AI in Drug Discovery and Development

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

Week-02
Lecture-08

Welcome to the course "AI in Drug Discovery and Development." In this session, we will talk about key applications of AI in drug discovery pipeline. So, by the end of this lecture, you will be able to understand the role of AI in transforming the drug discovery pipeline from target identification to post marketing surveillance. So, let us start. So, we have already seen that drug discovery is highly challenging and difficult process and it is a very lengthy process it takes around 10 to 15 years to successfully discover and develop a drug. So, that it can be available in the market for the use by the patients.

So, how what are those key applications the AI has to solve all these problems associated with the drug discovery. So, let us have a look at that. So, starting with the target discovery. So, what we can do is that we can analyze the omics data using the AI.

So, the AI integrates multi omics data, genomics, transcriptomics, proteomics, metabolomics, epigenomics to uncover disease associated genes and proteins identifying potential drug targets more efficiently. And we can also use the AI for text mining and making knowledge graphs, especially the NLP, the tools like GPT. So, what they can do is they can extract the insights from scientific literature, patents, databases to reveal emerging targets and unknown disease mechanism. Because it is always better to start with the new target and to identify a new target what we have to do is we can use the AI for analyzing all the published literature through the text mining and through the NLP tools. And then we can use deep learning for target prediction as well.

So, the AI models, they can predict novel targets by analyzing complex biological data, prioritizing proteins or pathways most likely linked to the disease progression. So, after target identification, the next step is target prioritization because we usually start with like identifying 10 to 20 targets. And then prioritize those targets based on the associated pathophysiology for the disease, whether the target is druggable or not. So, all these things we need to analyze. So, what we can do with the help of AI is we can do the AI driven network analysis.

So, ML can identify critical nodes in the protein-protein interaction network helping prioritizing target based on their centrality, connectivity and role in the disease pathways.

Also, we can do the structural druggability assessment like we can use the AI based structure prediction and pocket analysis tools Which can assess whether a target has suitable binding sites supporting druggability evaluation even for previously undruggable targets. So, some of those targets, they might have some allosteric binding pockets or they might have some cryptic pockets as well. So, those cryptic pockets, they are not kind of permanent pockets. So, they just appear for a short time during the dynamics of that protein ligand interaction.

So, with the help of these structure prediction tools like AlphaFold and other tools, so we can even identify those cryptic pockets as well, which can be druggable as well. So, once we have prioritized the target the next step is the experimental validation of the target. So, with the help of AI what we can do is we can use the molecular docking and drug target interaction prediction. So, these AI models they can predict binding affinity and drug target interaction which can accelerate early-stage validation by simulating how well potential drugs bind to the target. And then we can use AI in functional genomics as well, like the deep learning.

It supports CRISPR design and RNA interference prediction to guide experimental knockouts, helping validate a target's role in the disease pathways. And then another important aspect is the phenotypic data interpretation. So, whenever we are using the phenotypic drug discovery so where we are treating you know for example the cells with the compounds with the screening libraries and then we see whether those compounds they can have the desired effect on those cells or cell lines. So, and that screening that imaging can be done with the help of the AI. So, the AI enhances high content imaging and phenotypic screening analysis, identifying cellular changes and confirming the target's biological relevance.

Because until and unless we validate that the target, so it is not a very good idea to start discovering small molecules which can engage that target in the case of drug discovery. So, cell painting is one of the techniques which is nowadays it is highly used for doing you know those phenotypic screening. So, it is a high content imaging assay that stain different cellular components like nuclear, cytoplasm, mitochondria with multiple dyes creating a fingerprint of cells morphology. So, this fingerprint captures a broad range of cellular features, shape, texture, intensity, spatial organization providing rich multi-dimensional data. So, the AI and machine learning models, they are increasingly integrated into cell painting workflow to extract and analyze complex morphological patterns.

