AI in Drug Discovery and Development

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

Week-02
Lecture-09

Welcome to the course "AI in Drug Discovery and Development." In this session, we will talk about available AI tools and platforms. In earlier session, we have seen the key applications of AI in drug discovery and development. In this session, we will have a look at the tools and techniques which are available for accelerating drug discovery and development. So, by the end of this lecture, you will be able to understand how AI tools like PandaOMIX, BioBERT, AlphaFold and DeepSEA accelerate target discovery and validation. Explore the role of DeepDock, Gnina, ChemProp and REINVENT in hit identification and lead optimization.

Recognize how ADMETlab, pkCSM, and DeepTox predict pharmacokinetic toxicity and drug-likeness during preclinical studies. And also learn how Trial Pathfinder, Antidote, and Deep6 AI optimize clinical trial design, patient recruitment, and data analysis. As well as discover how Aetion, MediData and Benevolent AI supports regulatory submissions, market surveillance and drug repurposing. The first step is the target discovery.

So, and there are multiple tools which are being used for, you know, identifying new targets with the help of AI. So, the first thing we can do is with the help of the, you know, multiomics data. So like AI driven multiomics analysis to identify new targets. And Pandomics is one of the tools which is developed by InSilico Medicine, which uses AI to integrate and analyze transcriptomics, proteomics and genomics data, as well as other omics data as well. And it helps uncover novel targets by ranking them based on factors like disease relevance, novelty and druggability.

So, this platform supports target hypothesis even in less studies diseases as well. And another tool is BioSymphony, which merges diverse omics data with machine learning to identify patterns associated with disease progression pinpointing potential therapeutic targets. And then we can use it for target mining and knowledge integration. So, some of the tools for target mining and knowledge integration are like the IBM Watson for drug discovery, which uses NLP to mine scientific literature, patents, clinical trials and internal data. It helps identify hidden connections between genes, proteins and diseases supporting the target discovery.

And then we have the Elsevier's Pharmapendium which extracts preclinical and clinical data from regulatory documents to identify emerging targets and pathways. And these are not the only tools actually because of the limitation we cannot cover all of them but these are some of the tools which are being used efficiently. And then you have the Benevolent AI. So, we see a machine learning platform that uses to accelerate drug discovery. It leverages AI to mine scientific literature, analyze biological data and predict which targets could be most relevant for a particular disease.

And then another thing we can do for the target identification is the protein structure prediction. And AlphaFold is one of the you know revolution which revolutionizes the structure of biology by predicting three dimensional protein structure from amino acid sequences. So, and understanding the protein conformation is very crucial for identifying druggable pockets especially in undruggable targets. And then we have another similar tool is the Rosetta fold, which is a complementary AI driven tool that predicts protein complexes and interactions, helping identify potential targets for protein-protein interaction modulators. So after the target identification, the next step is target prioritization.

And here we have several tools like we can use the data integration platforms such as open targets. So, which combines diverse datasets like genomics, chemical, clinical to systematically score and rank potential drug targets based on evidence strength, safety and biological relevance. Then we have the PHAROS, which is an NIH tool, which prioritizes the understudied targets by integrating data from the illuminating the druggable genome project, highlighting less explored but promising proteins. And then we can, another thing we can do is like machine learning based target ranking. And then we have a tool like DeepTarget, which uses deep learning algorithm to assess and rank disease associated proteins, incorporating data on expression levels, network connectivity and tissue specificity.

And then we have TargetDB as well, which applies AI to rank targets based on clinical relevance, novelty and competitive landscape, helping researchers focus on the most promising ones. And then the next thing is, you know, the network analysis for target selection. So, where you can use the Cytoscape and AutoML. These are the tools which combines biological network analysis with automated machine learning to identify key nodes, proteins or genes that can act as central hubs in the disease pathways. And then for the network analysis for target selection, we have the Cytoscape and then we have the StringDB as well, which uses the AI to predict functional interactions between proteins supporting the prioritization of proteins involved in disease pathways.

