and inorganics from those molecules. And then we have to, you know, convert them into, maybe you wanted to have them in canonical SMILES, or you wanted to have them in SDF or MOL2 files. So, we have to, when we are converting this structure, remove the salts, clean the structure, and then normalize the specific chemotypes.

For example, the nitrogen group, for example, the nitro group can be represented in different ways, actually. So, if we wanted to use those molecules containing a nitro group in the library. So, we need to make sure that all the nitro groups in all the molecules are represented in a similar way. So, we need to specify, we need to normalize those specific chemotypes, and then we need to remove the duplicates because There might be one molecule that is present in the salt form in the library and without the salt form as well. So, when we are removing the salt, we will have duplication of that molecule, and that duplication needs to be removed.

Usually, we do it by, you know, similarity search methods; we compare the similarity of molecules and remove the duplicates. And finally, we need to manually inspect the molecules to determine whether all these changes we have made are correct or not. However, it is not possible in the case of when you are using a million-molecule-size library or an ultra-large library with billions of compounds. But if you are procuring, you are curating a small dataset. So, these are the steps you must perform in order to get a reliable data set.

So, this was just the chemical curation. And then the next step is the duplicate analysis, and in the next step, we have to analyze the intra- and inter-lab experimental variability. For example, we have now removed the duplicate structures, and if there is biological activity or any other physicochemical data associated with these libraries, we will consider these molecules. For example, we can talk about solubility. So, if there is solubility data associated with these molecules, So, we need to take care of the intra- and inter-lab experimental variability in the solubility data, and we need to fix that, actually.

And then we also need to exclude the unreliable data sources because many times all this data is curated from the public literature, which can be in the form of, you know, patents or research articles published in different journals. So, it might be that some of the data is not, you know, supported by the outcomes in that paper, or it might not be reliable, actually. So, in that case, we need to exclude those unreliable data sources; otherwise, what will happen is they will contribute to the errors in the libraries whenever we are using them for modeling purposes. And then another important thing is that we need to detect and verify the activity cliffs. Activity cliffs are a very important concept, and what it says is that if there is a molecule, a very small change in the structure can show a very big change in bioactivity.

So, that is called an activity cliff. For example, there is a benzimidazole derivative. And then that benzimidazole derivative has a very high affinity for the target. We can take the example of acetylcholinesterase. And if we change, if we make a very small change in that molecule, like if we just put a methoxy group in the fifth position of the benzimidazole ring.

So, the introduction of that methoxy group is leading to a complete loss of activity. So, that is called an activity cliff because a very small change led to the complete abolishment of the activity. So, we need to take care, detect, verify, and sometimes remove those activity cliffs because they ultimately lead to problems in the modeling. And then we also need to calculate and turn the dataset model ability index, and in the end, we have to use consensus QSAR prediction to curate mislabeled compounds. Sometimes it happens that a compound, for example, is not soluble, but somehow it is, you know, reported as a soluble compound.

So, that can also be fixed by using a consensus QSAR prediction so that we can remove those mislabeled compounds as well. And finally, what we get is, you know, a curated set, and you can see that from the original set of compounds, we got curated steps after all these processes. And then what you can see here is that on the x-axis, you have the data set size, so by removing all those and doing all these steps. Removing molecules at each step will reduce the size of the library a little bit. However, we will reduce the error rate tremendously.

So, that should actually be our purpose because if our data is not, you know, good. So, whatever we do with this data, you know we will not get something useful. In conclusion, data curation is, you know, a very important step, especially in virtual screening. Okay, so here one of the examples is shown of how or what kind of problems you can face. For example, there are two compounds, compound A and compound B.

Their PubChem IDs are different; compound A has a PubChem ID of 5219, while compound B has 123596. So that these compounds are not duplicates according to the PubChemID. If we look at the SMILES as well, these compounds are not duplicates. If you look at the chemical name as well, So, this compound is registered as imatinib, and this one is registered as Gleevec. So, these are not duplicating even if we look at the INCHI key.

So, these compounds are, you know, different according to the INCHI key. However, if we look at the structure and then do a 2D similarity from the curated structure. So, we come to a point where we can see that both of these compounds are exactly the same. However, they are registered with different IDs and all those properties are different. So, this is why this one example can help you understand the importance of all those steps in data curation to obtain very good data.