Like deep learning models like CNN, they can identify subtle biological relevant phenotypic changes that may not be visible to the human eye. And it can help in clustering and classifying the cell states as well. So, it can help cluster similar cellular phenotypes,

identifying how different compounds or genetic modifications affect the cells. And further, it can also help with the predicting the target activity. So, those ML models trained on the cell painting data can predict which target or pathways are likely affected by a drug or genetic perturbation.

So, it is like target deconvolution identifying if there is an active compound. So, what is the target of that, that active compound? So, when you do this phenotypic screening, you know that this compound is working, but you do not know the target of that compound. So, cell painting with the help of AI can be useful in that case, in the case of target prioritization, target deconvolution. So, it can also help in feature extraction from cell painting images and here the goal is to transform raw images into a set of quantitative features representing cell morphology, texture, intensity and spatial organization. So, AI supports this in two key ways one is the traditional feature extraction like there are tools like cell profiler which extract predefined features like nuclear size, cytoplasmic granularity, mitochondrial texture etc and this creates a structured data set of hundreds to thousands of features per cell.

And then we have the DL based feature extraction, where we use the CNN, which can learn image representation directly without relying on predefined features. And models like DeepCell or EfficientNet, they capture more complex high dimensional cellular patterns that may be missed by the traditional pipelines. In the case of AI powered phenotypic profiling and clustering, so once the features are extracted, the AI models, they cluster and classify phenotypes using, for example, unsupervised learning Where one can use TSNE, UMAP or k-means clustering where it groups cells with similar morphological patterns helping identify distinct phenotypic classes. And this is key to the HIT triaging, identifying compounds that induce a desired phenotypic signature. So, and then the supervised learning trains models on the label data to recognize specific cellular states like mitotic arrest or autophagy.

For example, AI can compare compounds fingerprint to known profile to predict its mechanism of action or the off-target effects. Okay, so now once we have done that, so the next step is the predictive modeling for target validation. So AI models, they trained on phenotypic data. They can predict target engagement. The AI learns phenotypic patterns associated with specific target inhibition.

For example, ChAT inhibition leads to altered vesicle morphology. or drug target interaction coupled with molecular doping data, it can predict whether the observed phenotypes correlate with the drug's intended target and then the off-target effects where the abnormal phenotypic fingerprints, example signs of cellular stress can suggest unwanted secondary interactions. Okay, so once we have identified and validated the

target, the next step is the hit identification and this is another important step. So how AI can, the key application of AI in the case of hit identification is the AI driven virtual screening. The deep learning models, they can rapidly screen massive compound libraries predicting binding affinity, drug likeness and ADMET properties to shortlist the potential hits.

So now, why do we need that virtual screening? Because if we wanted to screen a very large library like 50,000 compounds or 100,000 compounds or even 1 million compounds, in a phenotypic screen or in a in a high throughput screen assay setup against a target as well. So, we will need a lot of time, a lot of money and you know, a lot of resources as well to screen those many compounds. So, instead can we can we develop some computer algorithm algorithms which can screen those compounds virtually using those computational models which can predict whether a compound will be active or not. So, that is the process you know called as virtual screening. So, in this case we have both like you know structure-based virtual screening and ligand-based virtual screening where in the structure based virtual screening we uses tools like molecular docking, molecular dynamics etcetera where we are utilizing the three dimensional structure of the target like protein or receptor or ion channel.

While in the ligand-based virtual screening tool, we usually use the structure of the small molecules which are known active compounds. So, by using their structure, we just try to identify similar compounds, compounds having similar structure to those initial active compounds by hypothesizing that if the structure is similar, then they will have the similar bioactivity as well. And then in the hit identification, we can also use generative models for designing novel compounds. So, the AI architectures like variational autoencoders and generative adversarial network, they can design drug like molecules with optimized pharmacophore and chemical diversity. And then we can of course, predict the drug target interaction.