Okay, after the prioritization, so the next thing is the target validation. So, we have, you know, some of the validation tools such as Schrodinger's drug discovery suite. So, it offers molecular docking, free energy perturbation, molecular dynamics, simulations to evaluate target ligand binding interactions and improving early validation accuracy. And then we have the Cresset flare as well, which uses AI driven electrostatic modeling to validate binding site properties and potential ligand interactions, ensuring the target's druggability. And again, these are not the only two tools, there are plenty of them actually, which can help in silico target validation.

And then by using the in vitro methods, so you have the CRISPR screening data analysis tools like DepMap, which is developed by Broad Institute. So it integrates AI with CRISPR Cas9 data to identify essential genes in cancer cells, providing functional validation of proposed targets. And then we have the CRISPR Brain, which is an AI enhanced platform that analyzes CRISPR screens in neuronal cells, supporting target validation in neurodegenerative diseases. And after the target validation, so the next step is the hit identification. So, here we can extensively use the AI, for example, for virtual screening.

So, some of the tools like DeepDock, which leverages deep learning to predict binding poses and affinities, screen massive compound libraries quickly for promising hit compounds. So, by using DeepDock, you can screen millions and billions of compounds very efficiently without any hurdles. Then we have the G-NINA, which is another molecular docking tool that has a deep learning-based scoring function. So, it announces docking simulations with CNN improving accuracy in post prediction and scoring. And then we have AutoDock and Smina; these are molecular docking tools, but with the help of AI plugins, you can accelerate the ligand-receptor interaction analysis, and you can use them for virtual screening.

And then we have several AI-assisted ligand-based virtual screening tools, like PyRMD and DeepChem, which offers a neural network model and graph convolution to predict compound activity based on molecular descriptors and chemical structure similarity. And then we have ChemProp, which is a message-parsing neural network tool for predicting the biological activity and physicochemical properties of novel molecules. And then we have tools for AI-guided de novo molecule generation. So we have the ReLeaSE, which is a reinforcement learning-based model that generates novel drug-like molecules optimized for biological activity and ADMET properties. We have the MoLGAN, which combines a generative adversarial network and reinforcement learning to design new molecular structures with desired activity profiles.

We have the Junction Tree VAE, which encodes chemical structures as graphs, enabling the design of synthesizable bioactive compounds; in addition to that, we have REINVENT; for example, we have DrugHive. So, all those tools are you know the AI guided de novo molecule generation tools which can help us to identify the hit compounds. So, after hit identification, the next step is the hit-to-lead. So here we can use AI tools for activities in property prediction. And some of the tools are like ChemProp, which predicts multiple molecular properties such as potency, selectivity, and toxicity, and can guide us in hit prioritization, early lead development, and early lead optimization.

And then we have AlphaML, which is a fast, automated ML tool for property prediction, offering a balance between performance and interpretability. And then, we can use these; then we have tools for drug target interaction prediction like GraphDTA, which uses graph neural networks for predicting binding affinity between small molecules and protein targets, helping refine the hit compounds. And we have a deep affinity that predicts interaction strengths by learning from protein-ligand sequences and structures, ranking hits based on binding likelihood. And then we have MATCH, which is a molecular ATC hierarchy that applies deep learning to classify and rank compounds based on similarity to known drugs and their targets. And then we can use the tools we have for binding mode and conformational analysis, such as the Schrodinger Suite and Cresset Forge.

And then again, we have many, many such tools that can be used for binding mode and conformational analysis using AI. So, once we have converted a hit into a lead, the next step is lead optimization. So, here we have several tools, for example, REINVENT, which are based on reinforcement learning for lead refinement and use reinforcement learning to design and optimize lead compounds, balancing multiple properties: potency, solubility, and synthetic feasibility. And then we have MoLOpt, which optimizes molecular structures by exploring chemical space iteratively, aiming to enhance key pharmacokinetic and pharmacodynamic properties. And then we have ChemBO, which is a Bayesian optimization framework that searches chemical space efficiently to improve lead candidates' drug-like properties.

