AI in Drug Discovery and Development

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

Week-05
Lecture-21

Welcome back to the course "AI in Drug Discovery and Development." So, today we will be talking about the introduction and approaches to virtual screening. By the end of this lecture, you will be able to understand the concept of virtual screening in drug discovery. Know about various approaches to virtual screening. Explore the use of hybrid approaches that combine the best features of SPVS and LBVS. And also learn about the tools and strategies commonly used in virtual screening workflows.

So, we have seen it earlier, as well. So, the method to identify hit compounds or to begin drug discovery is that once you start, once you have identified a target. So, the next step is to identify the small molecules or biologicals that can engage with their target. And then they can modulate that target.

So, the classical way is to do high-throughput screening. So, this is an automated large-scale experimental screening of compound libraries to identify active or hit compounds. And usually, it is done automatically using robots, micro-plate readers, and high-throughput assays. And what we do here is expose the compound to the biological target in parallel and measure the biological response using assays. Such as these could be fluorescence-based assays, absorbance-based assays, or enzyme activities, or they can be luminescence-based assays or any other kind of in vitro assays.

So, here you can see this flow diagram. So, you start with the compound library, and then you perform the in vitro assays. And then you identify the hit compounds, which are the compounds showing high activity or at least some activity in that assay. So, some of the challenges with high throughput screening are that they are expensive and require large compound libraries. Because if I want to screen around 100,000 or half a million compounds in high-throughput screening, I have to procure all those compounds.

And then I have to develop an assay system that can handle those 100,000 compounds. So, it will take a lot of time. So, it is time-consuming, labor-intensive, and also requires a lot of investment. It also has the chance of giving you high false positive and false negative results, which means it can identify compounds that will be false positives. It means that those compounds are not active, but they will be identified as active in this assay, or it will miss the compounds, which means a false negative.

So, it can also miss the compounds that are active. So, these are some of the challenges with the high-throughput assays. So, the alternative is to perform the virtual screening. So, the virtual screening is a computational technique that is largely used in drug discovery to identify potential bioactive compounds from large chemical libraries. So, alternative to high throughput screening if we have a large library and if we can if we have for example, this large library of compounds contains 100,000, 1 million, or 1 billion compounds, and then we can perform in silico screening using computer algorithms.

like machine learning or AI that we will be going through in this course, as well. So, can we identify hit compounds with high accuracy? So, that is the whole idea behind virtual screening. So, what we do is screen those small molecules against biological targets such as proteins to predict their binding affinity and therapeutic potential. So, the key advantages of this method are that it reduces the time and cost compared to wet lab methods like high-throughput screening methods. And you can expand the chemical space that can be explored because it is very difficult or challenging to screen millions of compounds in high-throughput screening assays.

But we can screen not millions but billions of compounds using virtual screening methods. And that is how it became a key component for accelerating early-stage drug discovery nowadays. So, if you look at the evolution of virtual screening, in the early days between 1980 and the 1990s. So, usually the method which were used to perform virtual screening was a simple rule-based filtering. For example, Lipinski's rule of five-based filtering, where the small molecule libraries are filtered using all those rules.

That the molecular weight of the compound should not be more than 500, and the hydrogen bond acceptors should not be more than 10. Hydrogen bond donors should not be more than 5; the calculated log P value should not be more than 5. So, all these are rules were used earlier and then the simple molecular docking with rigid structures were also utilized during that time. Another method was the structure-activity relationship models, which were derived after synthesizing analogues of a hit compound. And then those analogues are synthesized, deduced, and based on that SAR, for example.

At the first position of the ring, an electronegative electron-donating group or an electron-withdrawing group is required. So, on the basis of that these virtual screenings were performed in that time. And then, between 2000 and 2010, the advancement in molecular docking saw the development of many flexible docking algorithms, such as AutoDock and Glide. So those cloud servers provide you with the computational time, computational power, and all the software requirements as well. So it is just like you can connect those cloud computers to your mobile phone and perform the job.

