AI in Drug Discovery and Development

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

Week-11
Lecture-53

Welcome to the course "AI in Drug Discovery and Development." In the earlier session, we talked about some of the successful case studies where AI has been used successfully to discover new drugs and targets. So, in this session, we will talk about some of the challenges in the drug discovery realm that are making the use of AI in drug discovery a little bit, you know, challenging. So, by the end of this lecture, you will be able to identify major challenges in AI drug discovery across different stages and understand limitations in data quality, representation, and model generalizability. Recognize issues in molecule generation, synthesis, and simulation validation; discuss regulatory, ethical, and translational hurdles in modern pipelines. as well as explore strategies for building reliable, scalable, and human-centric AI systems in drug discovery.

So, we have seen that AI has promised a lot of things to the pharma. For example, these could be an accelerated timeline. So, AI can compress the early stage discovery from years to months or weeks. Cost efficiency where AI promises that it can reduce, it has the potential to cut the R&D costs by up to 30 to 40 percent, especially in the preclinical phases.

It is, you know, promising the novelty as well, where AI can explore chemical space beyond human intuition, designing first-in-class and previously unexplored scaffolds. It has the possibility of, you know, personalizing the medicine and integrating with the patient-specific genomics for targeted precision therapeutics. And, of course, automation scalability, where it enables high-throughput virtual screening and data-driven optimization at scale. Drug discovery and development are highly challenging, starting from target identification to phase 4 clinical trials. There are hundreds of hurdles that can make a drug fail at any stage.

Despite the promising in silico data, most AI-designed compounds fail at synthesis, efficacy, or toxicity levels. So, the regions could be the translational gap where strong computational scores do not represent the real-world pharmacodynamics or pharmacokinetics. And the biological complexity where AI can't yet fully capture pathway crosstalks, adaptive resistance, or the off-target cascades, and the validation bottlenecks. Where the lack of robust FATLAB validation pipelines and feedback loops from failures and regulatory uncertainties means that the model lacks auditability and regulators lack guidelines for the adoption of AI tools. So, this is one of the examples, for example, where

the Sumitomo schizophrenia drug discovered with AI technology failed in two phase three clinical trials.

So, you can see how a drug is failing in late-stage clinical trials. So, the company is losing a lot of money because they have invested a lot of, you know, money in that project; if it is failing in the clinical trial, then it means that the company can go bankrupt as well. So, in this talk, we will discuss several challenges, and we will see what can be done to overcome those challenges. So, those challenges can be, you know, represented as they can come from the data representation and feature engineering. Where the data is sparse, biased, and imbalanced, there are molecular representation problems, activity cliffs, and experimental variability.

Those could be related to the model development and training, where poor generalization across the diseases, the black box nature of the models, or the predictive power versus the explanatory insights presents a limited benchmarking for those models. It could be with the molecule generation optimization simulation as well, where the methods cannot, you know, especially those de novo generative models, design molecules that are synthetically difficult to synthesize. And then the difference between the simulation and reality. And then the multi-objective trade-offs and generative AI disconnect, integrating AI and medicinal chemistry, while exploring chemical space, are still template-bound. And then the challenge could be in preclinical and clinical translations due to biological heterogeneity and toxicity blind spots.

It could be due to the cross-cutting and miscellaneous hurdles, such as the real-world integration of R&D culture, infrastructure inequalities, sea-load experience, and regulatory and ethical hurdles, such as the explainability versus performance dilemma. And then the target identification validation, where we have the experimental bottlenecks and the knowledge gap. So we will talk about all those challenges one by one, and then we will see what can be done to overcome them, as well as provide a kind of solution to these challenges. Okay, starting with the data, there could be challenges with the data representation and feature engineering. And there is a very, you know, common saying in computer science: "garbage in, garbage out.

" So if the quality of your input data is not good, of course your outcome will not be reliable. So the low-quality assay data, conflicting results across databases, and manual errors in literature mining lead to either misleading inputs for the AI models or wasted resources. So many times, when we are collecting data from literature, the quality of the data is actually an issue. Because if we are collecting data from multiple sources, we cannot ensure that all of the data is generated using the same experimental protocol. So, what is needed is how we can improve it.

