AI in Drug Discovery and Development

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

Week-09
Lecture-44

Welcome to the course "AI in Drug Discovery and Development." In this session, we will talk about molecule optimization with generative AI. So, in the earlier sessions, we have seen what generative AI is and the de novo drug design method, and then how generative AI works. What are the different models being used for generating new lead molecules or drug-like molecules? So, in this session, we will discuss a couple of, you know, tools that are being extensively used for generative AI. So, by the end of this lecture, you will be able to explore the AI tools to generate drug-like molecules. while satisfying ADMET, binding affinity and synthetic feasibility constraints and integrate generative AI tools into the drug discovery pipeline.

So, what is molecular optimization? So, molecule optimization lead optimization is a very important step in drug discovery where, once we have identified a molecule known as a hit molecule, we can say. So, we try to optimize its properties. Those properties are actually many-fold. So, there are hundreds, or at least dozens, of properties that we need to, you know, optimize.

So, optimizing those properties while retaining the efficacy or potency is one of the big challenges. So, optimization is a vital phase in drug discovery where lead compounds or initial hits are systematically refined to improve their drug-like properties, including binding affinity, solubility, and overall bioavailability. So, this step involves intelligently navigating the vast chemical space to design molecules that are not only more potent and selective but also safe and synthetically feasible. So, the process is an iterative modification of the chemical structure. where the optimization focuses on potency, enhances the target binding, selectivity, and minimizes the off-target effects, As well as ADMET properties, improving the absorption, distribution, metabolism, excretion, and toxicity.

However, there are key challenges, like multi-objective optimization, where we need to optimize multiple properties at the same time. So, there are balancing trade-offs between multiple desirable properties, like if one property we are improving on, another might actually go down or get affected. And synthetic feasibility is an important challenge, especially in de novo generative modeling, where we design the molecules that are synthetically feasible. So, ensuring that compounds can be practically and efficiently synthesized in the lab. So, this we have already seen, like there we use the representational

learning where we use the smile string, which is a linear text representation.

Graph-based representations that capture the molecular structure or 3D structure, including spatial information. And the advanced techniques, like molecular motif graph neural networks or the continuous latent space representation. Here, the objective functions are the multi-parametric optimization for drug-like properties. The key objectives include improving biological activity, enhancing drug likeness as indicated by the quantitative estimate of the drug likeness score, and improving ADMET properties and synthetic accessibility. The challenge here is balancing the conflicting objectives and integrating expert knowledge with the chemical rules.

So, let us discuss some of the tools that are being recently developed and are being used for molecule optimization. So, drug assist is one such tool that is an LLM for molecule optimization. It is an LLM-based large language model interactive optimization tool that leverages LLM for molecular design, integrates chemical knowledge with natural language processing, and enables intuitive interaction between researchers and the AI. So, everyone has used ChatGPT, actually, so it is something like ChatGPT, which can even design or optimize molecules. Like you have one starting molecule, and you can ask it to generate a molecule that is, you know, more soluble.

So, it will generate a molecule, append some groups to it, and then it will generate a molecule that is more soluble, more permeable, or something like that. So, it enhances the experience of a chemist by integrating large language models as well as generative models, specifically generative chemistry-based models, which can generate new chemical structures. So, this is an example of the iterative optimization capability of drug assistance. So, when the model provides a molecule that does not fully meet the requirements, it can correct the error and generate a new compliant molecule based on human-provided input as well. For example, this is a molecule that was generated, which has a quantitative estimate of a drug likeness score of 0.56, meaning it is a good molecule to start with.

So, the user is saying how we can modify this molecule to increase the QED value by at least this much. So, then it optimized it generated a new molecule and increase the QED value, but then the user if the user is not satisfied with it. So, then it can go back again and it can tell that, okay, this is not correct actually, or I need something else, or I need, you know, a better molecule or another molecule. So that it can generate another molecule as well.

So, it is like you know interacting with an assistant that is kind of working on generating molecular structures. And then, this is an example of AISCC discovering the AntiIBD lead targeting the CXCR4. So, synthetically feasible de novo molecular design of leads based on a reinforcement learning model bridges the gap between the theory of de novo molecular

generation and the practical aspects of drug discovery. So, they utilize this; in this case, what the authors did was. So, they designed a library of, you know, molecules based on, you know, the starting materials, and they did it using generative AI.

