AI in Drug Discovery and Development

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

Week-12
Lecture-60

Welcome to the course "AI in Drug Discovery and Development." This week, we have seen several hands-on sessions, where in the beginning, we saw how we can use RDKit to represent a molecular structure and also calculate different properties. Like molecular descriptors or fingerprints, all sorts of, you know, calculations with the molecules we saw. And then in another hands-on session, we saw how we can use, you know, regression-based modeling for predicting the solubility of molecules. So, we trained a model and then we used that model to predict the solubility of new molecules as well. In another session, we saw how we can use classification models to predict the bioactivity of a compound, whether it could be active or inactive against certain targets.

So, we trained a model, and then we predicted the bioactivity of new molecules. And then, in another hands-on session, we saw how we can use, you know, pharmacophore-based screening. So, that is what we saw, and in continuation of that, in today's session, we will see how we can use similarity-based screening tools. So, I will show you two tools: one is SWISS similarity, and the other one is Smart Small World.

So, I will show you both of those tools, and we will see how we can use them to screen based on the similarities. So, how can we screen ultra-large libraries for identifying molecules that are similar to our input query? So, what we have to do is visit this link first; we will go for the Swiss similarity, and then we will use another link, swdocking.org. So, let us open this in a browser. Once you open it in a browser, you will end up on this page, Swiss similarity, which is, you know, from the Swiss Institute of Bioinformatics, actually.

So, they have developed several tools like SwissDoc, SwissParam, SwissSidechain, SwissBioISOsteer, SwissTargetPrediction, SwissAdme, and then, you know, SwissSimilarity. So, you have seen all these tools. There are multiple things you can do. You can do the docking, and you can do calculations. Your molecules, and then you can, you know, predict the side chains of your structures, which you are, you know, generating from homology modeling or some other tools.

And then you can, you know, identify the bioisostere replacements as well, like which parts of the molecule can be replaced with some other bioisosteres. You can make this target prediction as well, such as whether a molecule will be active against a target or not.

Or you can, you know, predict the pharmacokinetic properties, which is where we use the Swiss ADME and then Swiss Similarity, which we will be exploring today, right? So, in Swiss, similarity is pretty simple; actually, what you do is just. Enter the SMILES of your molecule here in this cell, and then select the class of compounds. So, for an example, we will use furosemide as an example.

So, what you can do is go to PubChem, and then you can get the structure in SMILES format from, you know, PubChem. This you can do for any molecule you know, any molecule of your choice. So, we go back to this with similarity, and then you can enter the SMILES here, right? Otherwise, you can always draw your molecule using the sketcher as well, but I would not suggest that because sketching is cumbersome and laborious. So, it's always better to go for the SMILES that are easy, and you can get them from, you know, those databases like PubChem. Okay, so the next thing is that we are using the SMILES, and this is our query, actually.

Then we are going to search this against some databases, and here are the databases to select from. So, in this step, we have to select a class of compound. OK, so here, for example, there are multiple possibilities. So you can either search for it in the drugs database where, you know, all those drugs are in this database. Or you can use, you know, a bioactive dataset that contains molecules likely to be bioactive.

And then you can, for example, search in the commercial space as well. So, that is a very huge space, and then you can search in the space synthesizable as well. So, those are, you know, both commercial, but they can be custom-made, they can be synthesized, and most of them are coming from, you know, the Zinc database, actually the Zinc 20, which is a large collection of molecules. So, for simplicity, we will just go for the, you know, the drugs, and then in the next step, what you see here is you have kind of different types of fingerprints, actually. For the methods you wanted to search for your molecule.

So, for example, you have the FP2, which is a fingerprint type of fingerprint, actually. And then you have the ECFP eccentric connectivity fingerprint, which is ECFP4; you can use the MHFP6, and you can also perform pharmacophore-based searches as well. Or you have, you know, scaffold-based search, generic scaffold, and then you know you can even have 3D search as well, and you can combine 2D and 3D together as well. For example, we will go just like ECFP 4, and then we are searching only, you know, the DrugBank. You know where we selected a class of compounds; that is, we selected drugs.

So, it has three possibilities: one is the drug bank, another is the ChemBL approved drugs, or the ChemBL drug candidates in the clinic. So, if we change from here, for example, if we change from drugs to, you know, bioactives. So, then you have other, you know, libraries that are there. So, you can have, for example, you can search this library or the ChEMBL library or, you know, all these ChEMBL libraries, and then if you collect on synthesizable. So, then you can you know search this Inamine or Innova_pharm you know databases.

So, as I said, I like to keep it simple. So, we will just go for the drugs. So, we search for the search in the drug bank using the ECFP4, and unfortunately, you cannot select multiple methods. So, you have to just go for one at a time, actually. So, we just select the ECFP4 fingerprint 2D fingerprint method and then the DrugBank, okay, and then we just click on start screening.