And then there are tools for structural activity relationship analysis, like DeepChem, which supports SAR modeling with deep learning algorithms and predicts how structural changes affect biological activity. And then we have ChEMBL as well, which is a bioactivity database integrated with AI models to analyze SAR patterns and guide lead optimization. We have the Cresset Spark as well, which generates novel lead analogues based on electrostatic similarity, helping chemists to explore the chemical space more effectively. And then we have tools for synthetic feasibility and retrosynthesis planning, where we can use OpenBabel, for example, to convert the fingerprints, convert the chemical files, and generate molecular fingerprints to assess synthetic feasibility and lead design. We have the RDKit, which is an open-source toolkit that supports molecular manipulation, property prediction, and retrosynthesis planning.

We have SynSpace, which is an AI-guided retrosynthesis platform that predicts reaction pathways and scores the synthetic accessibility of lead candidates. And then we have many such tools, many more tools like these, which, of course, I couldn't cover in this session. And then there are, you know, tools for free energy and stability prediction, like the Schrodinger FAB plus free energy perturbation tool, which uses MD simulation with free energy perturbation to predict binding free energy changes, helping to fine-tune lead compounds. And then you have the ASYNPA, which predicts ligand binding stability and conformational refining leads to improved bioavailability and efficacy. And then you have AlphaFold, which is for protein-ligand interactions, primarily for structure prediction.

The integration of AlphaFold with docking workflows is helping to assess conformational dynamics during lead refinement as well. So, AlphaFold is not the only tool for protein structure prediction; you can also perform molecular docking with the help of AlphaFold, and you can determine the conformational dynamics of the protein during lead refinement as well. And then, coming to the preclinical studies, these are some of the AI-based tools that can be used in preclinical studies, such as in silico toxicity prediction. Like DeepTox, which uses deep learning and deep neural networks to predict various toxicities such as mutagenicity and carcinogenicity from molecular structure. You have the ProTox2, which is a web-based platform predicting toxicity endpoints like hepatotoxicity and immunotoxicity, integrating machine learning with extensive data sets.

You have the TOX21 data challenge models, where the machine learning models are trained on the NIH TOX21 dataset for predicting compound toxicity across multiple pathways. And then you can use it; we have tools for AI-powered animal model simulation, where we can use it for virtual physiological human, which simulates human organs and systems to predict pharmacokinetics, pharmacodynamics, and toxicity, reducing reliance on animal models. And we have DILIsym, which is a mechanistic modeling tool simulating drug-induced liver injury in virtual patients, helping predict hepatotoxicity. We have the PK-Sim, which supports physiologically based pharmacokinetic modeling, enabling AI-driven simulation of drug behavior in different species. We have the tools for AI-guided biomarker discovery, like we have the Panda, which is pathway and data analysis that uses AI to identify predictive biomarkers from multi-omics data supporting preclinical study design.

And then we have the DEEPBGC, which detects biosynthetic gene clusters and potential drug candidates from microbiome data, aiding in identifying novel lead-like molecules. And then, of course, we have BioGPT, which is a transformer-based model trained on biomedical literature to accelerate biomarker discovery by extracting complex insights. OK, coming to the next one. So, we have the tools that can be used for the ADME studies: absorption, distribution, metabolism, and excretion. And some of the AI tools, ADME property prediction, are the ADMETLAB 2.0, which predicts multiple ADME properties like solubility, permeability, metabolism, and toxicity using a deep learning algorithm.

And then we have SwissADME, which is an online tool for predicting absorption, bioavailability, and physicochemical properties of drug candidates. And then we have PKCSM, which is a machine learning-based platform that predicts ADME properties from molecular structures, including intestinal absorption and blood-brain barrier permeability. And then we have drug transporter interaction prediction tools as well, like TransPred, which can predict interactions between key drug transporters such as PGP, BCRP, and OATPs, guiding ADME profiling for oral bioavailability and brain penetration. And then we have the DeepTrans, which uses a deep neural network to predict substrate specificity and inhibitor potential for key transporters involved in drug disposition.