So, we need to ensure the data integrity through standardized protocols and better quality control. So, if you look at those, you know, consortia-based data collection. So, they publish a guideline that states how to collect the data based on a protocol, and if the data is collected according to that protocol, then you can ensure that at least you can improve the reliability of that data. And then cross-validation is another thing that can be done; we are using multiple data sources to validate results and minimize errors in the training data. Because if your input data is not good, of course, your outcomes or predictions will be of no use.

And then there is sparse, biased, and imbalanced data. There are data availability issues such as sparse data, a large number of inactive compounds, and confirmed active compounds. For example, for a new target, there might be very few known active hits and active compounds. So, that can be a challenge, as well. And then you can have the biased data, where there is overrepresentation of well-studied targets with limited data for rare diseases or minority populations.

What does it lead to in model generation and overfitting? Okay, so what is needed here is a balanced data set. We need to increase the data diversity by including rare and neglected compound diseases and diverse populations. Or we can use, you know, synthetic data generation tools; we can use data augmentation techniques to expand data sets for better generalization. So another problem is the molecular representation problem; molecules, especially in cheminformatics, can be represented using SMILES, graphs, and 3D conformers, but none provide a perfect solution. For example, this is the structure of aspirin, this is aspirin represented as a graph, this is a 3D conformer of aspirin, this is the SMILES notation of aspirin, this is the InChI key of aspirin, and this is again our structure representation.

So, AI struggles with chemical equivalency and stereospecificity, leading to inconsistent model performance and bias in learning. So, in the QSAR chapter, we have already talked about how to process the data and how to deal with data preparation. So, that is you know very important. So, what is needed is for us to have unified standards to develop and adopt standardized molecular representations to improve model consistency. And we need to use the chemistry-aware models where we can incorporate more chemical intuition and expert knowledge into the AI models.

Then there are these activity cliffs, and that is again an important issue along with the experimental variability. So, an activity cliff is when a molecule having a very small change in structure can lead to a very large change in biological activity. For example, you have one molecule in which there is a substituted benzene ring. So, if you just replace a

methyl in that benzene ring with a chloro, so that can lead to a completely abolishment of the activity. So, that molecule with, you know, a methyl was super active, but as soon as you replaced that methyl with a chloro, it became inactive.

So, a small change in structure can lead to a large change in the bioactivity that is called the activity cliff. Okay, so those are mostly affecting the model quality whenever you are making a predictive model. So, these kinds of molecules can affect the model's accuracy, or we can say, predictability. Experimental variability is another issue where differences in assay conditions or lab protocols can cause inconsistent results as well. And this leads to model instability and failed predictions.

So what is needed here is an enhanced model, as we need to incorporate experimental variability and activity cliffs into model training to better predict real-world results. And we need standardized assays that use standardized experimental conditions to reduce variability and improve the quality of the training data. And then there are issues with the model development and training as well, like one of the issues is overfitting due to the limited labeled data. So, deep learning models require a large amount of labeled data for accurate training, and in drug discovery, high-quality labeled data are often scarce, leading to overfitting, where models perform well on training data but fail on unseen data. So, what is needed here is that we need to have, you know, data augmentation, where we use synthetic data and transfer learning to expand the labeled data set, and then we use cross-validation to implement robust validation techniques to test the model's generalization.

Okay, then another issue can be the poor generalization across diseases or populations. So the AI models trained on data sets from specific diseases or populations struggle to generalize across diverse biological contexts, leading to models that fail in novel or underrepresented scenarios. And hence, they lead to bias in predictions. The models trained on limited data sets may miss important genetic, ethnic, or disease-specific patterns, and they also lead to missed opportunities. The AI may overlook potential drug candidates for rare or under-served diseases.

So what is needed here is that we need to have diverse data sets, as well as use federated learning to improve the generalization of those models across the disease or across the population. Okay, coming to another important limitation of the models, it is the black box nature of these models, which lack interpretability. So, especially those deep learning models often operate as black boxes, providing predictions without clear explanations. So, this lack of interpretability complicates the regulatory approval process and trust from the stakeholders. So, because we don't know which of those features are important for the endpoint of predicting those properties.