And then they were able to identify as well as validate those molecules in the wet lab assays. Another tool is DrugHIVE, which is again a very good model that is a structure-based drug design model using a deep hierarchical generative model. So, it represents a paradigm shift in molecule generation. So, it surpasses current autoregressive and diffusion based models in both efficiency and output quality across the common benchmarks. So, the key feature of Drug Hive is that it uses a hierarchical architecture that enables fine-grain control over molecule generation.

and has superior speed and performance for scalable applications and is designed to handle multi-objective optimization scenarios with ease. And then the applications in drug design could be de novo molecule generation with the desired physicochemical properties. or lead optimization to enhance potency and selectivity in ADMET, scaffold hoping for novelty and patentability, and linker design in fragment-based drug discovery High-throughput substructure replacement for rapid SAR exploration. So this is how the DrugHIVE works.

Okay. And then you can see here, for example, it has all sorts of capabilities, like you can do the substructure modification. Like you have one molecule where you want to modify the substructure, a certain part of that molecule. So that can be done. It can actually be used for the linker design. So, you have two molecules, and then you want to, you know, link them by using a generative model.

So, it can generate the linkers and link these two molecules, and this is especially useful for PROTAC designs or molecular glues, which are becoming quite popular in recent days. And then it can also be used for the fragment growing, where you have a small fragment that you want to develop into a full-scale molecule or a kind of large molecule. So that can be done. So, this is typically useful in fragment-based drug discovery or optimization as well. And then it can be used for the pattern replacement where you want to replace part of the molecule or the molecule itself with similar molecules.

So, all sorts of capabilities it has got, and the most important thing is it is taking into account the protein structure as well. So, that is why it is a structure-based de novo generative modeling tool that can take into account the information regarding the binding pocket of the protein. And then, based on the ligand structure, it can generate molecules that fit well into the binding pocket. So, that is one of the beauties of this model. And then another model is DrugFlow, which generates models for structure-based drug design that integrates continuous flow matching with discrete Markov bridges.

Demonstrating state-of-the-art performance in learning the chemical, geometric, and physical aspects of three-dimensional protein-ligand data. So, by seamlessly integrating a range of innovative AI algorithms covering molecular docking, QSAR, Molecule generation, ADMET prediction and virtual screening, DrugFlow can offer effective AI solutions for almost all critical stages in early drug discovery, including hit identification and lead optimization. So, this can be used for, you know, molecular docking, ADMET prediction, molecule generation; this is what we are interested in, and QSAR modeling, and some other tools are also there. Yes, let us talk about a case study where these generative models have been able to, you know, predict the accurate binding mode of this glucagon GLP-1R receptor agonist, LY3502970. So, this glucagon-like peptide 1 receptor (GLP-1R) agonist is a frontline treatment for diabetes, enhancing glucose-dependent insulin secretion and improving energy balance.

And the current therapies are peptide-based and require subcutaneous injection, limiting their clinical utility. However, this new molecule LY3502970 is a promising oral peptide known as a GLP-1R agonist currently in phase 3 trials. So, its binding mode has been resolved via Cryo EM in complex with GLP-1R providing valuable structural insights. However, due to its high molecular weight, accurately predicting its binding mode using standard docking algorithms remains challenging. However, you can see that by using drug flow, the authors in this paper have been able to accurately predict the binding mode of this challenging molecule, whereas the other tools were not able to predict that correctly.

Ok, and then coming to the most popular and, I would say, recent tool, which is GraphINVENT. So, it is a platform for graph-based molecule generation using graph neural nets. So, it uses a tired deep neural network architecture to probabilistically generate new molecules one bond at a time. So, all the models implemented in GraphINVENT can quickly learn to build molecules resembling the training set molecules without any explicit programming of the chemical tool. So, this is a tool developed by AstraZeneca's molecular AI team, and they have been doing excellent work in the field of generative AI for drug discovery.

So, coming to another tool again from the AstraZeneca is the REINVENT. So, it is an open-source generative tool for designing small molecules, and it uses RNNs and transformers to generate novel chemical structures. It supports tasks like de novo drug design, scaffold hopping, linker and R group design, and it uses reinforcement learning to optimize molecules for the desired property. And this is also one of the popular tools. So, you can see that it works with, you know, reinforcement learning, where the different scoring subsystems can be used, okay.