So by using all these tools, we can perform ADME predictive modeling. And then for the metabolism and cytochrome P450 interaction, we can use FAME-3, which is an AI-driven metabolic pathway predictor focusing on CYP450 enzyme interactions that help predict drug metabolism and metabolite formation. We have the METAPRED, which is again a deep learning model for predicting sites of metabolism and likely metabolites. And then we have the SMARTCyp, which is a fast prediction tool for CYP-mediated metabolic sites based on molecular fingerprints and AI predictions. After the metabolic and ADMET studies, we have the tools for safety pharmacology as well.

So, for example, we have tools for cardiotoxicity and hERG channel prediction, where we have hERGNet, which is an AI-powered model trained on ion channel data to predict hERG inhibition, a critical marker for cardiac toxicity. And then we have the PRED-hERG, which is a machine-learning platform for predicting hERG blockage and classifying compounds as safe or risky regarding QT prolongation. And then we have the CARDIOTOX, which predicts the arrhythmia potential using in silico models of cardiac electrophysiology, helping to prioritize the safer compounds. And then we have the neurotoxicity and CNS safety prediction tools like DeepNeuro, which is an AI-based model predicting neurotoxicity by analyzing structural patterns linked to CNS side effects. And then we have ToxCast, which is a high-throughput screening tool integrated with AI for predicting neurotoxic outcomes based on bioactivity profiles.

And we have DeepCNS, which predicts blood-brain barrier permeability and potential CNS side effects, guiding early safety decisions by using artificial intelligence. And then we have tools such as hepatotoxicity and renal toxicity prediction, like DILIrank, which is an ML-based tool for predicting drug-induced liver injury that is trained on the FDA-approved drug data. And then we have DeepHep, which is an AI tool that predicts hepatotoxic potential by analyzing structural alerts and pathway disruptions. And then we have the ToxKidney, which is an AI-driven platform for predicting nephrotoxicity based on chemical structures and known toxicity data sets. We have the tools for immunotoxicity prediction as well, such as Immunotox, which uses an AI model to predict immunomodulatory effects based on compound structure and known immune response pathways.

And then we have the ToxiM, which is a deep learning framework predicting immunotoxicity outcomes, including hypersensitivity and immunosuppression risks. Okay, then we have several tools that can be used for process development and manufacturing scale-up. So, some of the tools for AI-guided chemical process optimization are Chemetica, which uses AI to design optimal synthesis pathways, minimizing steps, costs, and waste. Then we have the IBM RxN for Chemistry, which predicts chemical reactions and retrosynthesis routes, streamlining process development. We have the ASKCOS, which is an open-source AI tool that predicts reaction outcomes and optimizes synthetic routes for scale-up.

And then we have tools for bioprocess development for biologics as well, like BioXpedia, which uses machine learning to optimize cell culture conditions, maximizing yields for monoclonal antibodies and recombinant proteins. We have the BioPredic, which is an AI platform for predicting protein expression and glycosylation patterns, accelerating upstream process development. And then we have the Xpert Bioprocess, which predicts fermentation kinetics and nutrient optimization to ensure high-yield, scalable production. Then we have some tools that can be used for, you know, the AI-driven continuous manufacturing optimization. Like we have Seeq, which is an AI platform for real-time monitoring and data analysis in continuous manufacturing processes, predicting equipment performance and ensuring product quality.

And then we have Aspen Tech, which uses machine learning models to optimize chemical processes, improve yield, and control batch-to-batch variability. As well as having the SkyTree, which is an ML model for predicting maintenance and fault detection in large-scale pharmaceutical manufacturing lines. So, until now we have seen tools that can be used for, you know, hit identification, hit to lead, lead optimization, target identification, and target validation. So, if we talk about development, we have those IND enabling studies, and for that, we need an IND, which is an Investigational New Drug. For the Investigational New Drug application, we need to perform some studies that convince the regulatory agencies that these compounds are not toxic and are effective.