Okay, so now let us have a look at the overall workflow of these ultra-large virtual screenings. So first, we will discuss the structure-based virtual screening workflow. And here it is not possible to cover all the tools because we have plenty of tools for doing all those ultra-large virtual screenings using structure-based methods. However, we will just focus on one of the tools called deep docking. So, it is a deep learning-based tool developed to accelerate docking based virtual screening.

So, usually what happens is when you want to screen millions of compounds using molecular docking. So, it takes a lot of computational power and a lot of time as well, and not everyone has access to, you know, supercomputers or GPU computing clusters. So, they can make use of a structure-based virtual screening like docking-based virtual screening. So, what this program proposes is that this algorithm says one can use deep learning-based QSAR prediction. And that is to predict the docking score of a small subset of the library, and then that model can be used to screen a large library containing maybe billions of compounds.

So, in this case, you can use any of the docking programs of your choice; it can be, you know, a Schrodinger or Glide. You can use the gold, or you can use the, you know, ICM Pro docking tool; you can use AutoDock Vina; you can use Dock 3.7. So, in order to prepare for DD virtual screening, the chemical library must be in this mile’s format. So, we need to download those molecules in the SMILES format.

So that we can download from maybe Zinc 20. And then it requires Morgan fingerprints of radius 2 and size 1024 bits for each molecule represented in a compressed form as a list of indexes of bits that are set to 1. So, and it works on the QSAR modeling protocol. So, what we do here is download the molecules, the library, and then we usually get a pre-prepared library from Zinc, for example. And the next step is the calculation of the fingerprints because those fingerprints are later used for developing this QSAR model.

And then it is also recommended to split the library of smiles into a number of evenly populated files to facilitate other steps, such as random sampling and inference, and place these files into a new folder. So, how it works is, for example, you have this ultra-large database; for example, we have a 1.3 billion compound library from the Zinc database. So, what you do is a random sampling of maybe 0.1 percent of these molecules, or you can even do 1 percent as well; for example, you can choose 1 million molecules out of it.

Randomly sample 1 million molecules, and then take those 1 million molecules and dock them into the binding pocket of any of your targets, using the docking score and their fingerprints. So, what you do is split them into validation, training, and test datasets and make a prediction model. And that DNN-based prediction, or you can say deep learning-based prediction model, QSAR model, and then that QSAR model you use to predict the, you know, docking score of this ultra-large library. So, this is iteration one. So, all these steps are followed, and then from this prediction, we get initial virtual hits, which are used for iteration two and the following iterations.

So, in the following iteration, you can see that once we get those molecules from virtual hits from this ultra-large library using the model. So, then we again do a random sampling on these updated virtual hits, and then we dock a small number of molecules, like 1 million compounds, again. And then we use these 1 million compounds to improve the model performance based on their docking scores. Again, this model, when it is obtained, has improved. So, it will again be used to predict the docking score of this ultra large library and update the virtual virtual hits.

So, we repeat this until we get a handful of compounds, maybe, for example, 1 million or 5 million, which you can handle easily. So, then these 1 million compounds can be used to identify top-scoring molecules using further molecular docking. So, this is how you are reduced: instead of docking 1.3 billion compounds, you are just docking maybe a couple of million compounds and then trying to identify hit compounds from this dataset.

So, this is actual work that we did recently. So, here you can see that we use this ultra-large library of zinc containing 1.3 billion molecules. So, we sampled around 900,000 compounds. which were split into training, test, and validation sets of around 300,000 each. And then those compounds were docked into the binding pocket of the enzyme called choline acetyltransferase.

And then the docking score that we obtained from this docking was calculated using AutoDock Vina. And then the docking score we obtained from this docking was used along with the fingerprints of these molecules to create a deep learning model. And then, by using this deep learning model, we predicted the initial virtual hits, and it was around 161 million compounds. And then we did the random sampling of 300,000 compounds again and repeated iteration 2, where we got around 3.

7 million compounds. Furthermore, we completed iteration 3, which gave us around 168,000 compounds. So, those 168,000 compounds were pre-filtered using, you know, different ADMET rules, and then we got around 94,000 compounds. And those were further docked into the binding site of choline acetyltransferase, and we selected the top-scoring compounds that had a docking score below minus 12 kilocalories per mole. So, we got 3,049 compounds, and then those compounds were further subjected to a physiochemical properties filter. Where we followed MMGBSA binding free energy prediction and molecular dynamic simulation, we identified the top 5 compounds as hit compounds for choline acetyltransferase.