So what is needed here is that we need to develop interpretable AI models. We need to develop hybrid models that balance predictive power with explainability. And we also need to develop some transparency frameworks. We need to establish industry standards for model documentation and explanation.

Okay. And then there is, you know, a confusion between predictive power and explanatory insights. So the AI models are excellent at making predictions, but they often lack the capability to provide clear biological insights. So those predictive models can forecast outcomes, but understanding why a model produces a result remains unclear. That's again like it can predict whether a molecule is toxic or not, but why it is toxic and how it is toxic, that is, you know causation means why it is doing that. So that is a little bit, you know, challenging to understand with these models.

So what is needed here is that we need to develop causal models that not only predict outcomes, but also offer biological explanations, and we need to integrate domain knowledge into those models. We need to combine AI's predictive power with domain expertise to uncover deeper insights into the biological mechanisms. Okay, another limitation is the limited benchmarking on real-world messy data. So these AI models are often benchmarked on clean, curated datasets, but real-world data is messy, incomplete, and noisy. So a model trained on this idealized dataset faces reduced model performance.

So those AI models may underperform when deployed in real-world settings due to overfitting on simplified, idealized data. so whenever we are developing those models we are using the limited data for training testing and even validation So it is showing very good results on the validation data set; we say that, okay, our model is very good. But it could, you know, be overfitted as well, which means whenever we are testing it on real-world data, that is nowhere close to reality; the models have been reduced. model performance. So, those AI models may underperform when deployed in real-world settings due to overfitting on simplified, idealized data.

and they lack the robustness, as well as those models may fail to account for the variability and noise present in actual experimental data. So, what is needed here is that we need to have real-world data sets as well as a robust evaluation of those models. Okay, some of the challenges are with the target identification validation. So, you know, as we say, prediction without biology is actually blind. So, we need to, you know, differentiate between the signal and the noise in omics data.

So, those high-throughput omics platforms generate vast amounts of data, but AI struggles to distinguish biologically meaningful signals from irrelevant noise. So, note that all of the statistically significant gene proteins are valid drug targets, and what it results in is false

leads and wasted resources. So, what is needed is how we can rectify this problem or how we can come up with a solution through biological grounding. we are integrating expert curated pathways and known drug target interactions to inform the model development. In addition to multiomics integration, we are using cross-validation across genomics, transcriptomics, and proteomics to improve target fidelity.

Okay, then there are challenges with the, you know, experiments as well; like there are experimental bottlenecks and knowledge gaps, and that is due to a lack of experimental validation pipelines. So AI predictions often stop at in silico output due to limited wet lab resources or expertise, stalling the validation of novel targets. And hence results in incomplete pathway understanding; diseases, especially complex ones like cancer and neurodegeneration, involve pathways that are not fully mapped. AI may suggest biologically implausible targets as well, so what is it leading to, leading to low translation as well as misleading results? What is needed here is that we need to develop AI tools along with the wet lab collaboration, as well as develop the knowledge graphs and causal inference. So all these computational techniques are not standalone techniques; actually, unless the outcomes from those models or methods are wet lab validated and confirmed in the wet labs, they are of no use.

Okay, and then some of the challenges related to the molecule generation optimization simulation are that AI can design the molecules, but can we synthesize them? That is, you know, one of the challenges. Synthetic feasibility is a challenge because generative models, especially de novo drug design models like General and Reinvent, often suggest molecules that are novel but practically unsynthesizable. So, the consequences are that they include toxicophores, unstable groups, or poor ADMET profiles, and they lack retrosynthesis planning, leading to a dead-end design. So, what is needed here is to integrate the retrosynthesis tools into molecule generation, as well as to have a human-in-the-loop system that can enable real-time input from synthetic chemists during AI-driven design. And then we can have the, you know, synthesis feasibility scoring as well, where those models can score whether a molecule will be synthetically feasible or not.

Okay, comparing the simulation versus reality, the biological inaccuracy in simulation is one of the challenges. Again, AI combined with MD simulations often fails to account for real biological variables. Like the solvent defects, pH or microenvironments, and the limitations are, you know, the force fields used may not capture all the relevant molecular dynamics, and the simulations may show binding, but outcomes do not always translate in vitro or in vivo. So, what we need to do here is develop improved force fields and perform multi-scale modeling. Again, the wet lab feedback loop is needed where the outcomes from these, you know, computational tools or in silico models need to be verified in the wet lab, actually.

