AI in Drug Discovery and Development

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

Week-11
Lecture-51

Welcome to the course "AI in Drug Discovery and Development." In this session, we will talk about some public AI resources for drug discovery and development. So, in earlier sessions, we have seen a number of tools that are being used for drug discovery and development at various stages, starting from target identification, its validation, hit identification, and hit to lead. Lead optimization, as well as ADMET predictive modeling and even clinical trial design and optimization. So, in this session, we will try to cover some of those tools that are both public and the latest. So, which have been released in the past six months or the past year, and they are making a big difference.

So, by the end of this lecture, you will be able to understand the value of open and public AI resources in accelerating drug discovery, as well as identify key AI tools and platforms available across different stages of the drug development pipeline. And explore core libraries, toolkits, and benchmarking platforms supporting AI-based drug design, and apply best practices for using public AI resources, ensuring reproducibility, transparency, and data integrity in research. So why do we need public AI resources? Because they have a lot of characteristics. The most important thing is the openness because it democratizes access by making data sets, models, and tools freely available to all, including under-resourced labs and startups.

So, because probably everyone knows that most of the data which is important for discovering new drugs and developing new drugs relies on the companies. And that data is their proprietary data, and they are reluctant to share it with academia and even with other companies as well. So, developing open resources and open tools so they provide access to the data and the tools that can be used by everyone in this research domain. So, it also promotes transparency and inclusivity in how algorithms are developed and used as well as fuels global innovation by allowing researchers worldwide to contribute and build upon open world. Another important characteristic of public tools is reproducibility because it allows independent verification of research outcomes, which strengthens scientific credibility, and standardized benchmarks and data sets enable fair comparison across models and studies.

It also builds trust and reliability in AI-driven drug discovery pipelines as well. The third important aspect of, you know, public resources is the collaboration because it encourages

interdisciplinary teamwork among AI developers, chemists, biologists, and clinicians. It facilitates shared innovation through open platforms and global initiatives and helps reduce duplication of efforts and speeds up the overall discovery process. So let us see some of the tools. As I said, we have seen a number of tools, like those databases, which contain, you know, the structure of drugs and structural proteins.

Right. And then we have seen several tools, like, you know, RDKit, which is one of the major tools being used extensively in drug discovery, especially in cheminformatics. And then you have deep chem, which is another excellent resource that is being used for predictive modeling and all those things. And then there are a lot of different servers that are being used for ADMET modeling as well. We have covered most of them in the earlier sessions.

So let us talk about this TXGemma, which is Google's foundation model for therapeutics. So TXGemma is Google's family of open-source foundation models designed to power therapeutic discovery through advanced biomolecular understanding and generative capabilities. So the advantages of this model are that it has multimodal and multitask capabilities. It can handle diverse biomedical tasks, including molecular property prediction, target identification, and drug-target interaction modeling. It is lightweight and open source, designed for easy deployment and fine-tuning in the real-world drug discovery pipeline.

As well as having state-of-the-art performance, it outperforms previous models across various benchmarks from the therapeutic data commons. So, how it is working is the technical approach behind the TXGemma; it is using the Gemma architecture built upon Google's Gemma large language model framework adapted for molecular and biological data. It has pre-trained and fine-tuned models available as a base model for general use and as a task-specific version for fine-tuned therapeutic applications. So, what it can do is enable faster, more scalable AI-driven drug discovery by bridging language modeling with chemical and biological reasoning. So, you can predict whether a molecule will be active or not; you can even generate new molecules as well with the help of this TXGemma.

Coming to another tool, Boltz-1, this is also becoming very popular nowadays. So, it is an open-source AI tool for biomolecular structure prediction, as a competitor of, you know, AlphaFold 3, or you can say as a substitute for AlphaFold 3. It has been developed. So, it is an open-source AI model developed by MIT's Jameel Clinic for predicting 3D biomolecular structures, including proteins, RNA, DNA, and more, with AlphaFold 3 level accuracy. So, the advantages of this Boltz 1 are that it is highly accurate.

So, it achieves AlphaFold 3 level structure prediction and is open source MIT licensed

with full access to code, model weights, and training data. And it has flexible input and output possibilities. It supports protein sequences, nucleic acids, and small molecules. So, it can model the 3D structure of all of these. And how it works is that it uses cutting-edge AI architecture to predict 3D structures with confidence scores trained on diverse biomolecular data.