So, we can use the graph neural net and transformer models to predict the binding probability and molecular interactions even for less studied target as well. Moving from hit identification to hit to lead because once we have got a hit compound, we need to convert it into a lead compound. And lead compound is an active compound with desired properties like desired pharmacokinetic properties, desired physicochemical properties. So, what we can do with the help of AI is like AI enhances structure activity relationship analysis. So, the machine learning it maps molecular modifications to change in bioactivity guiding chemists towards improved potency and selectivity.

So, with the help of AI we can analyze the structure activity relationship data. And we can come to a conclusion that okay, if we substitute this R2 substitution with some hydrophobic

groups, so that will be improving our compounds activity. And then we can use the AI for multi-objective property prediction as well and multi-parametric optimization as well. So, the AI models, they can simultaneously predict potency, toxicity, and pharmacokinetic, balancing multiple properties to prioritize lead candidates. And of course, using the De Novo design for lead expansion.

So, we have different tools which we will be going, which we will discuss in later sessions, like reinforcement learning algorithm, which can propose new analogs with optimized functional groups guided by feedback from virtual screening and SAR models. Okay once we have got a lead compound, so from the hit to lead process, the next step is the lead optimization. So, here we use the AI for molecular property optimization. So what we can do is we can use the Bayesian optimization and deep reinforcement learning iteratively to improve the potency, solubility and synthetic accessibility. And also we can use it for binding free energy and stability prediction where the AI enhanced physics based models they can estimate binding affinity, stability and conformational flexibility refining leads to leads for better target engagement.

And then with the help of AI so we can also design the synthesis route actually that is known as retro synthesis planning and this is a very you know important field of medicinal chemistry where you design a synthetic route for any new chemical entity. So we can use neural network which can predict optimal synthetic route ensuring leads are chemically feasible and scalable for synthesis. Okay, moving to the next. So, once we have got an optimized lead, so next step is the preclinical studies.

So, we can use AI. So, the key application of AI in preclinical studies is to do the in-vivo, ex-vivo prediction. So, the deep learning models, they simulate disease progression and they can predict how a drug modulates biological pathways, reducing reliance on animal models. So, this is kind of, you know, we can reduce by using the artificial intelligence or machine learning tools. And then we can predict the bioactivity and also target engagement. So, it can analyze chemical structures and protein ligand interactions to predict efficacy against intended targets.

And then in silico PK/PD modeling as well because PK/PD is another important pharmacokinetic, pharmacodynamic is another important aspect of preclinical drug discovery. So, where we can use machine learning which can integrate physicochemical properties, preclinical data and biological system models to simulate drug concentration, half-life and efficacy over time. And then we can use AI for biomarker discovery, which can identify potential biomarkers from multi-omics data, helping monitor drug response during preclinical testing. And then we can also use it for automated report generation. So these NLP models, especially like GPT, so they can compile preclinical findings into

structured reports, accelerating the documentation for regulatory filing.

Because when we are filing it for IND application or NDA application, so we need to provide a detailed report of all those data which we have obtained from all these studies. So here we can use the NLP for automated report generation as well. And then ADMET studies, non-compliance to the ADMET is one of the major reasons why most of the drugs they fail in clinical trials actually. So, because they, those drugs they might have like side effects or they have, you know, they do not have selectivity or they are not, you know, bioavailable.

So, all those kinds of problems. So, what we can do with the help of AI is we can predict those ADMET properties. And then we can use them as a screening model so we can even before doing all those studies. We can predict whether a molecule will be going will be going to have a good bioavailability or not or toxicity or not. So, the a we so the key applications are we can use AI for absorption prediction so we have the models like DeepADMET and ADMETlab which can predict oral bioavailability CACO2 permeability and gut transport interactions. And then we can use it for distributing distribution modeling.

So, AI can forecast tissue penetration, protein binding like albumin binding. And blood brain barrier permeability which is a very important very important criteria for developing CNS drugs because those drugs they must have must be permeable to the blood brain barrier. And then it can also analyze the metabolic pathway prediction. So, AI can also predict the metabolic pathway. So, these ML models, they can predict which enzymes, example CYP450 family, they metabolize the drug highlighting potential actives, inactive metabolites and drug-drug interaction risks.