Okay, coming to the multi-objective trade-offs and generative AI disconnect, the AI must optimize across multiple conflicting parameters like efficacy, toxicity, pharmacokinetics, and more. So, due to this disconnect, what is happening is that the models often favor one objective, like potency, at the cost of others, like safety. For example, we wanted to optimize the molecule to make it more potent. The model can design a potent molecule, but on the other hand, that molecule can be toxic because it has not taken into account the safety profile of those designed molecules. And also, the generated candidates may look promising on paper, but they fail holistic drug-likeness checks, so what is needed is to develop those multi-objective optimization frameworks where they can simultaneously optimize multiple properties.

It can be called multi-parametric optimization, or MPO, where the models can optimize the molecule based on multiple parameters such as pharmacogenetic parameters, efficacy, potency, and toxicity. So then we can use them to balance the trade-offs across the efficacy safety and PKPD. And also, we need to have an integrated evaluation pipeline so as to resolve this disconnect between multi-objective and generative AI. Okay, another disconnect is between AI and medicinal chemistry. So, it has been observed that there is, you know, a lack of integration between AI and medicinal chemistry; they often work in isolation, creating a disconnect between the design and the feasibility.

So what is it leading to? So we have promising AI-designed molecules, but they may have been deprioritized due to synthetic complexity or overlooked SAR trends. And then lost the opportunity for feedback that improves both the model and the medicinal relevance. What is needed is a co-design framework where we develop collaborative platforms for AI and chemists to iterate and work together. And then we need to, you know, train AI on real census decisions, not on the idealized data sets, because we need to understand those real-world constraints as well and train the models on those constraints. Okay, another challenge is that you know the chemical space, which is limited exploration because, if you look at the chemical space, it's actually huge.

So the AI models often explore known or slightly modified scaffolds, leading to marginal gains. So, the consequences are that you know true novelty is rare, and the chemical diversity remains unexplored. So, if you look at the chemical space, the search space is huge, but those AI models are often exploring molecules that are very close to the known space. So, that is how they lack novelty, and another problem is redundancy in the generated libraries, which reduces the discovery potential. So, what is needed here is how we can resolve this by using template-free generative models, where we encourage exploration beyond the traditional reaction rules.

And feedback from the novelty matrices, where we can guide the generation using chemical diversity and unexplored space indicators. Okay, then there are challenges with the preclinical and clinical translation as well, where we cannot say that the in silico success is more or less like clinical success. Because there is a lot of translational gap, actually many AI-designed or simulated molecules succeed in silico. But they fail in animal models or early clinical trials because the biology is so complex, and in silico, if you are simply predicting that a molecule will be active. So, that may fail at multiple steps because, as I said, the biological system is very complex, and drug discovery and development have hundreds of points where any molecule can fail.

So, what is needed here is that we need to have better physiological modeling. Integrating in vivo-like conditions into early-stage simulation and evaluation, we need to have early experimental cross-checks. We are combining AI designs with iterative in vitro and in vivo screening that can lead to improving or translating the in silico success to clinical success. And then there are biological heterogeneity and toxicity blind spots as well. So the challenge here is that we generally have oversimplified models.

The current AI tools often ignore inter-individual variability, such as sex, age, genetics, comorbidities, and the safety signals missed due to a lack of longitudinal or population-level data. So, what we need here is a population-aware model that incorporates stratified data to reflect biological diversity. And then we need to have the longitudinal data sets where we can train toxicity and safety models on time series data from real-world evidence. Then there are regulatory and ethical hurdles as well, like there is, you know, a regulatory gray zone for AI indirect discovery. Because regulators are catching up, they are not actually leading.

