AI in Drug Discovery and Development

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

Week-10
Lecture-48

Welcome to the course "AI in Drug Discovery and Development." In today's session, we will talk about the overview and data sources for AI in drug repurposing. So by the end of this lecture, you will be able to define what drug repurposing is and its advantages over traditional drug discovery. Distinguish between on-target and off-target repurposing strategies. Describe the key experimental and computational methods for drug repurposing. Identify major data sources used in AI-driven drug repurposing, as well as interpret AI models and platforms like repurposed drugs and TXGNN for drug repurposing.

So, just think about it as we know that drug discovery and development is a highly challenging and long process, and it is very costly. So, just think about it: if we can discover a drug in, you know, one-fifth of the normal drug discovery pipeline and with a very low cost as well. So, that is what you kind of know about the role or the application of drug repurposing. So, drug repurposing, also called drug repositioning, is a strategy for finding new therapeutic indications for existing drugs, including approved, withdrawn, failed, or investigational compounds that were originally developed for different diseases.

So, it leverages the existing knowledge of a drug's pharmacology, safety profile, dosing, and formulation. This approach helps bypass early phase development hurdles such as toxicity and ADME testing. So, this is like using our approved, or you know, experimental drug for another indication for another disease. So, in that case, we do not have to, you know, do all those optimizations and the toxicity studies. So, we are bypassing that.

So, we are saving a lot of time and a lot of money as well. So, the advantages of drug repurposing are that the first thing is the time saving, as it shortens development time by several years. And it is cost-effective because it reduces the R&D costs as we are, you know, skipping the optimization and toxicity studies. So, it saves at least 40 to 60 percent, and then it, you know, has an improved success rate. So, it uses drugs with known safety profiles.

At least the drugs that are failing due to safety concerns. These molecules that are being repurposed will not have that problem. And then there are regulatory benefits because sometimes they may qualify for accelerated approval pathways, as those drugs have already been tested in clinical trials. So, if we compare it with traditional drug discovery, which

starts with target identification, hit identification, lead optimization, preclinical development, and clinical development, finally you get a drug. So, here we are skipping all these optimization things, using molecules with established safety instead of using a new molecular entity in traditional drug discovery.

So, in drug repurposing, we already have, you know, a drug with a molecule with established safety, and we are directly jumping to clinical development. where we are testing those molecules in the clinical trial and checking their efficacy in the desired disease. So, you can reduce the development time from at least 15 years to maybe 3 to 12 years, and then the cost, which is around 2.6 billion on average, can be reduced to about 10 to 300 million US dollars. So, you can see that we are saving time as well as reducing costs.

That is why it is one of the biggest advantages of drug repurposing. And talking about the risk-reward profile for drug repurposing over traditional drug discovery. So, here in this plot, you can see that on the x-axis we have the risk and on the y-axis we have the reward, where we are saving time and money. So, in the case of traditional drug discovery, it lies here in this corner where the risk is very high and the reward is somewhere between high and low, actually. However, in the case of drug repurposing, the risk is very low because these molecules have already been tested in preclinical and clinical studies.

So, the risk is very low; therefore, the reward is very high. So, if it succeeds, if the drug repurposing strategy is successful, then that can be easily converted into a clinical drug. So, these are some of the approaches for drug repurposing. So there are experimental approaches like phenotypic screening or binding assays to identify target interactions. And then there are computational approaches like docking, pathway mapping, retrospective clinical analysis, novel data sources, and genetic associations.

Signature matching, molecular similarity, network-based approaches, text mining approaches, semantic approaches, pharmacophore modeling and screening, and MD simulations. And then there is this serendipitous aspect as well, because serendipity is one of the most successful drug discovery approaches. So, the drugs are discovered accidentally through clinical observations. So, then we can have this repurposing as on-targets and off-targets as well, where the on-target repurposing indicates that the drug is repurposed for a new disease, but it acts through the originally known biological target for that molecule. This means the molecule is binding to the target where it was binding before, but the therapeutic effect we are getting is in the case of a new disease.