And how can we use it? We can use it for, you know, identifying new drug targets and designing therapeutics. It can be used for protein engineering as well, where it facilitates protein design and interaction studies. And then it is being used for structural biology in education, where it enhances molecular understanding and serves as a teaching tool as well. Okay, coming to another tool known as AIRCHECK, which is an artificial intelligence-ready chemical knowledge base. The availability of well-curated data sets for predictive modeling is very important in the case of using AI for predictive modeling.

So, if the data we are using is not reliable, we cannot achieve better performance from the model; we cannot obtain the desired predictive power from the model. And that is how this air check is generating high-quality chemical data sets and experimental insights. So, Aircheck is a cloud-based platform developed by Postera to power AI-driven drug discovery through high-quality chemical data sets and experimental insights. So, the advantages of the AI check are that it ensures the data is AI-ready and provides curated clean datasets optimized for ML tasks in chemistry and drug design. It also has benchmarking support that includes datasets under active benchmarking to ensure reproducibility and relevance.

Reproducibility is a very big concern for all these predictive models. If a model is reproducible and the data is also reproducible, then that is the best quality data. It is scalable and accessible. The cloud hosted platform that simplifies integration with the AI model and pipelines. So how it works is that it aggregates experimental results, chemical structures, and annotations into standardized formats, as well as enabling dataset search, download, and model evaluation directly through the platform.

It is built to complement AI tools used in structure-based and ligand-based drug discovery as well. So, how can we use it? We can accelerate the virtual screening, whether it is ligand-based virtual screening or structure-based virtual screening. We can improve the predictive modeling of the existing models or the new models by using the datasets from this source and support the faster therapeutic candidate identification. Okay, another, you know, similar kind of tool is Plinder. So it's a kind of large-scale open-source data set and evaluation resource designed to advance protein-ligand interaction prediction through high-quality annotations, rigorous benchmarks, and real-world relevance.

So, the advantages of Plinder are that it has extensive coverage, comprising over 449,000 protein-ligand systems across more than 11,000 scope domains and 50,000 unique small molecules. It has a very rich annotation as well; each system includes over 500 annotations encompassing protein and ligand properties, quality matrices, and more. And then it is regularly updated. That is also one of the important characteristics of any data set. If it is regularly updated, it means that it brings quality to it.

So it features an automated curation pipeline to stay current with the Protein Data Bank. And how it works is that it aggregates experimental results, chemical structures, and annotations into standardized formats. and then it enables dataset search, download and model evaluation directly through the platform and is also built to complement AI tools using structure based and ligand based drug discovery. And we can use it for, you know, accelerating virtual screening, improving predictive modeling, and supporting faster therapeutic candidate identification. And this is, you know, from their website, so you can see how much, you know, data it has compared to the other servers.

Okay, coming to another tool, that is the Polaris Hub. It is an open-source benchmarking platform for AI-driven drug discovery by providing standardized data sets, evaluation tools, and community-driven best practices. So, the advantages of Polaris Hub are is it is having standardized benchmarks. So, it offers curated datasets and unified evaluation protocols to ensure consistency across the studies. It has a collaborative framework as well, encouraging contributions from academia and industry to expand resources and improve the model assessment.

It is transparent and reproducible, promotes open science by making datasets, code, and results freely accessible and reproducible. And how it works is that it hosts diverse datasets relevant to drug discovery, including molecular properties, bioactivity, and reaction outcomes. It provides tools for model training, testing, and comparison under a common evaluation framework, as well as supports leaderboards and collaborative challenges to drive innovation and transparency. So, how we can use it is that it enables fair benchmarking of AI models, improves reproducibility in computational drug discovery, and supports informed model selection for therapeutic development. And because benchmarking is very important for predicting or evaluating any AI model, most of the time those AI models work very well on test data sets.

But in the real-world data, they actually lack that confidence. So, we need to properly benchmark the AI models and in that process of benchmarking, this Polaris Hub can be highly useful. Coming to another tool, which is the BioEMU1. So, you know, the earlier tools we discussed were about data sets and benchmarking. This tool is specifically for understanding protein dynamics.

So, normally when we determine a 3D structure by using, you know, any method such as single crystal X-ray crystallography or single particle cryo-electron microscopy. So, what we are getting is that we are getting only one conformation of the protein molecule, but that is not the reality. Because that can exist in different conformations and because of the dynamics, it changes shape over time, and that is actually known as the dynamics. So, can we develop a tool that can be an alternative to molecular dynamics simulations because molecular dynamics simulations take a lot of computational power? Because you have to, you know, do a lot of computations, then only can you simulate or produce the dynamic ensemble of a protein. So, can we develop something with the help of AI, and that the BioEMU-1 is a solution to that, actually? So, by using AI, it predicts the ensemble dynamics and the confirmations that can exist during a time frame.