So, once we click on start screening, it will open it in another tab. So, once we do that, you can see that these were the run parameters. So, we selected the ECFP4, and then the library we screened is the molecules from the Drug Bank, and this was our query molecule, which is furosemide. Then, these are the results. Okay, so now you can see that a score of one means it is completely similar to our query, and then you can see that it is furosemide as well, which means that the tool was able to identify a similar molecule to this, actually.

So, this is you know now it has successfully identified furosemide from the database. And then you have other molecules as well which have been identified. So, those molecules are like, you know, the score is 0.3, for example, and then the score is varying. So, it means that those are not very similar.

So, if the score is close to, you know, 0. So, it means that compounds are not similar to the query, and if the score is close to 1, it means the compounds have high similarity. So, you can see that it has screened several compounds; we can see that along with the score as well. This is one of the simplest ways by which we can do the screening, and then we can screen all those libraries. So, this is how switch similarity works, okay.

So, the other tool which we were talking about is the small world you know tool. So, which uses another approach where they use the graph-based, you know, similarity search. So, where they split the molecules into smaller graphs and then they compare the similarity of those smaller graphs with the query molecule. So, when you open this link, you will end up on this page, and here you have the possibility to draw the structure. You can also just copy the SMILES from the PubChem database as well.

So, we will do the same. So, we will go to PubChem, and in this case, we can take the example of aspirin, a simple drug. So, we just go to the SMILES and then copy the SMILES. Okay, and then come back to the SW small world and paste that SMILES into the identifier, okay. So, now you can see that it got identified, and it immediately starts, you know, searching the molecules. So, you can see that you have here showing 1 to 6 of the top 24 entries, right? And then you can see that it has searched them to 0.76.

And what we have used here is the Enamine SC stock from May 2022 to search the library. The beauty of this method is that it is even faster than other comparative methods. So, it can actually do the search in milliseconds. So, for searching billions of molecules, it is very, very fast. Compared to the Swiss similarity, you have a large number of libraries here.

For example, you have this Mcule ultimate, which is like billions of molecules. You have the real database, which again contains billions of molecules. You have the Zinc 20, around 1.3 billion molecules, all Zinc in stock, in the former set: interesting, Mcule-V, Mcule, Mcule full, Mcule purchasable. Old in stock, weight okay, from the zinc, Wuxi; these are mostly like ultra-large libraries.

So, you are searching these ultra-large libraries using this tool in a very short period of time. So, what it gives you is the identified compound, and then it gives you the distance. Like these, all are exactly the same; these are exactly similar compounds to the query. So, that is why you see 1, and then the ECFP is 0.61, the daylight similarity is 1, and then the unknown distance is 0; likewise, you can see all these properties actually.

And then, not only that, you can always change the search type from not only solver technology for similarity searching. But you can use, for example, maximum common substructure, substructure, superstructure, the Bemis-Murco framework, or the element graph as well. So, these are, you know, all the possibilities you can do with this search wheel, and then you can have these advanced options as well. Where you wanted to control the distance, the anon distance, the terminal, the rings, the linkers, mutations, substitutions, and everything.

So, you can do all that as well. And then there are, you know, the scoring options where, for example, you can do the multi-source scoring, where multiple component queries indicate multiple start points. And then you can do the search for the top only, where only the top results of the first page will be retrieved. And then you can use different scoring methods like atom alignment, where you will align and score each hit related to the query atom types, and the differences are categorized as may, min, HYB, and SUB. And then you can do the SMARTS alignment as well, where you treat the input as SMARTS and align and score each hit related to the query atom expression in the input. So, the important thing is that you have to turn off the atom alignment because you cannot use both of them together, actually.

And then you can use the ECFP4, which is an extended connectivity circular fingerprint, and you can also use the Daylight fingerprint, which is a path-based fingerprint. So, the advantage of this tool is that you can screen billions of molecules in a short period of time. And then, especially this small world technique that they have developed is very nice. And then you can get molecules that are, you know, similar, and you can also identify novel molecules as well. So this is a quick introduction to how we can use similarity-based screening tools, like the Swiss similarity and this small world similarity tool.

OK. So this is the time to wrap up the whole course. So, from the beginning to the end. So we have seen many things. So, in the beginning, we saw what drug discovery is and then how it works. We saw machine learning tools and techniques that are being used for drug discovery and development, and then we saw them individually as well, like how we can use AI.

So, then we saw how AI can be used for hit identification, for ADMET prediction, for lead optimization, and then we saw how we can use AI for even clinical trial design and optimization. And then in the end, we saw that we had several hands-on sessions where I showed you how we could use most of them, which were open-source tools. So how we can use those open source tools for making discovery and development easier. So I hope you enjoyed this course, and it will be followed by the examination. So I wish all of you the best and hope you will utilize the skills you have learned in this course for your future endeavors. Thank you so much.