And then they can allow us to do the clinical trials. So, there are tools that can be used for formulation development. So, we have Formulation.ai, which predicts excipient compatibility, solubility, and stability to accelerate formulation design.

We have Molecule.one, which is an AI-guided formulation prediction focusing on solubility, dissolution, and drug delivery pathways, such as the formulation of nanoparticles or liposomes. And then we have the SeeSAR, which is developed by BiosolvetIT and supports formulation scientists with in silico stability prediction and protein-ligand interaction analysis to optimize bioavailability. And then we have, you know, the predictive toxicology and safety assessment tools like DeepTox, ToxCast, and Tox21, which is an AI-enhanced high-throughput screening data set helping to identify potential adverse effects before the clinical trials. And then we have the ADMET predictor from SimulationPlus, which is an AI-driven prediction of human-relevant ADMET properties to ensure regulatory compliance. Then we have tools for PK/PD analysis and dosing strategy simulation, such as GASTROPLUS, which uses physiologically based pharmacokinetic modeling to predict drug absorption, distribution, metabolism, and excretion.

And we have the Simcyp simulator, which is an AI-enhanced PKPD modeling tool for predicting drug interactions, poor absorption, and population-specific dosing strategies. And then we have the PK-SIM, which supports whole-body pharmacokinetics, enabling rapid evaluation of dosing strategies in virtual patients. So, after IND enabling studies, so there are tools for IND submission preparation as well. So, we have AI-driven regulatory document generation tools like RegDoc365, which automate the regulatory documentation generation, ensuring compliance with the IND requirement. We have the DocuSign CLM, which is an AI-powered contract lifecycle management tool that automates data collection and formatting for regulatory filings.

We have the Master Control, which streamlines IND submission preparation by integrating the quality management system with regulatory documentation workflows. And then we have the data curation and integrity verification tools like Benchling, which centralizes experimental data and ensures traceability, helping compile accurate non-clinical study reports and manufacturing data for the IND. We have LabKey, which is an AI-powered data integration platform that verifies data integrity and consistency across preclinical and manufacturing stages. And then we have the PerkinElmer Signals Notebook, which supports AI-enhanced data analysis and aggregation to ensure comprehensive IND dossier preparation. Then we have some tools for regulatory intelligence and compliance monitoring, such as Pharma Intelligence, which tracks regulatory trends and helps align IND submissions with evolving FDA and EMA requirements using AI-driven regulatory analytics.

And then we have IQVIA SmartSolve, which predicts potential compliance risks and flags missing data points before submission. And then we have Celegence Captis, which is an AI-based regulatory content intelligence platform to ensure submissions that meet the global requirements, such as the FDA eCTD standards. And coming to the clinical trial design and planning, we have several tools, for example, that can help us in designing the trial protocol. So like we have the Trial Pathfinder, which uses real-world data to simulate clinical scenarios, helping design adaptive trials and optimize endpoints. We have the Phesi, which is an AI model that predicts trial feasibility, site selection, and patient stratification based on historical and real-world data.

We have the antidote technologies, which support the design of decentralized trials and identify underrepresented patient populations for inclusion. And then we have tools for predictive recruitment feasibility as well, like Deep6 AI, which scans patient records to predict how many eligible participants a trial can recruit at different sites, speeding up the planning. And then we have Medidata AI, which can forecast enrollment timelines and dropout rates by analyzing past data, past trial data, and real-world evidence. And then we have tools for, you know, patient recruitment and retention. So, for the AI-driven patient matching, we have IBM Watson for clinical trial matching.

So, it analyzes the EMR to match patients with complex eligibility criteria in real time. As well as having Deep6 AI, which uses NLP and deep learning to parse unstructured medical records, we find eligible participants faster. And then we have patient trial retention and engagement tools as well, like TrinetX, which identifies high-risk dropout patients and suggests engagement strategies based on behavioral data. And then we have SaamaAI, which tracks patient engagement and adherence patterns, predicting and preventing dropouts. And then once the clinical trial is finished, we have to analyze the data.