So, the pharmacological mechanism will remain the same, and dosing will also likely be similar to the original use. An example is minoxidil, which was originally developed as an

antihypertensive vasodilator and was later repurposed for hair loss. It acts on the same target, which is the potential channel that improves the blood and nutrient flow to the hair follicle. And then you can have off-target repurposing as well, where the drug is used for a completely new indication via an unexpected target. So, the target is different from the original, and often it is discovered through screening, docking, or data mining.

And then because the target is now new, we need to establish the dose. So, the dose may need adjustment, and safety and efficacy must be reassessed because now the target is not the same. So, then that target might be involved in some beneficial, you know, pathways as well. So, inhibiting the target or the binding of the drug to that target can lead to side effects or adverse effects. So, in this case, efficacy and safety will need to be reassessed.

So, an example of off-target repurposing is aspirin, which was initially discovered as a non-steroidal anti-inflammatory drug, but later it was found to inhibit platelet aggregation. So, it was repurposed as an anti-platelet agent for preventing heart attacks and strokes and also explored as a molecule to inhibit prostate cancer as well. So, if we look at the general workflow of drug repurposing, So, we can have you know these four options: in strategy one, we explore the known off-targets for the new indication, where we have, for example, off-target one, off-target two, off-target three, and then we can use structure-based screening methods like docking and binding energy estimation. And in strategy 2, we do not know the new target structure. So, in that case, what we do is we can, you know, generate the target space by identifying the binding pocket, doing a comparative search, and trying to target the hits with high binding pocket similarity.

We can perform the docking and mining energy estimation, and finally, we can identify a new target for the existing drug. And then study 3 could be based on drug-target complexes, where we can use, for example, pharmacophore models and perform pharmacophore screening to identify the target. And study 4 could be when we know the new target structure, so when we have the target structures. We can, you know, use it for binding pocket detection, and then we can use docking and binding energy estimation to determine the affinity of those molecules with the new target. So, where does the data come from in the case of drug repurposing? This data can come from multiple sources.

So, these are some sources of data for drug repurposing, such as chemical and structural data that can come from various databases like PubChem, BindingDB, and PDB. Or it can be genomic and transcriptomic data; it can be proteomic and interactome data; it can be phenotypic screening and high-throughput assay data. It can be the clinical trial and FDA data, the literature and patent mining data, or the electronic health record and real-world evidence data. So, all this data is usually, you know, gathered from public resources. So, these are some examples of the data sources for repurposable chemicals, like DrugBank.

So, it contains detailed chemical, pharmacological, and pharmaceutical data of drugs, as well as sequences, structures, and pathway information of drug targets. And then you have the TCM, which contains 170,000 traditional Chinese medicine compounds that pass the ADMET filters along with the 3D structures. And then you have the e-Drug3D, which contains 1,822 compounds similar to the US Pharmacopeia of small drugs. Then you have the super drugs 2, which contain 4,600 active pharmaceutical ingredients. You have the DNP, which is the natural products subset of the dictionary of organic compounds.

And then you have the CAC drugs, which are the data set of drugs approved to be marketed in Europe, the USA, and Japan, with information about their targets and other molecular interaction networks. These are some of the data sources to explore new targets, pathways, or indications for repurposing, such as the therapeutic target database. So, which contains the information about the studied and reported proteins, RNA, DNA drug targets, as well as pathways involved in targeted diseases, and there is STITCH, which is known for predicting interactions of chemicals and proteins. Then you have the small molecule pathway database which contains information on at least 350 human small molecule pathways. And then you have the transformer, which contains the data on enzymatic and non-enzymatic transformations of various genobiotics in humans, as well as the interactions and processes of transport of drugs, prodrugs, and traditional Chinese medicines.

And then you have the Human Metabolome Database, which contains small molecule metabolites in the human body. And then you have the KEGG pathway database, which contains detailed information on targets, molecular interaction networks, and enzymes involved in the metabolism of known drugs, with reference to several relevant databases and web-based tools. Then you have the databases to train and test ML models and for predicting the binding affinity. So, you have the PDB, which contains the experimental structures of biomolecules such as proteins, nucleic acids, ribosomes, etcetera. You have the PDB bind, which contains experimentally measured IC50, KD, KI, and other binding affinity data for those known protein-ligand complexes.