So it is a deep learning model developed by Microsoft to predict the structural ensemble of proteins, capturing how proteins flex and change shape. So the advantages of BioEMU-1 are that it is ultra fast. It generates thousands of structures per hour on a single GPU, up to 100,000 times faster than molecular dynamics simulations. And it is accurate as well; it predicts both known and unseen confirmations, and then it provides useful insights as well as estimates of folding free energies to excess protein stability. And how it works is that it is trained on, you know, AlphaFold data, MD simulations, and experimental folding stabilities.

It uses a diffusion-based generative model to sample equilibrium conformations. And talking about the applications, it enables faster drug discovery, protein design, and a deeper understanding of biological functions. So, if we can, you know, get access to that structural ensemble of a protein, what we can do is use it for structure-based drug design, where we can design novel molecules. Or we can even identify those, you know, allosteric pockets or cryptic pockets as well. So, the cryptic pockets are those pockets that are not normally seen, but with time, they appear and then they disappear.

So, with the help of these models, we can even identify the cryptic pockets as new targets for drug discovery and development. And here in this figure, you can see that it is compared with the MD molecular dynamics simulation. So, BioEMU-1 is nicely able to cover or access all those conformational ensembles that we can get from the MD simulation. Okay, so DyNA1 is another such tool. It is, you know, for inferring millisecond protein dynamics from NMR data.

So, structural NMR is another tool to determine the, you know, conformation ensembles. And, sometimes you know it can actually miss those transitions from one conformation to another. Dyna-1 is such a tool that can predict those conformational dynamics derived from

the NMR data. So, Dyna-1 is a deep learning model designed to predict protein dynamics on the micro 2-millisecond time scale by analyzing missing information in the nuclear magnetic resonance data. So, the advantages are that it uses an innovative approach, utilizes the absence of certain NMR signals to infer dynamic conformational changes in proteins, and is highly accurate.

It demonstrates a strong correlation with experimental data, validating its predictive capabilities. And then how it does this is by using data integration; it is trained on a diverse set of NMR datasets, capturing a wide range of protein motions. As well as using the deep learning framework, it employs advanced neural nets to model and predict dynamic behaviors from incomplete data. And the overall impact of DyNA1 is that it enhances our understanding of protein flexibility, which is crucial for drug discovery and understanding biological functions. So you can see here that by using the dyna-1, we can get the conformational dynamics of a protein that is missing in the structure NMR study.

Okay, so TamGen is another tool that is used for, you know, AI-driven target-aware molecule generation. So this is a kind of generative AI model developed by Microsoft Research to accelerate drug discovery by generating novel molecules tailored to specific biological targets. And you might have seen that generative AI is becoming much more popular, not only in other fields but in drug discovery as well. Because the advantage of generative AI is that it generates novel molecules, which have a lot of, you know, intellectual property value. So, the advantages of Tamgen are that it uses a target-aware generation where it utilizes a GPT-like chemical language model to design molecules with high binding affinity to designated protein targets, and it enhances the molecular quality as well.

It produces compounds with improved drug-like properties and synthetic accessibility. So, these are two odd challenges with, you know, generative modeling where getting a drug-like molecule and a synthetically accessible molecule are sometimes, you know, what the models are struggling with, but this model, Tamgen, is really good at it. So, the technical approach it uses is data integration, where it is trained on extensive chemical and biological datasets to understand the structure-activity relationships. as well as it uses a generative modeling which employs advanced neural nets to generate and refine candidate molecules efficiently. So the impact is that it has successfully identified potent inhibitors against targets like the tuberculosis CLPP protease, demonstrating their potential to expedite the drug development process.

And here you can see how it actually works. So it is taking, you know, information from the protein as well as from the ligands and then generating high-affinity molecules for the target. Coming to another tool, Ligand MPNN, which is an AI-driven protein-ligand

interaction design tool. So, it is a deep learning-based protein sequence design method that explicitly models interactions with monoprotein components such as small molecules. So, it builds upon the protein MPNN by incorporating spatial and chemical context for accurate protein-ligand binding prediction. So the advantages of ligand MPNN are that it has ligand-specific designs.