And that is how it led to the development of high-throughput virtual screening, which means you can screen billions of compounds in a very short period of time. The integration of molecular dynamics for enhanced accuracy occurred where the molecular dynamics method was integrated into virtual screening methods to enhance the accuracy of the identified hit compounds. And large-scale chemical library screening, for example, of the ZINC, Enamine, or REAL databases, also happened during this time. And if you look at the post-2020 era, we can largely say that it is an AI and machine learning era. where the AI driven docking like DeepDock or deep docking or DiffDock or GNINA, so all these developments have happened.

And then, there are also generative models being developed for novel molecule design, like Re-Invent, which is a graph neural network-based generative model from AstraZeneca. And then, multi-omics and deep learning integration for target prediction has also occurred, which has increased the accuracy and the capacity of virtual screening methods. So, if we talk about the applications of virtual screening, it helps us in hit identification and lead optimization; it can rapidly help us screen large compound libraries to identify hits. We can also use it to refine the lead compounds by predicting binding affinity and their ADMET properties. Also, we can optimize drug-like characteristics using quantitative structure-activity relationships and molecular docking.

So, we can use virtual screening methods for drug repurposing, which is a way to identify new therapeutic applications for existing drugs. In the FDA-approved drug database, there are more than 1,500 drugs. So, those 1500 drugs, can we find another use for them? Can those drugs also engage other targets in our bodies and be developed as therapeutics for other diseases? So, that is the idea of drug repurposing, and virtual screening can also help us identify alternate targets to expand drug utility as well. So, it means that one drug can have two or more therapeutic indications that we can achieve through drug repurposing. Then it also helps us in target-based drug discovery, where we can use it as a structure-based virtual screening for precise binding predictions.

And the LBVS we can use for similarity-based compound selection. And then we can do the integration of docking, MD simulation, and free energy calculations to increase the accuracy and outcome of these virtual screening methods. And then we can use those virtual screening methods for screening natural products and for fragment-based drug design as well. So, it can be used to identify bioactive natural products for drug discovery and also for fragment-based screening to assemble optimized drug candidates. It can also enhance the efficiency and reduce the experimental cost.

Additionally, it can enhance the efficiency and reduce the experimental cost of identifying

the hit compounds. When we talk about, you know, virtual screening. So, the first thing we need is the ligand databases because we want to identify hit compounds. So, we have to screen some compounds that are commercially available, for example, or they can be custom synthesized. So, these are some of the ligand databases that can be used for performing virtual screening, the most widely used of which is ZINC.

which is a free database of commercially available compounds and contains around 1 billion compounds and then it has all those, you know, they have classified those compounds into drug-like, lead-like, fragment-like molecules. Or they can be, you know, in stock, or they can be purchasable; they can be like customs in size. So, they have all those; you can split that database into all these based on all these factors, and then you can pick the molecule. You can also cherry-pick the molecules based on the similarity of your desired molecule. And then we have ChEMBL, which contains the bioactivity data as well.

So, this is a database that contains molecules along with their bioactivity, and it currently contains more than 2.3 million compounds. So, the beauty of this ChEMBL database is that it is annotated with bioactivity. So that you can use it for making QSAR models, for example. So, you can download the molecular structure, and you can download the associated bioactivity.

And then you can use those bioactivities and those structures to make QSAR models, and then you can use them for screening purposes. Then we have PubChem, which is again an open database of small molecules; it contains around 100 million compounds, and it also includes the experimental and computed properties of those molecules. Then we have DrugBank, which contains molecules that are FDA approved for use as medicine in the clinic, and it also has investigational drugs, meaning the drugs that are under clinical trials. So, they might be under phase 1, phase 2, or phase 3 clinical trials, or they might be in the preclinical proof of concepts stage as well. So, it has around 15,000 compounds, and the beauty of the Drug Bank is that it contains a lot of information along with detailed drug-target interactions as well.