Then you have the binding DB, which contains measured binding affinities of small drug-like molecules and drugs with known drug targets. And then you have the Scorpio, which consists of structurally resolved and thermodynamically characterized protein-ligand complexes. Then you have the KI database, where it has information about published and internally derived affinity values for around 50,000 to 55,000 drugs and drug candidates with GPCRs, ion channel transporters, and enzymes. Then you have the BAPPL complex sets, which contain 161 protein-ligand complexes with experimental and predicted free energies of binding. And then you have the DNA-drug complex data set, which contains

DNA-drug complexes comprising 16 minimized crystal structures and 34 model-built structures, along with experimental affinities.

And then DUDE is also there, which is, you know, a data set of decoys that you can use to train the models we discussed in earlier sessions as well. So, after talking about those databases, let us see how you handle the data for drug repurposing. So, the first step is actually data cleaning methods. This is similar to what we discussed during the library preparation. For example, when we are doing the screening, we prepare a small molecule library using a method that is similar to that.

Here the first step is removing duplicates, identifying and eliminating duplicate entries in the dataset to ensure that each data point is unique, and then we handle the missing values. For example, either we remove those missing values, you know, or we replace them with some constant. And then we need to remove the outliers; actually, we need to identify the outliers. Identifying and addressing outliers that may skew results, either by removing them or adjusting their values based on statistical methods. Data standardization ensures that data formats are consistent across different data sets, such as standardizing units of measurement or date formats.

And then we usually need to normalize the data. So, we are scaling numerical values to a common range to prevent bias in machine learning algorithms, which can be sensitive to the scale of input features. And then we need to preprocess the data. So these are some of the data preprocessing techniques we can use if we have text data. So, for unstructured data such as scientific literature, NLP techniques are used to clean and prepare text data.

This includes tokenization, stemming, and removing stop words. And then we do the feature extraction by selecting relevant features from raw data that contribute significantly to the predictive modeling process. And this may involve dimensionality reduction techniques such as principal component analysis. And then we encode categorical variables by converting them into numerical formats using techniques like one-hot encoding or label encoding to make them suitable for ML algorithms. And then we need to do the data integration, combining data from multiple sources, such as clinical trials, genomic databases, and electronic health records, into a unified dataset for comprehensive analysis.

And then for the data transformation, we apply mathematical transformations like logarithmic polynomial transformations to improve the distribution of the data or enhance the relationship between the variables. So, after that, let us see how AI can help us in drug repurposing. It can help us with target identification. So, AI algorithms can analyze large datasets to identify novel proteins or genes that can be targeted to counteract diseases.

It can help us with molecular simulations. AI enables high-fidelity molecular simulations to be run entirely by computers, reducing the need for costly physical testing. And then it can help us with candidate prioritization as well. AI can rank and prioritize lead drug compounds for further assessment, outperforming the traditional ranking techniques. It can help us with hypothesis generation, where large language models like ChatGPT can generate quality hypotheses for drug repurposing by synthesizing vast amounts of literature. And then the data integration can integrate heterogeneous biomedical data and complex relationships to predict new disease-drug links.

So, how does AI analyze omics data for drug repurposing? So, it helps with pattern recognition. So, AI algorithms can scan large-scale genomics and proteomics datasets to identify patterns indicating how drugs might interact with different molecular targets linked to the diseases. And then it can help with predictive modeling, where these models can predict target interactions and potential efficacy by analyzing genomics and proteomics data, allowing researchers to focus on promising repurposing candidates. and can help us with the network analysis where it can construct and analyze biological networks from genomics and proteomics data to uncover complex relationships among drug targets and diseases. And it can help us with the gene expression analysis, where the AI systems can compare gene expression profiles of diseases with drug-induced changes in expression to identify potential repurposing opportunities.

And of course, the multi-omics data integration allows these techniques to integrate diverse omics data to build a more comprehensive model for drug repurposing. So let us see an example of a platform called Repurpose Drugs; it is an AI-powered platform for drug repurposing. so it's an open source ml based web portal designed to identify novel therapeutic uses for existing drugs both single agents and in the combination. So, it offers researchers a systematic and scalable way to explore drug disease relationship by leveraging real world clinical data and advanced predictive modeling. So, the purpose is to accelerate drug repurposing by identifying new indications for existing drugs, exploring synergistic drug combinations for complex or resistant diseases, and reducing the time, cost, and risk associated with traditional drug development.