So it models interactions between protein sequences and the ligands, and it is context-aware, as it considers spatial and chemical properties of known protein atoms for precision. It outperforms the traditional methods, as it is more accurate in designing ligand-based proteins compared to classical tools like Rosetta. And how it works is that it uses neural nets to generate protein sequences optimized for binding specific ligands. It leverages protein sequence data and ligand binding patterns to design proteins with enhanced interaction capabilities. So where we can use it is that it accelerates the design of protein-based therapeutics and also optimizes protein-ligand interactions for drug discovery and development.

So this tool is basically for designing proteins that have a specific high affinity for specific ligands. So it takes the information from the ligand and design a protein which cater to that ligand only. So that is how it is, you know, different and useful from the other similar tools. Okay, coming to another tool, BoltzDesign 1, which is again a biomolecular structure predictor. So, Boltz Design is an AI-based tool developed by Zeroth for the design and optimization of proteins with desired structures and functional properties.

So, it uses ML to generate novel protein sequences based on specific goals. So again, this tool is for generating, you know, we can call it protein engineering, where we are designing a new protein because we are actually interested in its function. So, the advantages of both design is that it is AI driven, it leverages deep learning to design proteins with optimized functionality. It has a tailored design approach that customizes proteins to meet specific structural and functional requirements. Speed and efficiency are another advantage as they enable rapid design and testing of novel protein candidates.

And it is open source, providing access to the tools for research and development purposes. And how it works is that it uses a generative model to create protein sequences that fold into stable structures, and it optimizes sequences for desired functional properties such as enzyme activity, binding affinity, or stability. For example, if I want to generate an artificial enzyme having an amylase-like activity. So, I can design those proteins and enzymes with the help of such tools.

The application is therapeutic development. It can design proteins and enzymes for drug discovery and disease treatment. It can be used for synthetic biology, where it develops

proteins for bioengineering and industrial applications, and for biotechnology, where it enables the creation of new molecular tools for research and development. Okay, another tool is ScopeDTI, which is a kind of deep learning DTI drug target interaction framework. So, ScopeDTI structural co-optimization of protein-ligand binding for drug-target interaction is a deep learning tool designed to predict and optimize drug-target interactions. It improves the accuracy of DTI predictions by considering both the protein and ligand structures.

So, the key features of Scope DTI are that it is AI-driven, using deep learning for protein-ligand interaction prediction. Structural co-optimization is another feature in which it optimizes both protein and ligand structures simultaneously. and then it is contextual aware so, it incorporates the 3D structural data for more accurate binding predictions. So, how it works is that it is being trained on large data sets of known protein-ligand interactions to predict how new ligands will bind to the target proteins. and it simultaneously optimizes proteins and ligand structures for better interaction predictions.

And we can use it for, you know, drug discovery, where we can identify potential drug candidates by predicting their binding affinities; we can use it for lead optimization, where we can improve the lead compounds through binding affinity assessments. And we can use it for target identification, even where it can help us find new drug targets by evaluating the ligand interactions. Okay, another tool is AutoML, which is basically used for, you know, ADMET modeling. So, it is an automated pipeline for the safety profiling of molecules. So, AutoML ADMET is an AI-driven platform designed to automate the prediction of ADMET properties, which are absorption, distribution, metabolism, excretion, and toxicity of drug candidates, facilitating early-stage profiling in drug discovery.

So, we can filter out a molecule in the early stage to determine, for example, whether this molecule has any chances of, you know, having toxicity if we can filter them in the early stage. So, we reduce the chances of failure in the clinical trials. And if a company is, again, I am saying if a company is losing or failing a molecule in clinical trial phase 2 or phase 3, then the company is actually losing a lot of money. And you know that the phase 2 and phase 3 clinical trial failure regions are safety and efficacy. So, most of the time, those molecules show toxic side effects or adverse events, and that is the reason.

So, if we can filter them earlier, we can reduce the time as well as save a lot of money during this drug discovery and development process, and AutoML is one of the tools that can be used for that. So, the advantages of AutoML ADMET is the automation. So, it streamlines ADMET prediction workflows, reducing manual interventions, and it is high throughput in nature, generating ADMET predictions at scale for large compound libraries.

And it has accurate predictions, which leverage ML to provide more accurate safety profiles for the drug candidates. And how it works is that it uses the AutoML technique to automate model training and selection for ADMET property prediction.