Here you can see the AUC curve for all three iterations, where you can see that iterations 1, 2, and 3 have a perfect AUC of nearly 1. And this was the final result, where you could see that we identified these 5 hit molecules as potential inhibitors of choline acetyltransferase. And these are further evaluated for their in vitro activity as well. Okay, so that was about structure-based virtual screening. So now let us discuss the ligand-based virtual screening workflow.

So the ligand-based virtual screening workflow works in a somewhat similar way, but here the beauty is that we can handle a very large number of compounds. and we can do the screening in a very small amount of time. Because mostly those you know, ligand-based virtual screening tools work on the principle of similarity calculations. So, we have a library in which we compare the similarity of the known active compounds with the screening library compounds. And then we select the compounds that are showing very high similarity and propose that those are potentially active compounds.

Today we will discuss this tool called PyRMD, which is a fully automated tool that utilizes a random matrix discrimination algorithm and high-performance AI methods specifically tailored for the identification of new ligands. So, what it can do is easily screen millions of compounds in a very short period of time. It can also be automatically trained using the ChEMBL data set, which is where you get the biological activity data from. So, we need to get the structure and its respective bioactivity data for known acetylcholinesterase inhibitors. So, ChEMBL is a very large database for obtaining all that bioactivity data.

So, you can see here, for example, it has around 2.5 million compounds with, you know, 1.7 million assays. So, what we do is first we go to ChEMBL and then we download the required data set. So, what we need can be any kind of activity, okay. So, you can see that this data set contains the bioactivity data as well.

So, the next thing is that once we get the training data set, we also need to get the decoy compounds. And decoys are the molecules that are structurally similar to your active compounds, but they are inactive in nature. And that is, you know, to train your model efficiently so that the model can differentiate between an active compound and an inactive compound. So, they are structurally quite similar, but they are, you know, inactive in nature.

So, then you can see DUD.E is one of the famous you know database for identity for getting the decoy compounds. And then it has a number of targets where you can, you know, get the decoys; for example, this is the adenosine A2A receptor, and there are around 3,000 substances that are decoy in nature for this target. And then these are curated by hand, which actually means they are manually curated; there are also auto-curated data sets for different targets. So, in this case, for example, we download the decoys for acetylcholinesterase (ACEs). So, after getting those decoy compounds, these decoy compounds will be used for training the model, actually.

So, once we get the decoy compound, what we do here now is. The first step is that we have this training DB which contains the bioactivity and the structure. So, from ChEMBL, we get the structure as well as the bioactivity. So, it can be either in the form of a CSV file or in the form of a SMILES file, and then once we get this file. So, we need to prepare this database; we need to separate these molecules because in ChEMBL, we do not have information on which ones are active and which ones are not. So, we will just get the, you know, the bioactivity, for example, the IC_{50} values.

So, some of those molecules might have like 10 nanomolar IC_{50}, while another molecule might have 10 millimolar IC_{50} values or even like 100 millimolar or 1 molar. So, now what we have to do is classify them into active and inactive; only then can we train the model. So, the next step is the database preparation, where we separate them into active, inactive, and discard the compounds. So, what we usually do is, for example, we take a cutoff of about 1000 nanomolar. So, if the IC_{50} is below 1000 nanomolar, we will consider it an active compound.

And if the IC_{50} value is, you know, above 100 molar or 100 millimolar, then we will consider them as inactive compounds. And the in-between compounds, we just discard them, okay? Because we have to segregate them into actives and inactives, only then can we train the model. So, once we do that, we also remove the duplicates from the prepared DB. So, the prepared DB is, you know, your database that you wanted to screen, actually. And then this prepared DB can be from anywhere; it can be a zinc library, or it can be from, you know, Enamine.

It can be from e-molecule, mcule, vitas, or asinex, and the next step is the featurization; in featurization, the first step we do is the SMILES standardization. And then smile standardization means that if the compounds are, you know, not in non-canonical SMILES, we have to convert them into canonical SMILES. And then the next step is to calculate the fingerprints because those fingerprints will be further used to make a prediction model. Okay, so we use, for example, ECFPs, FPFPs, or MHFPs.