And then you can see here that those scoring systems can use either RDKit for that; based on RDKit, you can predict some properties and use them for scoring. Or you can use a QSAR model for predicting the property of your choice; the endpoints, and you can use that for scoring. Or you can use pharmacophore-based scoring methods, or you can use something like ChemProp for predicting some properties and use them for scoring. So, the idea is that the generator model generates molecules okay, and then those generated molecules go to the scoring method. And then they score those molecules, and then after scoring, a reward or punishment is given to the generator model, which changes the way it generates the new molecules.

It can also be that we can use transfer learning in this case, where you can use information from already existing molecules to design new ones. And then it has different generators where the REINVENT can generate, you know, the molecules atom by atom. And then you have the LibINVENT, where you can generate a library of molecules that can design the library, and you can also do the scaffold decoration. And then you have the LinkINVENT where the fragment linking can be done or scaffold hopping can be done. And then you have the MoltoMol, which is a transformer-based molecule optimization model.

So, this is one of the advanced models that is being recently explored by researchers and developed by the AstraZeneca molecular AI team. So, another quite interesting tool is DockStream from AstraZeneca's molecular AI team. So, it is a flexible standalone molecular docking wrapper that provides access to a collection of ligand embedders and docking backends. So, it integrates with de novo drug design platform like reinvent enabling reinforcement learning agent to generate molecules with the favorable docking scores. So, Dockstream automates the execution and analysis of various docking configurations, facilitating the identification of optimal docking protocols for specific targets and ligand sets.

So, you can see here that it uses Reinvent to generate the molecules, and then you can use various open-source tools. As well as you know the commercial tools like Schrodinger Glide, OpenEye Hybrid, or CCDC Gold, you can use open-source tools like AutoDock Vina to build those molecules. So, how do you do it? You prepare the target structure, the protein structure, and then you do, you know, the ligand embeddings. And then you perform the docking so it is fully parallelized, and you can have different write-out modes. And then you can also use the H bonds or only the core constraints as well.

And then it docks the molecules into it, and this docking score, pose, and docking score information are being used by Reinvent for generating the molecules. So, the reinventing of generating molecules and then these are being docked into, you know, the binding

pocket of a protein. And then further, the information is used as, you know, the reward and punishment for reinforcement learning to design better molecules. And then you have drug X, which is an open-source software for de novo design of small molecules using deep learning generative models in a multi-objective reinforcement learning framework. So it includes various generator architectures, scoring tools, and optimization methods to generate novel compounds with desired properties.

So, the software offers a flexible API accessible via the command line or a graphical interface suitable for users with different technical expertise. And this is what you can see here, like we have the compound data. So, which is standardized and cleaned and encoded into the RNN seek translation network and then it is further you know used to generate the focused virtual library. And then this is the correspondence between input and encoding types with generator models. This is an example of how drug X works to generate new molecules.

Okay, and then another, talking about another popular model which is being used in silico medicine for their discovery pipeline. So, GENTREL which is Generative Tensorial Reinforcement Learning. So, it is a machine learning approach for de novo design focusing on synthetic feasibility, biological target effectiveness and molecular uniqueness. So, it combines reinforcement learning, variational inference and tensor decomposition into a two-step generative algorithm. So, it uses three self-organizing maps (SOMs) as reward functions: the trending SOM, general kinase SOM, and specific kinase SOM.

And you can see that it is using, you know, generative AI. So, where the chemical space is represented by using the encoders, the chemical space is represented as a latent space, and then the generator is generating molecules from this latent space. And where you can use, you know, these three reward functions for reinforcement learning. And this is an example where these guys at In Silico Medicine used general methods to discover potent inhibitors of the discoidin domain receptor DDR1 kinase target, implicating fibrosis and other diseases in 21 days. So, six of these compounds, each complying with Lipinski's rule, were designed, synthesized, and experimentally tested in 46 days which demonstrates the potential of this approach to provide rapid and effective molecular design. So, these were some of the tools that we discussed.

So, this can be explored further for your own research projects as well as to design novel molecules having drug-like properties. So, let's come to the summary. So, generative AI is revolutionizing molecule optimization by significantly reducing development timelines by up to 73 percent in some cases.

And it enables the exploration of a vast chemical space covering millions of compounds. And the AI-driven insight allows for faster and smarter drug discovery by efficiently

navigating millions of compounds. So, I have some suggestions for further reading where you can go through these papers to learn more about some of those tools. And with that, thank you.