And then it integrates various data sources to predict the compounds' behavior and toxicity in the human body. And the applications of this tool are for early safety profiling; it can assess the ADMET properties of drugs indicated early in the discovery process. And we can use it for lead optimization as well, where it optimizes the compounds by predicting their pharmacokinetic and toxicological properties. And we can use it for risk reduction, as it can help us reduce the likelihood of failure due to poor ADMET properties in later drug development stages.

Okay, then we will have a look at another tool called MolPhoenix. So this is a kind of specific tool which is used for, you know, phenotypic screening of molecules. There are two ways: one is, you know, a target-based approach, where we first decide the target, that okay, this enzyme is responsible for, you know, the symptoms in the disease or the pathophysiology of the disease. So just target this enzyme, and we will get rid of the disease. Another approach is that we are not looking at the target; we are just looking at the effect. So, we just treat a cell line, or you know, a tissue with the enzyme, or even an animal, actually, with the molecule and see whether the molecule is giving us an effect or not.

So, this tool is specifically used for predicting the effect of molecules on cell morphology. So, MOLPHEONIX is a foundation model developed by Recursion's Valence Lab that predicts the impact of any given molecule and concentration pair on phenotypic cell assays and cell morphology. So, it leverages contrastive machine learning to map molecular structures and phenonics images into a shared latent space, enabling accurate predictions of the cellular responses to molecular perturbation. So, the key features of molpheonics are that it uses contrastive learning, aligning molecular structures with phenomics images using a dual encoder architecture. And then it is concentration-aware as well, where it explicitly models the effect of molecular concentration on cellular morphology.

And then it is, you know, having the possibility of high throughput screening. So, it utilizes high throughput microscopy to capture cellular response to millions of perturbations. And it has the pre-trained phenomics model. It employs a pre-trained phenomics model known as PHENOM-1 to enhance the image representation as well.

So, talking about this performance, so it has achieved an 8.1 times improvement in 0 short molecular retrieval accuracy over the previous methods and it reached 77.33% top 1% accuracy in identifying active molecules. So, you can see how much big changes this model

is bringing, so how much useful it can be. How we can use it? So, we can use it for you know drug discovery where it can facilitate virtual phenomics screening to identify potential drug candidates. Or we can use it for lead optimization where we can enhance the understanding of molecular effects on cell morphology for lead optimization.

And we can use it for biological research where it can provide us insights into the relationship between molecular structure and the cellular functions. Okay so now we have had a look at like some of those tools and to be honest there are hundreds or even thousands of tools and we cannot cover all of them of course. So let us have a look at some of the best practices for using those public AI resources. So the first thing is that we need to follow the FAIR guidelines, we need to use the findable, accessible, interoperable and reusable data with proper metadata and open formats and this ensures that the data is you know reproducible and it is reliable okay and that is that is one of the you know advantage of having those open source tools another thing is that it boosts collaboration transparency and long-term reuse so we need to follow the fair principles which boost collaboration transparency and long-term reuse another thing is we need to control the version as well as we need to cite so cite all data sets tools and models with DOIs, track software and model versions which enables reproducibility and proper credit.

Another thing is we need to use the trusted and curated data sets. So, we need to rely on the platforms like TDC, MoleculeNet, ChemBL and avoid unverified sources as well as what it will do is it will minimize the bias and improves the performance. And then we need to validate beyond the benchmarks as well because sometimes those benchmarks can also be biased actually. So, we need to test models on external and diverse data, not just the leaderboard sets. So, which will enhance the generalizability and robustness of those you know tools. And then we need to cross check with experimental data always where possible validate AI outputs with wet lab or clinical data, which will ensure the scientific rigor and the real world relevance as well.

Okay, coming to the summary. So, these public AI tools, they accelerate drug discovery from target identification to clinical trials, enhancing the efficiency. And these open source platforms like TxGemma, Boltz1, Ligand MPNN improves biomolecular design. Aircheck and Polaris Hub offers curated datasets and support model benchmarking. Tools like AutoML and MolPhoenix, they enable safety profiling and phenotypic prediction. And we need to you know if we follow the best practices which ensures responsible AI use and that is you know very important if we are using the AI for any you know task.

So, with that I have a small activity for you. So, explore an open source AI tool in drug discovery, identify its capabilities and assess how it can be applied to a specific stage in the drug development pipeline. And I have some suggestion for further reading which you

can go through if you want to learn more about, you know, this area. And with that, thank you.