Then we have Enamine, which is the largest commercially available compound library, containing around 30 billion compounds. And then what you can do is have all those physically available compounds as well as the compounds that are virtual in nature, which can actually be synthesized on custom order. And then you have MolPort with similarly 7 million compounds and then Mcule, which contains high-quality purchasable compounds, somewhere around 500 million compounds. So, it is a vendor-curated library, which means they have aggregated multiple vendors into their database and then provide all those molecules. And then we have ChemSpace, which is another ligand database that contains extensive molecular space for drug discovery, with around 15 billion molecules.

So, the focus of ChemSpace is on the diversity screening libraries. Likewise, once we have obtained the compounds, the next step is that we need the target structure if we are using structure-based drug design or structure-based virtual screening. So, the Protein Data Bank is primarily a resource for experimentally determined 3D structures, including 200,000 structures of proteins and nucleic acids. So, it contains the 3D structures from X-ray crystallography, NMR, and cryo-EM. And then we have AlphaFold, which is, you know, a great thing where they have resolved the three-dimensional structure of around 200 million protein molecules.

And these are AI-predicted, deep learning-based protein structures from DeepMind, which is a subsidiary of Google. And then we have a Swiss model repository which contains user-submitted homology model protein structures. We have a mode base that contains comparative protein structure models with around 10 million molecules, and these are structural models based on sequence homology. Then we have PDB redo. So, it is, you know, similar to the Protein Data Bank.

So, in this PDB redo, what they have done is refine the Protein Data Bank structures. Sometimes the Protein Data Bank (PDB) structures have some problems; they are not resolved correctly. So, in PDB redo, they assign correct, you know, electron density based on the maps, and then they fit the model into the electron density and refine the structure. So, then it leads to improved resolution and model accuracy.

So, it has around 100,000 molecules. And then we have the OPM database, which is a membrane protein structure along with the orientation data on how the membrane proteins are oriented in the membrane. It contains around 2000 molecules, which are useful for studying all those membrane proteins like G protein-coupled receptors. And then we have SCOP and CATH, which are protein classification databases on structure. And then we have the cryo-EM data bank, EMDB, which is a repository for 3D electron microscopy maps containing around 25,000 molecules. It provides a volumetric map of large biomolecules where the structures are resolved by 3D electron microscopy.

So, now we have seen how we can obtain small molecule ligands and then get the protein structures or the target structure. So, next is what kind of virtual screenings we can use. So, there are basically two types of virtual screening: one is called structure-based virtual screening, and the other is called ligand-based virtual screening. And then the third one is actually a combination of both of them called a hybrid approach, where we use both structure-based virtual screening and ligand-based virtual screening techniques. So, okay then, let us have a look at the structure-based virtual screening methods.

So, basically, in structure-based virtual screening methods, as we saw earlier, we use a 3D structure of the target protein to predict how potential ligands will bind to it. So, it is mainly a green docking-based screen where we use tools such as AutoDock, VinaMPI, Glide, Gold, FlexX, etc. So, there are a plethora of those docking-based tools, actually. So, what we do is take the building blocks and make the on-demand libraries. And then we use the target structure, which is obtained from X-ray, either from X-ray, from AlphaFold, or from cryo-electron microscopy.

And then perform the molecular docking and evaluate these top-ranked molecules. So, you know, after experimental evaluation, we will be able to identify the hit compounds. So, in this, we use molecular docking, and molecular docking is a structure-based virtual screening method that predicts the preferred binding orientation of a ligand within the target binding site. So, it uses a scoring function to estimate binding affinity, and some methods can include either rigid docking, where the receptor is kept fixed. and the flexible docking where you can have flexible side chains of the receptor as well.

And the key software we have seen are Auto Dock, Glide, and Gold Dock. So, these are some of the key software that are being used. Then we can use refinement techniques to enhance the docking results, such as by using molecular dynamics simulation. So, which can assess ligand stability, flexibility, and binding mode refinement over time. And then we can also use free energy calculation methods like MMPBSA, MMGBSA, or free energy perturbation methods, which can more accurately estimate the binding energy post-docking.