So, there is no clear guidelines for adaptive AI that results into delays in clinical development of promising tools and hesitation among pharma stakeholders to invest in uncertain regulatory territories. So what we need here is a regulatory sandbox where there are controlled environments to test AI tools with real-world oversight, and we need to have specific regulatory frameworks that provide tailored guidelines for adaptive algorithms, continuous learning systems, and real-time monitoring. Okay, and then there is, you know, this explainability-versus-performance dilemma. So where accuracy alone does not earn trust, it is actually because of the black box nature of deep learning models. Like the high-performance models, such as transformers, GNNs often lack human-interpretable outcomes, which cause clinicians and regulators to struggle to validate AI decisions, as well as the risk of rejecting useful tools simply due to a lack of interpretability.

So what we need here is explainable AI and a transparent model reporting system to increase accuracy and earn the trust of these end users. And then there are some challenges

with real-world integration and the R&D culture. So, the problem isn't always with the model; it is with the mindset as well. The R&D adoption is more than just algorithms, actually. There is a lot of resistance to algorithmic decision-making, so the bench scientists and clinical teams often distrust AI recommendations, especially when they are not explainable.

So, the consequences it has are partial or tokenistic AI adoption in R&D workflows, and AI tools remain isolated from experimental validation cycles. So, what we need here is a human-in-the-loop system as well as cross-team collaboration models. And then the infrastructure inequality is also one of the challenges because there is unequal access to resources. Like pharma giants, they have access to vast datasets, high-performance computing infrastructure, and cloud services, enabling large-scale AI projects. In contrast, startups and academic labs often face limited budgets and lack access to essential resources, hindering their ability to compete.

So this leads to bias in the models, as well as slow innovation, especially in academia and startups. So what we need to do here is develop, you know, those open access platforms. Not only the data sets, but even the computing resources can be shared. And then we need federated learning, as well as global partnerships between research institutions, whether in academia or the corporate world. And then that can fill this inequality in the infrastructure so that everyone can exploit AI for accelerating drug discovery and development.

And then there are, you know, the cross-cutting challenges, like the sealed expertise in AI-driven drug discovery. So usually, chemists and ML engineers live in separate worlds, where chemists often lack ML expertise, while ML engineers may lack knowledge of pharmacology or medicinal chemistry. So this misalignment results in models that are poorly validated and fail during synthesis or biological testing. So the consequences it has are the innovation block as well as slow translation. So, what we need here is hybrid teams where both ML engineers and chemists are working in collaboration on a single project, and we need to have common goals.

And then we need to bridge this language issue, which means the language of the computer, actually the language of AI, which is sometimes difficult for chemists to understand. The biggest solution to overcome all the challenges is interdisciplinary collaboration. So, if you know the pharma companies that have most of the data associated with them, they collaborate with academia or the ML engineers who are experts in AI. They collaborate with the biotech companies that have expertise in biology, assay development, or preclinical discovery. So this interdisciplinary collaboration is key to overcoming all the challenges, and we need to have better data practices as well; we need to invest in standardized, high-quality multimodal data sets.

And then we also need to ensure diversity in the data, and that diversity should come from demographics, diseases, and populations. And then we need to implement the FAIR guidelines as well, which ensure findable, accessible, interoperable, and reusable data, so we need to adapt these FAIR principles. And then we need to develop human-centric AI, as well as robust validation pipelines. We need to take care of ethical and inclusive innovation, as well as global access and capacity building. So, these are some of the solutions to the challenges that are being faced in the adaptation of AI or the successful use of AI in drug discovery and development.

Coming to the summary, our modern drug discovery is evolving with AI, but significant challenges remain across every stage from data curation to clinical translation. And the key issues include poor data quality, model overfitting, synthetic infeasibility of generated compounds, and weak real-world validation. Regulatory uncertainties, limited explainability, and infrastructure gaps further complicate the integration of AI into practice. And addressing these challenges requires interdisciplinary collaboration, robust validation and responsible AI design focused on scalability, equity and biological relevance.

So, in the end, I have an open question for you. As AI becomes more powerful in drug discovery, who should be held accountable when an AI-predicted drug fails in the clinical trial: the model, the developer, or the decision maker? And I have suggested some of the seminal articles here. So especially these two by Andreas Bender. These are excellent reviews where you can find all those limitations in detail. So what is realistic? What are illusions? Part one and part two. And then you can go through this article as well if you want to learn more about the challenges of implementing AI in indirect discovery and development. And with that, thank you.