And then we can model the excretion and clearance as well. So, AI can anticipate renal through the urine or hepatic through the bile or other excretion routes helping predict the systemic exposure. And then we can forecast the toxicity. So, toxicity risk forecasting. So, DL models like TOX21, DeepTOX, they can predict hepatotoxicity, cardiotoxicity, genotoxicity, minimizing failure rates in later stages and these are some of the major reason why those drugs they fail in the clinical trial in later phases like phase 2 or phase 3.

And then we can also do the transporter interaction modeling where machine learning can help us predict the interactions with key transporters like P-gp, OATP that affect the drug bioavailability and resistance mechanisms. After the ADMET prediction, we can use it for doing the safety pharmacological studies like we can do the cardiotoxicity risk prediction, AI models like hERGNet which can simulate ion channel interactions to predict QT prolongation, arrhythmia risk and cardiac arrest potential. We can use it for CNS safety prediction. Deep learning models, they can forecast neurological side effects like sedation,

seizures or cognitive impairment by modeling blood brain barrier penetration and neurotransmitter interactions.

We can do the off-target liability screening. So, the AI screens for unintended protein binding to identify safety risks, like the kinase of targets causing toxicity. And then we can have the, we can use the AI for reproductive toxicity, where it can predict the impact on fertility, embryo development and teratogenicity from chemical structure data. And then we can use it for respiratory safety assessment. We can simulate the dry effects on lung functions like bronchoconstriction, respiratory rate changes to anticipate pulmonary side effects. And then we can also do the immune system interaction modeling to predict the risk of immunotoxicity or cytokine storm by evaluating the drug's potential for immune activation or suppression.

Okay, after that so we can because process development manufacturing scale up is another important area in drug development So, we can use we can use the key applications of AI in this area for the route optimization So, as we discussed earlier as well that we can use reinforcement learning algorithms to optimize the synthetic pathways, selecting cost-effective high yield route with minimum waste. We can also do this crystallization and formulation prediction. These models, they can predict crystallization conditions and excipients to ensure ideal drug stability, solubility and bioavailability. Then we can do the process parameter optimization. So, the AI driven simulations, they fine tune temperature, pressure, pH and solvent composition for consistent large-scale manufacturing.

And then we can do the AI driven scale up simulation where it can predict chemical behavior during scale up from lab to pilot to the industrial batches which can avoid surprises in yield or purity. Because when we are scaling it up from the lab to pilot and to the industrial batches, so yield could be, yield and purity could be issues because you cannot expect that if a method is working in the lab environment, that will be also be giving the same yield and purity in the industrial batches. So the AI can help us to, you know, model that scale up as well. And then we can use the defect and impurity detection. So, the computer vision system, it can monitor production lines for particle aggregation, impurity formation or physical defects in real time.

And then we have the, we can use it for, use the AI for supply chain optimization as well, where it predicts the sourcing risk and optimizes raw material procurement to ensure uninterrupted cost-effective production. Okay, once we have done this, the next step is the IND enabling studies. What we can do with the help of the key applications of AI in this, in the IND enabling studies are AI-driven chronic and reproductive toxicity prediction. These models, they can predict long-term toxicities from short-term data, reducing the need for extended animal studies. And then we can do the bioanalytical method development,

where ML models, they support development of assays for accurate quantification of drugs and metabolites in plasma, tissue and urine samples.

And we can use the AI for stability studies as well, where it predicts the chemical degradation under temperature, humidity and light variation, which is also, you know, known as accelerating stability testing. And then we can do the toxicokinetic prediction where AI links systemic exposure to toxic effects correlating plasma concentration time curves with adverse event probability. And then we can do the immunogenicity risk modeling, where the ML models, they predict if the drug might trigger immune response, for example, protein aggregation leading to the formation of antibodies. And we can also use it for the dose selection optimization. So, AI integrates PK/PD, toxicology and animal data to suggest human equivalent doses for first in human trials as well.