It employs an XGBoost-based regression model, which is a popular ML algorithm known for its high accuracy and scalability. So, it includes two kinds of data: a single drug prediction model that predicts the likelihood that an individual drug is effective for a specific disease. And then it generates another model that is a combination drug prediction model, which estimates the potential of drug combinations for treating a disease. So, the training data source that this tool is using is curated from publicly available clinical trials data from clinicaltrials.gov, ensuring high relevance to real-world applications.

And then the single drug model it contains positive data sets having 382 approved drugs and 190 diseases, and a negative data set where information on 409 drugs in 175 diseases is included. And then the combination model has, you know, 60 in the positive data set, which has 65 approved drugs in 55 diseases, and the negative data set contains 62 combinations and 39 diseases. So, the advantages it offers are that it expands the landscape of potential repurposing candidates beyond what traditional methods offer and leverages existing clinical trial outcomes rather than preclinical or omics data alone. So, this is the beauty of this tool: it leverages the clinical trial outcomes instead of just the preclinical data, because even if we go for the preclinical omics data. So, there might be chances that it can fail in the clinical trial, and it aids in hypothesis generation for experimental validation or new clinical trial design as well.

So, it can be used for, you know, academic and translational research for AI-driven hypothesis generation by pharma or biotech pipelines. And then it can act as a support tool for regulatory submissions involving drug repositioning. And it prioritizes, you know, candidates for neglected or rare diseases. Then we have another tool called TXGNN. It is a graph neural network for zero-shot drug repurposing.

So, the foundational AI model built for large-scale drug repurposing is capable of making short predictions across 1,780 diseases, including many with no known treatments. So, the main purpose is to accelerate the identification of novel drug-disease associations by transferring knowledge from well-studied diseases to poorly understood or untreated diseases. So the architecture and performance of TXGNN were validated against real-world prescribed data and existing drug-disease associations, showing high accuracy and generalizability. So the key features are that it is built on decades of curated biological, chemical, and clinical data linking drugs, diseases, genes, pathways, symptoms, and more through the medical knowledge graph. And then it uses the GNN embedding, which learns rich latent representations of drugs and diseases within the KG, capturing multimodal relationships.

And then the metric learning module trains the model to measure similarity between drugs and diseases, enabling knowledge transfer from known to unknown conditions. And the zero-shot interference capability can predict drug-disease links for new or rare diseases without any retraining or fine-tuning, a major advantage in urgent or emerging health scenarios. And then it has this TxGNN explainer as well as a built-in explainability tool that reveals interpretable multi-hop reasoning paths, helping users understand why a drug is predicted to be effective for a disease. And you know this explainability is a key issue with all those advanced models, but this tool has, you know, this explainability tool. So, it means that it is able to tell which molecule is going to be effective and why.

And then we have the Clarivate ML-driven pipeline for drug repurposing. So, Clarivate has developed an ML-powered pipeline designed to discover novel drug-disease associations using minimal starting information such as a disease name or a set of disease-related genes. So, the key features of this tool are that it has flexible input options. It accepts simple disease terms associated with the genes or related molecular features as input. And then it uses a multi-layered algorithmic framework where the molecular network analysis detects target-target and drug-target disease relationships across complex biological interaction networks.

And then, the molecular pathway assessment identifies shared pathways between drugs and diseases. And the disease similarity modeling finds analogues of known drug indications across similar disease profiles. And then the advanced feature selection ranking it is using is PLS regression to model drug-disease association. And it also implements recursive feature elimination to optimize predictive features and rank repurposing candidates. And then the integration of proprietary data sources uses Cortelli's drug discovery intelligence.