So, then we come to the similarity searching, which is basically a ligand-based virtual screening method. So, what we do here is identify structural similarity between compounds based on molecular fingerprints; we use either 2D or 3D molecular fingerprints. And then we retrieve the similar compounds using indices such as Tanimoto for similarity or other indices. So, you can see here in this example that we take a query compound and then we calculate the features, which are basically fingerprints.

So, you can use all these ECFP, FP2, MH, and FP6. Or we can even calculate and determine the pharmacophore. We can determine the, you know, like maximum common substructure of the scaffold, or we can determine the 3D fingerprints, like this E3FP, or we can determine the electro shape. And after converting these query compounds into all these features, we use these features to compare their features with the features of the library. So, that is our, you know, screening library, and by using all these indices, such as the Tanimoto coefficient or Soergel distance or Tanimoto dissimilarity coefficient, Soergel distance-1, Euclidean distance, Manhattan distance, Dice coefficient, Tversky, or cosine. So, the idea is that you compare the similarity of your query compound with the target molecules.

So, the idea is that you determine the similarity of your query compound with the screening library compounds and whether those two molecules have, you know, similarity between each other. So, then you identify them as hit compounds. So, then finally, you get the hit compounds, and those can be tested and validated in the in vitro assays. The next ligand-based virtual screening technique is pharmacophore modeling, where we identify the essential molecular features required for biological activity. We have seen that we have also studied earlier that a pharmacophore is basically a three-dimensional spatial arrangement of functional groups that are responsible for biological activity.

So, we can define those pharmacophores using all those pharmacophoric features like H-bond donors, acceptors, hydrogen bonds, hydrophobic regions, or, you know, aromatic regions. And then this we can use for either ligand-based or structure-based virtual screening methods as well. So, this is a workflow for pharmacophore modeling in ligand-based virtual screening. So, what you do is take a database of non-active ligands and then explore the common chemical features. You generate a pharmacophore like this, and then you compare this pharmacophore with the, you know, then you validate.

And then refine the pharmacophore to generate the final model, and then use this final model to screen a database of maybe 100,000 or 1 million compounds. And then the idea is that you identify molecules that have a similar pharmacophore to this query molecule and the query pharmacophore. And once we identify similar pharmacophore-containing molecules, we pick those molecules, experimentally validate them, and then identify the hit compounds. And then the third ligand-based virtual screening method is, you know, quantitative structure-activity relationship, where we develop a mathematical model linking the molecular structure to the biological activity. So, it uses either statistical and machine learning techniques or random forest and SVM; this we will see during the course.

So, then we can convert those molecules into descriptors like physicochemical, topological, or quantum mechanical properties, and we can make a correlation model or regression model using those descriptors. Those features and this as the structural representation of the molecule and the biological activity, which can be any activity that we are actually interested in. So, we develop a model, and then we use that model to screen the molecules. So, this is the workflow of the QSR modeling in LVBS. So, we take the chemical database, and then we generate the QSAR models.

And then, by using this QSAR model, we screen the large chemical library. We predict these properties of those large chemical molecules in this large chemical library, and then we identify the hit compounds; again, those hit compounds are evaluated in the in vitro assays and in vivo assays. And then this chemical database, you know, is improved by

adding that in vitro and in vivo activity into this database. So, we are enriching their chemical database, which can be further used to improve the QSAR model as well. So, then we have these hybrid virtual screening methods where we can use docking-guided pharmacophore modeling, utilizing the docking results to define pharmacophoric features.

So, you can generate a pharmacophore without always needing the ligand. So, what you can do is if you know about the binding pocket of your receptor, enzyme, or target. So, you can actually generate a pharmacophore from the target as well. So, this is how we use docking-guided pharmacophore modeling, where we use the docking results to define pharmacophoric features, like PharmaDoc, which integrates molecular docking with pharmacophore modeling.