And then here we use the decoy compounds and the decoy database as well. Okay, so then further we use the RMSD classifier, perform the repeated stratified k-fold cross-validation, carry out the fitting process, and conduct the test set classification. And there are several matrices that we use to evaluate the model's performance. Like we use the true positive rate, false positive rate, ROC, AUC, all these bedrocks, all these things, all these metrics we use to evaluate model performance. Okay, so here you can see this is one of the examples that we ran. So, you can see here the benchmarking results where the TPR rate, which is recall, is 0.442, the FPR, or false positive rate, is 0.010, precision is 0.651, F score is 0.526, and ROC AUC is 0.890. And then in the second fold, you have got those, you know.

So, you can see the precision got a value of 0.996, the F score is 0.604, and the ROC is 0.875. So, you can see that this model is working quite well, and it can differentiate between active and inactive based on the active and inactive classification that we did on the training dataset. So once the benchmarking is done, we get all these, you know, precision-recall curves as well.

So if it is closer to the top right corner, then it is considered good. And the ROC curve, in the case of ROC, shall be closer to the top left corner. So, it shall be like this. So, it is considered the best.

So you can see here that the model has an AUC value of 0.87. So, once we get a good working model, the next step is the screening, actually. So, in the screening mode, what you do is use the training database, which was obtained from ChEMBL, either in CSV format or in the SMI file. And then you prepare the database for active, inactive, and discard the compounds. Then you have to remove the duplicate compounds as well. And here you again use the prepared DB file, which is the database you wanted to screen.

So this database can contain millions of compounds; this prepared DB is what you wanted to screen, actually. And after doing that, you again do this featurization, smile standardization, fingerprint calculation, and use the decoys, and then you use the RMD classifier. Do the fitting process and screen the database classification, and predict the molecules as either actives or inactives by using the RMD score from this prepared database. And this is, for example, an output from the screening process. So, here you can see the screening output file, which contains the SMILES of those molecules, includes the titles of those molecules as well, and then you have the RMD score and the similarity.

Then you have the most similar compound in the ChEMBL, and it also flags a compound if it is a pain compound, meaning if it is going to have the property of a pan assay interference compound. So, it will also flag it, so in the end, you get the molecules that are potential hits against this target, the target of your choice. Okay, so this is again another example where we use this technique to screen ligands for, you know, tau binders. So, we used, you know, a training data set of active tau inhibitors, tau aggregation inhibitors, we can say, from ChEMBL. And then we used a library of 12 million compounds from ZINC, which was, you know, available for purchase.

So, commercially available compounds are available. So, we used PyRMD to identify 8915 active compounds using this approach, and after removing the duplicates with RDKIT, we got around 2367 ligands. Meanwhile, there were, you know, ADMET filters that were also used. So, after getting this many compounds, we docked them into the tau monomer binding pocket. And this tau monomer binding pocket we obtained from replica exchange molecular dynamics, where we used a random coil structure of tau from a crystal structure, and then We subjected it to replica exchange molecular dynamics to obtain the stable single tau monomer three-dimensional structure because tau is, you know, an intrinsically disordered protein.

So, it does not exist in a single conformation of, you know, 3D structure. So, we identified this using the REMD, and then we used what we identified about the binding pocket using FTMAP, and then we docked those compounds in this binding pocket. and finally, we do the in silico ADMET studies using DruMap and then we further did the MD simulation of the top hits followed by MMPBSA calculation and we identified some of the compounds which are potential inhibitors of the tau aggregation. Now again, these compounds are in the process of being procured, and we will conduct the in vitro evaluation of those compounds against our aggregation. So, this is an example of how one can use, you know, ligand-based virtual screening to screen ultra-large libraries.

So here, instead of 12 million, we could have used about 1.3 billion compounds as well without any issue. OK, so coming to the summary. So, automation in high-throughput virtual screening speeds up the screening of millions of compounds while reducing manual work. And AI integration improves accuracy by using smart docking algorithms and machine learning to find better drug candidates. And AI also helps reduce false positives and continuously improve screening methods through adaptive learning.

And parallel computing enables faster virtual screening by using cloud-based platforms to handle large datasets efficiently. And it also allows multi-target screening, making it possible to evaluate many drug-like compounds at the same time. So, there are these papers. So, this first one is for the deep docking algorithm which has been developed by the Sherkov deep docking protocol.

And then you have the PyRMD paper developed by S. Cosconati. So, you know, you can go through these papers, and you can learn more about these protocols and workflows. And in the end, I have an open question for you. So, how does automation in high-throughput virtual screening improve the efficiency and reliability of drug discovery compared to traditional screening methods? So, you just like ponder over it a little bit. And with that, thank you.