Okay after the IND enabling studies so it can help us in the IND submission preparation as well. where it can help us in you know automated regulatory documentation where the NLP system it compiled complex data into structured IND sections reducing manual efforts and improving consistency. We can use it for the gap analysis where it can identify the missing data, statistical inconsistencies or weak justifications in the IND dossier reducing the rejection risk. We can do the virtual patient cohort simulation where it can predict human PK/PD profiles and clinical responses in diverse virtual population supporting those justification and safety margins. we can use it for predictive trial design as well where the machine learning models can recommend trial parameters example population size inclusion criteria to maximize the statistical power while minimizing the cost and then we can use it for drug labeling assistant as well and NLP extracts insights from preclinical data to propose preliminary drug labeling content like warnings or dosing guidelines And then another application is for regulatory intelligence, where it can analyze historical approvals, competitors, INDs and regulatory trends to align submission strategies for faster clearance.

So, once we are done with the IND application, so the next step is the clinical trial design and planning. So, where the AI plays very crucial role because if a drug is failing, you know, in clinical trials, then the company is losing a lot of, you know, lot of investment. So, the key applications of AI in clinical trial design and planning are the patient recruitment optimization. So, it can analyze patient record, genetic data and real-world evidence to identify eligible participants faster improving recruitment timelines. We can use it for virtual cohort simulation where those models, they can simulate diverse virtual patient populations to optimize inclusion exclusion criteria and reduce enrollment bias.

We can do the adaptive trial design where it supports flexible trial design that adjusts doses, sample size or arms based on interim data, increasing efficiency and reducing cost. We can also identify the site. We can also use it for site selection and feasibility prediction, where

it identifies optimal trial sites by evaluating historical site performance, investigator expertise and local patient pools. And then, we can use it for synthetic control arms as well. Instead of a traditional placebo group, AI models, they create virtual control arms using historical data, reducing the number of patients needing a placebo.

And then another application of AI in clinical trial design and planning is the risk-based monitoring strategy where the ML models they can predict sites or patients at higher risk or protocol deviation or dropouts allowing proactive resource allocation. Okay, once we have designed the trial, so the next step is the patient recruitment and retention. So, and here also AI plays very important role. So, the key applications are we can do the AI driven patient matching. Those NLP natural language processing, it can extract data from electronic health records to match patient with complex inclusion/exclusion criteria.

And then we can do the real-world data integration where AI leverages data from variable devices, social media and health apps to locate untapped patient populations. And then we can use it for personalized recruitment outreach like those machine learning models. They optimize messaging and communication channels, emails, SMS, social media to improve engagement and enrollment rates. we can do the dropout prediction and prevention. So, where AI identifies patients at risk of dropping out, enabling early interventions like reminders, nurse follow ups to boost the retention.

Because recruitment and patient recruitment and retention, especially retention is a key challenge in clinical trials. And then we can use it for virtual and decentralized trial support, where the AI powers remote monitoring platforms, reducing travel burdens on participants while maintaining the data integrity. And then it can provide us language and accessibility support as well. The NLP supports multilingual outreach and simplifies complex medical terms, improving the accessibility for diverse populations. Okay once the clinical trial is done, so we have the clinical trial data.

So, the AI has key applications in clinical trial data analysis as well where we can use it for real-time data monitoring. So, AI continuously monitors patient data streams, flagging adverse events or efficacy trends early for faster decision-making. And we can use the AI for biomarker response analysis, where deep learning models, they correlate biomarkers with treatment response, helping identify patient subgroups which are most likely to benefit from the treatment. And then we can use it for adaptive dose response prediction where the ML models, they refine dosing regimen mid-trial by evaluating patient outcomes and pharmacokinetics. We can use it for unstructured data analysis as well where NLP extracts insight from clinical notes, imaging reports and patient diaries to complement structured trial data.