Then we can use the machine-learning-assisted docking as well. We will see during the course that we are using AI model filters or ranked docking results to enhance accuracy, like we use DeepDock, Deep Virtual Screening, Deep Docking, or G-Nina. And then we can use QSAR-enhanced molecular docking, where we use the QSAR models that prioritize the molecules before docking simulations. And then we can use the consensus scoring application, where instead of using a single scoring function to identify the hit molecule We can use multiple scoring functions and bring a consensus based on those multiple scoring functions. So, it combines multiple virtual screening techniques like docking, QSAR, similarity searching etcetera. And then, in an example, you can use docking and fingerprint similarity to improve hit selection or hit identification.

And then, finally, AI-driven hybrid screening, where we can use deep learning models trained on structure-based and ligand-based virtual screening datasets. So, we can use the DeepChem library with generative models that can generate structures for docking. If we compare these virtual screening techniques, such as the structure-based virtual screening methods like the docking method. So, the data requirement for the docking method are.

So, we need the structure of the protein in this case. The computational cost is a little bit high; that is why the throughput is moderate, but the accuracy is actually high. So, scalability is also limited, and the experimental data dependency is low. So, if we compare all these virtual screening techniques based on the data requirement, computational cost, their throughput accuracy, scalability and experimental data dependency. So, the docking a technique which is you know structure based virtual screening method.

So, it relies on the protein's structure. So, we need the protein structure in our hands to perform docking, and computationally it is quite intensive. So, we need a lot of computational power if we want to screen, for example, 1 million compounds. So, we need a lot of computational power for that, and that is why the throughput is moderate. But the

accuracy is high.

So, there is always a balance between accuracy and throughput. So, if the accuracy is high, then throughput will always be on the lower side, and scalability is also limited in this case. And then it will be, and the dependency on external experimental data sets is low because we do not need any, you know, for example, experimentally derived binding affinity or all those properties. And, if we look at the pharmacophore-based virtual screening, which is a ligand-based virtual screening method. So, what we need is the ligand structure or at least the features, and the computational cost is medium while the throughput is very high. So, we can, you know, go for screening billions of compounds using this pharmacophore-based modeling.

And then the accuracy is moderate, scalability is very high, and we need some experimental data. For example, we need the three-dimensional structure of the ligands bound to the receptor for the best case because then we know exactly how the ligand is binding to the receptor in its binding pocket. And if we go for the similarity-based searches. So, we need the structure of the compounds again from the chemical library, the computational cost is low, the throughput is very high, the accuracy is moderate, and the scalability is high. The experimental data dependence is high because we need the compound structures, and then we need to calculate those features from those compound structures as well.

Then, if we talk about the QSAR, we need the ligands as well as the ligand descriptors, which we can calculate as well. So, the computational cost is medium, throughput is moderate, accuracy is high, and scalability is moderate. But experimental data dependence is very high because we need the associated biological activity or the physicochemical properties. Then only will we be able to make the QSAR model. And if we talk about hybrid models, if we combine all these, you know, ligand-based and virtual screening-based models.

So, we need to have, you know, a lot of data required, like we need the protein structure as well, we need the ligand structure as well, and we need the descriptors as well. And then, if the computational cost is high, the throughput can depend on which two methods we are combining or which three methods we are combining. But the accuracy is high compared to the individually used methods, especially the ligand-based virtual screening methods, and the scalability is high. and then experimental data dependency will varies based on which techniques we are combining.

So, in the end, if we summarize this session. So, virtual screening accelerates drug discovery by computationally filtering large compound libraries. And then, structure-based

virtual screening relies on protein-ligand interactions; largely, it involves molecular docking methods. And then, the ligand-based virtual screening identifies hits based on the known active molecules. So, we have different techniques like similarity searching, quantitative structure-activity relationship, or pharmacophore modeling. And then there are hybrid methods that can integrate structure-based virtual screening and ligand-based virtual screening to improve the accuracy and efficiency of virtual screening.

So, you can go through all these, you know, nice publications that are seminal in their nature. So, you can go through them, read more about virtual screening, and learn more. And in the end, I have an open question for you: how could AI-powered virtual screening change the future of drug discovery? And with that, thank you.