It's a drug pipeline mechanism of action and indication data. And then a Metabase, which is a curated pathway, and the target information, and OFFXTM, which is a drug target safety, efficacy, and off-target profile data. Okay, so after that, let us see some of the practical and commercial challenges in drug repurposing. Because it seems you know very simply and very attractively, we can find a drug in a very short time and with very little money, but there are actually some challenges. So the major challenge is the intellectual property issue because repurposing known drugs may limit patentability, reducing profit opportunities and discouraging pharmaceutical companies. And then you need additional clinical trials, so if higher doses or different administration routes are required, a new phase 1 trial may be necessary, and that will increase the cost of development.

And then commercial viability; the lack of patentability for repurposed drugs can make them less attractive for pharmaceutical companies to pursue. So, these are some of the major challenges with drug repurposing. So, let us discuss this case study of baricitinib, which was repurposed for the treatment of COVID-19. So, baricitinib is a JAK1 and JAK2 inhibitor originally approved for treating, you know, rheumatoid arthritis. So, in 2020, it was repurposed as a treatment for COVID-19 to reduce hyperinflammation and cytokine storms in severe cases.

So, the repurposing was driven by Benevolent AI, a UK-based AI company specializing in biomedical knowledge graphs and drug discovery. So, the benevolent AI platform used the ML-enhanced knowledge graph of biological pathways, gene interactions, and drug mechanisms, and the AI algorithm scanned thousands of drugs to identify candidates that

could modulate viral entry and inflammatory pathways. So, in the in the by using that you know technique. So, they identified baricitinib as a dual action candidate, where it had an anti-inflammatory effect by inhibiting the JAK-STAT pathway to reduce the immune overreaction, and it also had antiviral potential. It was predicted to inhibit AP2-associated protein kinase 1, which is a regulator of viral endocytosis.

So, further clinical validation was done where in the ACTT2 trial baricitinib plus remdesivir versus remdesivir alone. So, in this case, the baricitinib and the remdesivir in combination led to faster recovery time and reduced disease progression. So, then it was, you know, the FDA emergency use authorization was granted in November 2020 for the use of baricitinib for the treatment of COVID-19. So, the full FDA approval for using hospitalized COVID-19 patients was granted for this trial. So, coming to that, you know how the drug repurposing area is shaped by artificial intelligence.

So, let us see some of the future trends that are, you know, emerging. So, the foundation models and multi-disease generation are one of the things that are coming up. The emergence of foundation models like TXGNN that can generalize across thousands of diseases, including rare and untreatable conditions. As well as zero short and few short learning, it will play a central role in predicting drug efficacy without disease-specific training data. Explainable AI and transparent predictions are, you know, the future where the demand for interpretable AI will grow, especially in regulated environments, and tools like TXGNN explainer and causal inference framework will help researchers and regulators understand why a drug is predicted for a disease. And then AI-augmented drug combinations and polypharmacology, where AI will move beyond single-drug prediction to design synergistic combinations, especially for cancer, neurodegeneration, and resistant infections, where the pathology is, you know, multimodal.

In these diseases, there is no single pathology; there are multiple pathologies involved. So, that is why, you know, multi-target and network-based drug actions will be, you know, the future, along with AI-human collaboration research pipelines. So, AI will act as a co-pilot for clinicians and researchers in generating, refining, and validating repurposing hypotheses, and the emphasis will shift from AI replacing experts to AI augmenting expert intuition and evidence. So, coming to the summary, drug repurposing offers a strategic shortcut in drug development by uncovering new uses for existing drugs, leveraging known pharmacological and safety profiles. So this approach reduces time, costs, and risks compared to traditional discovery, especially when aided by regulatory advantages and pre-existing clinical data.

So, experimental and computational methods include docking, pathway mapping, and text mining, which enable efficient identification of repurposing candidates. And AI

revolutionizes drug repurposing through advanced analytics, predictive modeling, and large-scale integration of clinical and omics data. Despite challenges in patentability and commercial uptake, AI-augmented repurposing sets a new paradigm for rapid data-driven therapeutic innovation. So, in the end, I have an open question for you. So if AI can identify new uses for existing drugs by analyzing patterns across genomics, clinical data, and real-world evidence, could future AI systems autonomously design adaptive treatments that evolve with a patient's disease in real time, adjusting doses and targets like a self-learning therapy? I have some suggestions for further reading.

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