And then we can use the AI for endpoint prediction as well, where it predicts the likely clinical outcome. Example, progression-free survival or symptom reduction based on early-stage data, accelerating go-no-go decisions. And then we can use it for bias and anomaly detection as well, where AI detect data anomalies, bias patterns or protocol deviations, ensuring data quality and regulatory compliance as well. Okay, once we have analyzed the clinical trial data, the next step is the regulatory submission and approval. where we can use AI for automated clinical study reports, where it can compile the trial data into submission-ready reports, reducing manual writing and formatting time.

We can do the AI-powered statistical analysis as well, like the ML model, they perform complex statistical analysis, like subgroup analysis or survival modeling, ensuring robust and defensible results. Then we can use it for comparative effectiveness simulation as well where the AI models they compare the new drugs performance to current standards of care strengthening market access arguments. And then we can use it for real-time audit trail analysis where it can track the data changes and maintain a transparent audit trail ensuring regulatory readiness for inspection as well. And then we have the, we can use it for safety signal detection where AI mines clinical data for unexpected adverse event patterns proactively addressing safety concerns before submission. And then we can use it for regulatory strategy optimization as well where the ML analyzes past approvals and regulatory feedbacks to fine tune submission strategies and reduce approval delays.

Okay, so once we have got the approval for the drug and now the drug is available in clinic for use by the patients. So, the next step is the phase four clinical trial, which is also called as post approval market surveillance to determine the, you know, the adverse effects actually. So, what we can do is we can use the AI for doing the pharmacovigilance. So, the machine learning, it continuously scans real world data, the EHRs or social media or case reports for emerging adverse events and safety signals.

And then we can use it for automated adverse event reporting as well. The NLP system extracts data from unstructured sources like physician notes to speed up the adverse event reporting compliance. And then we can use it for real world effectiveness analysis. where the AI compares trial results to real-world outcomes, identifying gaps or new insights into the drug's performance in diverse populations. And then we can use it for, you know, the market uptake prediction. The ML models, they can forecast patient adoption rates, regional demand and competitor positioning, guiding post-launch strategy.

And we can use the AI for label expansion as well where the deep learning models they can suggest new therapeutic areas or subpopulations that may benefit from the approved drug supporting label expansion submissions. And then we can use it for personalized treatment pathways as well. The AI integrates patient data, biomarkers and treatment

history to recommend personalized dosing regimen for improved outcomes. Okay and then there is another you know area where we can where the AI has key applications and that is the life cycle management and drug repositioning. So, we can we can use the AI for you know the line extensions like these models they identify opportunities for new formulations or like sustained release or combination therapies to extend the product life cycle.

And then we can use it for indication expansion discovery as well, where it can analyze multiomics data, literature data, real world data to uncover new therapeutic indication for the drug. And then we can use it for competitive intelligence monitoring, where the machine learning it tracks competitor activity, market trends and clinical trial landscapes, guiding lifecycle strategy adjustments. And then we can use it for formulation and delivery innovation as well. It predicts the improved drug delivery mechanism like nanoparticles or patches to enhance patient convenience and adherence. And then we can use the AI for generic defense strategies where the ML anticipates generic entry timelines, suggests patent extension or reformulation strategies.

Additionally, we can use the AI for the real-time cost optimization where it optimizes manufacturing distribution and supply chain to ensure ongoing cost effectiveness in mature markets Coming to the summary so what we have seen in this in this session is that AI identifies novel drug targets from multi-omics and literature data and it can accelerate the hit identification and optimize leads for potency and safety and it can predict the ADMET properties and toxicity to refine the candidate structures And then it improves the patient selection, trial design and data analysis as well as it supports safety monitoring, drug repurposing and lifecycle management. So, from this we can see that AI is a very powerful tool which is revolutionizing the drug discovery and development and you can see the applications of AI starting from the target identification to the post-parked surveillance and drug life cycle management. Okay, and then I have an open question for you. What is the next frontier for AI in drug development? Could AI eventually replace human-led decisions or will human expertise always be essential? And you can go through this, you know, this literature for knowing more about this. And with that, thank you.