So, we also have tools for analyzing the data. So, we have the real-time data monitoring tools. Like Medidata Rave, which uses machine learning to monitor trial data in real time, detecting data anomalies and ensuring data integrity. And we have the Covance Xcellerate as well, which is an AI-powered risk-based monitoring system to detect protocol deviations and improve trial oversight. We have several tools for AI-powered biomarker discovery as well, like Tempus, which analyzes genomic, transcriptomic, and clinical data to identify predictive biomarkers during trials. And then we have GNS Healthcare, which leverages causal AI to uncover hidden biomarkers and stratify patient response data.

Okay, then the next step is the regulatory submission approval. So, here we have several tools that can be used for AI-powered regulatory documentation, like Cunesoft, which automates the preparation of regulatory documents and submission files to meet the FDA and EMA guidelines. We have the Docubee, which is again an AI-driven document analysis tool that ensures consistency between clinical study reports and submission data. And then we have the tools for compliance monitoring as well, like IQVIA Regulatory Intelligence, which tracks evolving global regulations and flags gaps in submission dossiers. We have the Celegence CAPTIS, which is an AI-based compliance tool ensuring accuracy and adherence to regulatory standards. All the data are with regulatory bodies and they have approved the drug.

Now the drug is available in clinics, being used by patients to treat their disease. So, the next is, you know, phase 4 trial, actually. So, that is also known as the post-market surveillance as well. So, now we have some tools here as well that can be used for, you know, the pharmacovigilance studies. Some tools, like the Argus safety, automate the adverse event detection from multiple data sources, such as EHRs and social media.

And then we have benevolent AI, which uses NLP to monitor literature, patient forums, and regulatory data for emerging safety signals. And then we can use them for, you know, real-world evidence generation as well, and some tools that can be used for real-world evidence generation are Flatiron Health, which aggregates oncology real-world data for post-market analysis and label expansion studies. And then we have the action, which supports regulatory-grade RWE (real-world evidence) generation, integrating claims data, EHRs, and trial data, along with some tools for lifecycle management and drug repositioning. So, we have tools for AI-powered drug repositioning like HealX, which uses graph-based AI to predict alternative disease indications for existing drugs. Recursion Pharmaceuticals has tools that they are using to combine high content imaging with AI to discover new applications for known compounds.

And then, for competitive intelligence and market strategy, we can use tools like IQVIA Commercial Intelligence, which track competitor pipeline pricing strategies and market entry data to support lifecycle planning. And then we have Clarivate Cortellis, which monitors patent landscapes, regulatory trends, and market opportunities for lifecycle management. So, coming to the summary, so we have seen that we have, you know, amazing tools which can be used to expedite the drug discovery and development phases. So, we have we have seen Pandaomics, BioBERT, DeepMinds, AlphaFold and DeepSea, which can accelerate target identification, pathway analysis and protein structure prediction. We have seen DeepDoc, Genina, ChemProp, Reinvent, which support molecular docking, QSR modeling and generative modeling for lead refinement.

We have seen ADMET Labs, PKCSM and DeepTox, which can predict the pharmacokinetics, toxicity, bioavailability, guiding safer, more effective candidates forward. And then we have seen Trial Pathfinder, Antidote, Deep6AI, which can enhance the trial design, patient matching and real-time data monitoring to optimize the clinical trial outcomes. And we have seen other tools as well, which can track the real-world data, monitor safety and uncover new indications for the drug repurposing. So, we have seen plenty of tools and if you are working in this field, you can use some of them for accelerating your research in the field of drug discovery and development as well.

So, in the end I have an open question for you. So, how do you decide which AI tool or platform is most suitable for a specific stage of drug development so just think about it. And then I have some you know some literature for you to go through if you wanted to know further about this topic. And with that thank you.